ALMA MATER STUDIORUM

UNIVERSITÀ DI BOLOGNA · SEDE DI CESENA

FACOLTÀ DI SCIENZE MATEMATICHE, FISICHE E NATURALI

Corso di Laurea Magistrale in Scienze e Tecnologie Informatiche

Realizzazione di un sistema di Social Business Intelligence basato sul motore SPSS

Tesi di Laurea in Data Mining

Relatore:

Chiar.mo Prof.
Matteo Golfarelli

Candidato:

Diego Lanzoni

Correlatore:

Dott. Matteo Francia

III Sessione

2011/2012

Indice

Elenco delle figure

Elenco delle tabelle

Introduzione

Negli ultimi anni, stiamo assistendo ad un enorme sviluppo delle reti sociali sul web, favorito anche dai recenti progressi in ambito tecnologico (pensiamo ad esempio alla diffusione dei dispositivi mobile), che consentono in tempo reale lo scambio continuo di informazioni. Le esperienze personali condivise nel web 2.0 influenzano migliaia di persone vicine o lontanissime, in modo diretto, immediato e soprattutto efficace. Questa mole di conversazioni e di informazioni presenti on-line, costituiscono una "miniera di conoscenza" incredibilmente utile, costituita da esperienze, opinioni, commenti che riguardano prodotti, servizi, brand, singoli individui ed organizzazioni, all'interno di blog, forum e per ultimi, di social network. Se non vogliamo farci sfuggire queste preziose informazioni dobbiamo ascoltare la rete, vale a dire, dobbiamo intercettare ed estrarre ogni giorno i dati rilevanti per il business, nella mole di contenuti che circola quotidianamente nel web. I dati estratti devono essere analizzati in modo continuo per poter produrre "conoscenza utile" per l'azienda. Il processo di ascolto della rete è complesso in quanto richiede tecnologie e strumenti software specifici unitamente all'adozione di metodologie comprovate. In riferimento a quest'ultimo aspetto, la sperimentazione di una metodologia standard da seguire per la realizzazione di un sistema di Social BI, è stata uno dei cardini principali di questa tesi, proprio per la

mancanza in letteratura di un approccio consolidato e certificato orientato all'adozione dei social media in ambito aziendale. Un vero sistema di Social Business Intelligence, deve fornire all'utente finale una serie di funzionalità che diano una visione completa della presenza sul web, dell'azienda o dell'organizzazione che sia. L'analisi dei dati quantitativi ricavabili dal Web non rappresenta un valore aggiunto importante, numerose aziende, che ne hanno intuito il grande valore informativo, utilizzano già questi dati per misurare la propria presenza sul web. L'impiego di misure come: il numero di "social mentions" (menzioni sui social nerwork) o come il numero di followers delle proprie pagine sociali (su Twitter o Facebook) o di altri parametri sempre orientati al conteggio del volume di traffico generato dagli utenti, non permette di estrarre il vero valore informativo contenuto nei social media. La sfida oggi si sta spostando verso gli aspetti qualitativi della presenza in rete di un'azienda, come l'individuazione delle opinioni, dei giudizi o più globalmente del cosiddetto "sentiment". Stiamo parlando di un'informazione, che ci permetta di valutare da un punto di vista qualitativo, ad esempio: la reputazione di un brand, l'opinione su un prodotto o una persona o più generalmente su una qualsiasi tematica, della quale si è interessati a conoscere l'opinione del Web. Sono queste informazioni, quelle considerate oggi con il maggior valore informativo ed in grado di fornire vantaggio competitivo per il business dell'azienda. Per dare validità a queste informazioni, occorre conoscere l'efficacia (cioè la qualità del risultato ottenuto) dei sistemi di analisi, impiegati nella rilevazione dell'opinione. Considerata la carenza di studi approfonditi che diano evidenza della validità di queste analisi, quantificandone le performance ottenibili in casi di studio reali, questo studio di tesi si propone anche di fornire una valutazione effettiva dell'efficacia del sistema di sentiment analysis (parte integrante del sistema di Social BI). Oltre a ciò, ci proponiamo di offrire indicazioni sperimentali sulla variazione delle performance riscontrate, nelle attività di verticalizzazione su un dominio di ascolto reale.

Per il raggiungimento degli obiettivi fin qui illustrati, lo studio di tesi ha previsto la realizzazione di un sistema di Social Business Intelligence, che ha riguardato sia la parte architetturale che implementativa. L'architettura (che ha incluso anche l'attività di reperimento automatico dei dati dalle fonti web e della loro memorizzazione sul database operazionale) è stata concepita nel suo complesso per mettere a disposizione dell'utente finale una serie di macro-funzionalità fondamentali per un sistema di Social BI. Le funzioni messe a disposizione vanno da quelle più semplici, come quelle basate sul conteggio (*Top Topics*,*Trend Topics e New Topics*), a quelle invece molto più complesse come l'opinion mining (o sentiment analysis), funzione chiave del prototipo realizzato. L'implementazione ha riguardato ogni aspetto del sistema: dalla realizzazione del database operazionale e dei relativi flussi automatizzati di lettura e scrittura, al processo di elaborazione del testo. In riferimento alle funzionalità citate, come precedentemente anticipato, una particolare attenzione è stata rivolta alla funzione di opinion mining, che ha richiesto, data la complessità, un'analisi e una sperimentazione ad essa dedicata, molto approfondita, su un caso di studio preciso: quello della politica nazionale italiana. Inoltre per la parte relativa all'analisi del testo, processo cardine del sistema, si è sperimentato l'utilizzo di una tecnologia basata su un approccio di tipo statistico, mediante l'impiego di un potente strumento di data mining: SPSS Modeler e del relativo modulo specifico per l'analisi dei testi, SPSS Text Analytics. In relazione a quest'ultimo aspetto, per il corretto impiego dello strumento di analisi del testo, si è resa necessaria l'implementazione di apposite strutture ("regole di TLA") non presenti nel motore di analisi iniziale, per l'estrazione di pattern nel testo al fine di garantire le macro-funzionalità sopra descritte.

Allo scopo di fornire una panoramica precisa sugli argomenti trattati, in questo studio di tesi verrà affrontato, nel primo capitolo, il tema della Social Business Intelligence e dell'importanza che social media possono ricoprire per il business, mentre il secondo capitolo sarà dedicato all'illustrazione delle tecnologie e

strumenti impiegati per il reperimento dei dati dal web (crawling dei documenti) e per l'analisi dei testi mediante il workbench SPSS. Nel terzo capitolo affronteremo la descrizione dell'architettura del sistema (compreso il database operazionale) e di tutti i dettagli implementativi ad essa connessi: come la memorizzazione dei dati sul database e tutti i flussi di lettura (dei dati grezzi) e scrittura (dei dati processati). Nel quarto capitolo, si descriveranno le caratteristiche del caso di studio affrontato (la politica italiana) e delle prime fasi della metodologia, come la creazione della tassonomia di dominio. Nel quinto capitolo, invece, si fornirà l'esperienza maturata nelle attività di arricchimento del sistema di analisi per l'ambito di ascolto, con la sperimentazione di una metodologia di verticalizzazione relativa all'approccio utilizzato. Infine nel sesto capitolo, si riporteranno tutti i test effettuati sulla valutazione dell'efficacia della funzionalità di sentiment analysis del sistema e l'osservazione dell'evoluzione delle performance, sperimentata nel processo di verticalizzazione del motore di analisi.

Capitolo 1

La Social Business Intelligence

1.1 Introduzione alla Social BI

La Social Business Intelligence è l'evoluzione della Business Intelligence applicata al Social Web. La Social BI è quindi un insieme di metodologie, processi, architetture e tecnologie che permette di trasformare dati grezzi provenienti dal web in informazioni preziose e significative per il business. A differenza della Business Intelligence tradizionale, i dati analizzati sono non strutturati e vanno ricercati al di fuori dei propri confini aziendali (o organizzativi).

L'esponenziale sviluppo che i social media hanno avuto negli ultimi anni ha aperto le porte a questo ramo della Business Intelligence. Sul web, ogni giorno, vengono comunicate esperienze, legami, idee ed opinioni, attraverso video, foto, messaggi, e aggiornamenti di stato e allo stesso modo avviene attraverso blogs, news on-line e documenti web. Questi dati, i cosiddetti "user generated contents" vengono condivisi con amici e conoscenti su social network e spesso anche in maniera pubblica su blogs, forum e portali web. Fin da subito, appare evidente come i social network, o piattaforme sociali in generale, tra cui Facebook, Twitter Youtube, LinkedIn, contengano al loro interno un immenso valore informativo, se

consideriamo la mole di informazioni che gli utenti della rete si scambiano quotidianamente (Grimes, 2010).

Nel mondo dell'impresa, ad esempio, questi dati potrebbero rappresentare un grande valore per tanti settori aziendali come: il marketing, il "customer services", "product design", la pianificazione dei servizi, l'approvvigionamento, la compliance aziendale, cioè la maggior parte dei processi di un'azienda o organizzazione, potrebbero beneficiare di tali informazioni. Inoltre, dal punto di vista aziendale, il canale con il mondo social è bidirezionale, in quanto la stesse aziende possono utilizzare questi nuovi mezzi di comunicazione per veicolare messaggi e comunicare sia con gli attuali, che potenziali nuovi clienti (Hinchcliffe, 2013).

1.2 Da BI a Social BI

Abbiamo già accennato alcune differenze tra la Business Intelligence e la Social Business Intelligence, in questa sezione vogliamo evidenziare nel dettaglio come deve evolvere la BI in Social BI, soprattutto in ambito aziendale.

Per anni, le aziende si sono servite della Business Intelligence tradizionale come supporto operazionale, tattico e strategico per il cosiddetto "decision making" (prendere decisioni) e per l'ottimizzazione di altri processi come quello delle vendite, del marketing, della produzione e della logistica, degli aspetti finanziari, ed di altri ancora. Tuttavia, i tradizionali sistemi di Business Intelligence progettati per operare su dati di tipo operazionale e transazionale, immagazzinati nei database e datawarehouse aziendali, non sono attrezzati per gestire il fiume di informazioni, proveniente dal mondo social, che potrebbe rivelarsi, così rilevante per il business (Grimes, 2010). Gli strumenti devono evolvere per portare i cosiddetti social-data al centro delle analisi aziendali, supportando gli attuali sistemi analitici presenti. Inoltre le stesse aziende devono evolvere, adattando il loro modello di business, estendendolo verso un nuovo mercato, diverso nel modo di comunicare e nel tipo di

consumatore presente. Tale modello, denominato anche come Social Business Model (SBM, rappresentato in figura 1.1) esalta l'impegno e la collaborazione con i clienti, partner e dipendenti. I processi aziendali sono così innescati o comunque influenzati dal comportamento degli utenti web (potenziali clienti) mediante gli user generated contents. La stessa azienda deve sfruttare le piattaforme social per individuare soluzioni innovative di "customer engagement" comunicando con il consumatore mediante lo stesso canale.

Figura 1.1 Social Business Model

Dall'altra parte l'azienda deve sfruttare queste interconnessioni, per ottenere una Business Intelligence sempre più collaborativa e condivisa da tutti i settori aziendali e dai propri partner. In breve, il passaggio da BI a Social BI deve prevedere un fenomeno di socializzazione dei dati e allo stesso tempo di socializzazione della Business Intelligence (Grimes, 2010).

La possibilità di ascoltare l'opinione espressa da ogni consumatore che la comunica attraverso la rete, fornisce grandi opportunità ma richiede allo stesso tempo importanti risorse organizzative e tecnologiche, che l'azienda deve necessariamente possedere. Dal lato organizzativo, la Social BI è un concetto trasversale a tutti i

processi aziendali e ai vari livelli decisionali. Questo aspetto richiede quindi una trasformazione, evoluzione anche e soprattutto nella cultura aziendale, e una comprensione dei fenomeni sociali in tutte le fasi del business: dalla progettazione di un prodotto, all'acquisizione del personale, allo sviluppo delle campagne di marketing (compresa la valutazione della loro stessa efficacia), ed infine alla gestione delle criticità improvvise. Dal lato tecnologico, il Social Business Model, richiede la memorizzazione e analisi di enormi quantità di dati, solitamente destrutturati, i cosiddetti Big Data. La gestione di una cosi grande mole di dati è quanto mai complessa, piena di problematiche che riguardano ogni fase del trattamento dei dati:

- L'estrazione dei dati dalle varie fonti come:
    - Social Networks
    - Blog
    - Forum
    - Portali web
    - Repository e documenti interni
- La manipolazione e memorizzazione
- La fase di analisi dei dati che include
    - L'estrazione di nuovi pattern
    - Riconoscimento di trend nuovi
    - Inferire informazioni rilevanti quali nomi, concetti di interesse, relazioni che tra queste intercorrono (di qualunque tipo) ed opinioni.

1.3 L'impatto della Social BI nel business

Fino adesso abbiamo parlato dell'impatto che i social-data potrebbero avere in tutti gli ambiti aziendali. Ma possiamo tradurre questo potenziale ipotetico in vero profitto per le aziende? Si consideri la seguente domanda: quali di queste informazioni possono portare ad un aumento delle rendite e dove possiamo trovare queste informazioni? Un esempio può essere l'analisi della quantità delle menzioni che il brand produce on-line, ma soprattutto del sentiment che viene manifestato su di esso, mediante opinioni espresse da parte del pubblico sui vari prodotti e servizi. Di conseguenza queste informazioni influenzeranno le attività aziendali correlate come la progettazione del prodotto, il servizio clienti,il marketing e potranno fornire valide indicazioni sulla bontà ed efficacia del proprio sistema di "advertising" e di altri processi aziendali. Ed è in questo modo che è possibile innalzare la qualità delle analisi, che saranno più complete, accurate e che produrranno risultati ancora più rilevanti (Grimes, 2010).

L'azienda, analizzando gli "user generated contents", può raccogliere fondamentali informazioni e opinioni espresse dai consumatori in rete. Il canale web può diventare uno strumento ad alta profittabilità che le aziende devono sfruttare se vogliono beneficiare delle enormi potenzialità che esso porta (Hinchcliffe, 2013). La conoscenza che possiamo ricavare da questi contenuti destrutturati è molto importante e allo stesso modo diversificata. Può riguardare la comprensione del contesto all'interno del quale l'azienda opera, l'individuazione delle best practices del settore, la verifica dei cosiddetti "unmet needs" della propria clientela e dei consumatori in generale, le opportunità che si possono creare e le potenziali criticità che devono essere gestite. Le conseguenze che l'acquisizione e il corretto sfruttamento dei social data, ha sul business aziendale sono di primaria importanza (figura 1.2) (Bughin, Byers & Chui, 2011).

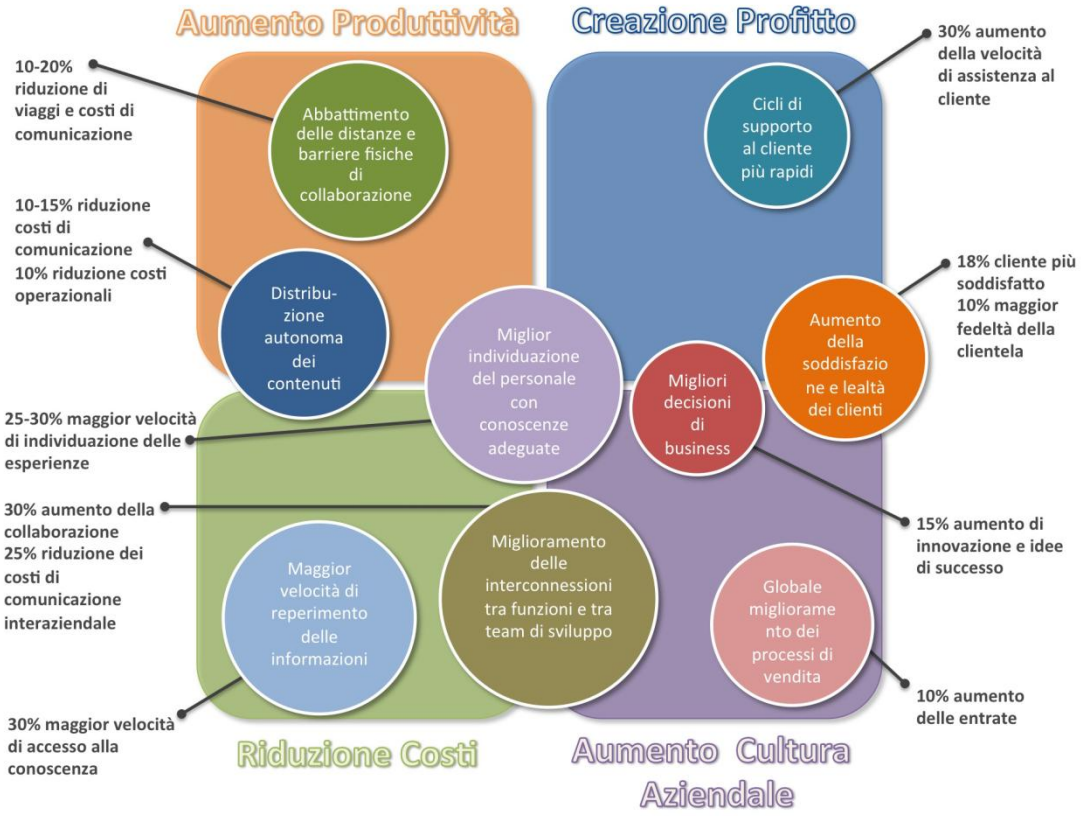

Figura 1.2 Vantaggi portati dalla comprensione dei social data sul business aziendale

Come è possibile osservare i principali vantaggi derivanti dall'impiego dei social media ricadono principalmente in due aspetti, nell'aumento:

- della velocità di accesso al know-how
- della velocità di individuazione del personale più adatto alle esigenze aziendali
- della collaborazione
- della velocità di assistenza al cliente e di conseguenza della:
    - soddisfazione del cliente
    - loyalty del cliente
- del tasso di innovazione in azienda

e l'abbattimento di una serie di costi che l'azienda può sostenere, come i:

- costi di viaggio
- costi di comunicazione esterna
- costi di comunicazione interna

1.3.1 Digital Marketing

Come già anticipato, il settore aziendale più vicino alla realtà dei social media è senza dubbio quello del Marketing, o più precisamente del Digital Marketing. Infatti, l'adozione di sistemi aperti ai social media è più tempestiva nei settori in cui il valore informativo dei social data è immediatamente compreso e subito convertito in valore competitivo sul mercato. La divisione di Digital Marketing racchiude le attività di marketing sui canali digitali. Perciò grazie alla recente diffusione, i social media sono solo l'ultimo canale, in ordine cronologico, aggiunto alla lista che comprende già la presenza di portali web, mailing list, sms come strumenti utili a massimizzare la visibilità e l'efficacia delle campagne pubblicitarie. Il ciclo operativo del Digital Marketing aggiunge alle funzioni già svolte una serie di attività specifiche, indirizzate ai canali digitali (blu in figura 1.3), e si espande con l'insieme di nuove attività indirizzate ai social media (nero in figura 1.3) (Malloy 2012). Lo sfruttamento del canale social nel DM permette di rinforzare il punto di contatto con il consumatore aprendo un rapporto di comunicazione bidirezionale, in cui l'approccio classico di diffusione "broadcast" viene sostituito da un modello interattivo, in cui anche l'utente può svolgere un ruolo attivo nella comunicazione. In questo senso il Marketing può fare da apripista verso la totale adozione del Social Business nell'azienda, diventando il principale contenitore di dati social e ritrovandosi quindi in una posizione privilegiata, dalla quale iniziare lo sviluppo delle strategie e architetture di SBI.

Figura 1.3 Il Task Breakdown del Digital Marketing

1.4 Social BI vista dalle aziende

Fin'ora abbiamo discusso approfonditamente dei social data, dell'alto valore informativo che potrebbero avere per una azienda e delle problematicità che l'adozione della Social BI può comportare. Ma come rispondono concretamente le aziende a tutto questo? Si stanno già muovendo in questa direzione o sono riluttanti al cambiamento perché ritengono che i tempi "non siano ancora maturi"? Lo studio denominato "Social Media and Enterprise BI-Analytics Connection" di Seth Grimes (BeyeNetwork), sulla base di un questionario sottoposto a più di 500 imprese di tutto il mondo, può essere utile per far luce su questi interrogativi (Grimes, 2010).

Innanzitutto occorre fare il punto della situazione sullo stato attuale della Social BI all'interno delle imprese, ed in particolare se i social data rappresentano già una

fonte di informazioni per l'azienda e principalmente per qual processo aziendale. Dall'intervista emerge, non inaspettatamente, che l'adozione dei social data è già una realtà per il settore di brand/reputation management con il 43.7%, come riportato nella figura 1.4.

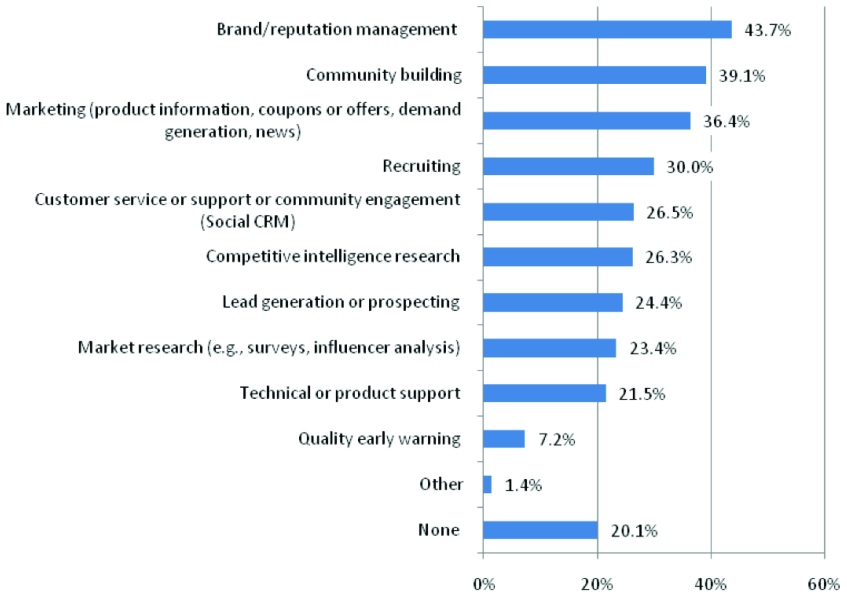

Figura 1.4 Come le aziende utilizzano attualmente i social data

Confortante che solo circa il 20% non utilizza in nessun modo i dati provenienti dal mondo social. Incuriositi da questo dato, ci domandiamo, come effettivamente l'azienda misuri la sua presenza nel mondo social. Dalle risposte (figura 1.5) emerge come le misure siano ancora piuttosto quantitative, orientate al volume di traffico generato dagli utenti, e risultino piuttosto deficitarii rilevamenti qualitativi sull'apprezzamento ad esempio del brand o dei prodotti/servizi dell'azienda. A questo punto, ci si chiede se siano già state predisposte vere e proprie piattaforme (con l'utilizzo di software) dedicate all'ascolto e all'analisi dei social media e in caso positivo, l'effettiva utilità percepita nel loro impiego. Il primo responso è piuttosto negativo, in quanto solo il 15% degli intervistati impiega già piattaforme indirizzate ai social-media.

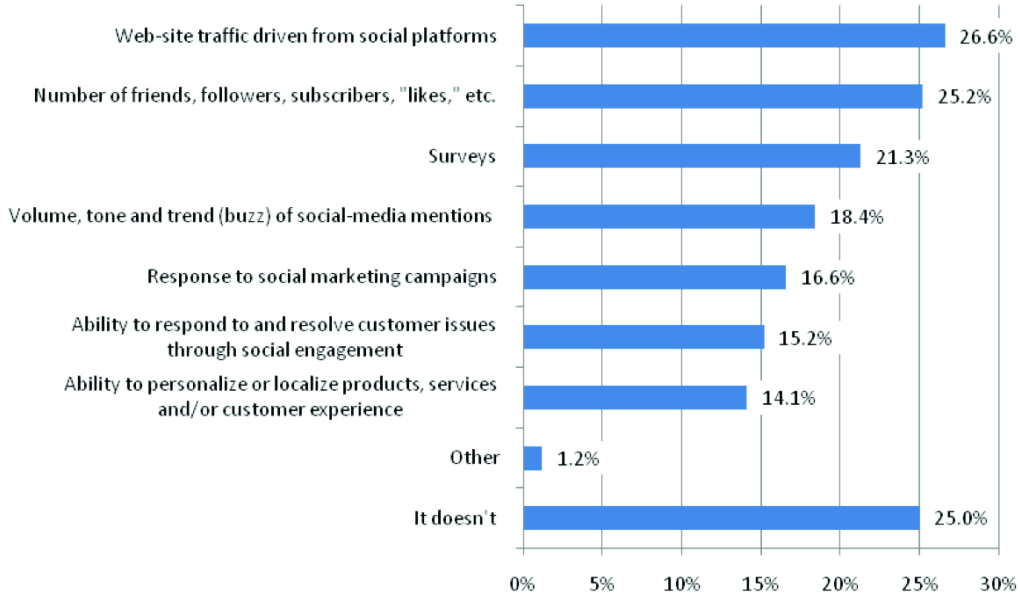

Figura 1.5 Quali misure si utilizzano per quantificare la presenza nel social

Al contempo l'utilità percepita è molto positiva, solo il 14% ritiene poco apprezzabile o ancora di non chiara importanza l'adozione di queste soluzioni (figura 1.6).

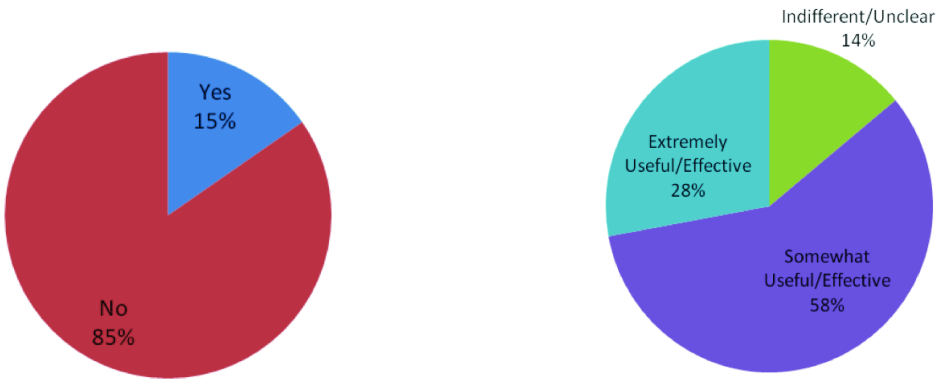

Figura 1.6 Impiego ed utilità percepita di piattaforme indirizzate ai Social media

Un altro aspetto importante più volte dibattuto in questo primo capitolo, è la socializzazione della Business Intelligence, quindi la condivisione on-line degli oggetti della BI come reports, table, e altre visualizzazioni. Ebbene una buona parte (46%) delle aziende intervistate, come si può vedere dalla figura 1.7, non è interessata alla condivisione della BI.

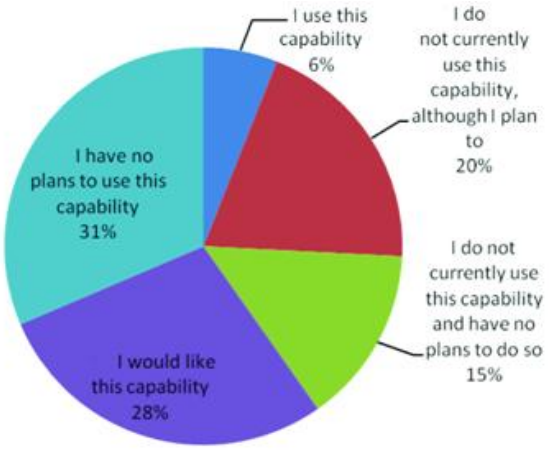

Figura 1.7 Condivisione della BI on-line

Per far si che i social-data siano veramente al centro delle analisi, è necessario che l'azienda impieghi software o servizi per l'analisi automatica del "natural language" o dei dati destrutturati. Nel capitolo 2 inizieremo a discutere delle tecnologie e strumenti per l'analisi dei "unstructured data" provenienti dal web, in particolare dei documenti testuali. Ebbene le aziende stanno già utilizzando questo tipo di tecnologie? La risposta è quasi totalmente negativa (figura 1.8), l'uso di software di Text Analytics è molto limitato e nemmeno sempre orientato all'analisi dei social media. Per concludere questa panoramica del punto di vista delle aziende verso la Social BI, è stato chiesto di esprimere tre "benefits" (vantaggi) nell'introdurre i social data nelle analisi di BI.

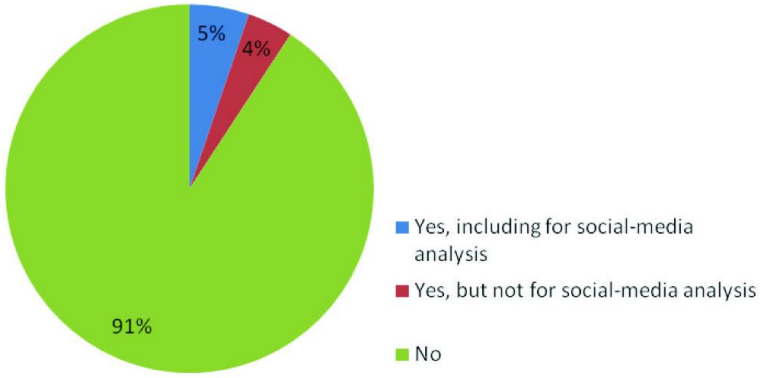

Figura 1.8 Uso di software di Text Analytics

Come possiamo vedere nella figura 1.9 i concetti maggiormente manifestati sono stati in senso assoluto quello di "customer" (cliente) e quelli relativi all'"understanding" (cioè alla comprensione di determinate esigenze del pubblico).

Figura 1.9 Termini più utilizzati

Da questo studio abbiamo potuto osservare come solamente una minoranza delle organizzazioni/aziende ascoltate, stia impiegando piattaforme social per i migliorare il proprio business, e tra queste la maggior parte non va oltre l'utilizzo di metodi analitici di base. Le maggiori applicazioni sono infatti orientate all'osservazione delle cosiddette "social-media mentions", della presenza dell' azienda sui social-network, e principalmente di tutte quelle iniziative di Digital Marketing (come aveva previsto nel paragrafo 1.3). Soltanto una piccola parte sta

giù utilizzando strumenti per l'analisi effettiva dei contenuti sociali, in particolar modo software di Text Analytics, che sono un "must" per le aziende che desiderano veramente automatizzare le analisi dei contenuti on-line. Un'altra constatazione è come la maggioranza delle aziende abbia ancora una certa diffidenza verso una struttura più sociale e condivisa dei processi di BI.

E' chiaro che la Social BI è ancora agli albori ma è altrettanto chiaro come esistano forti potenzialità di crescita. La previsione che dà Seth Grimes è chiara, e non è stabilire se la Social BI diverrà una realtà, bensì quando e come lo diventerà, e i risultati riportati in questo studio sono un primo passo per dare risposta a questi interrogativi (Grimes, 2010).

Capitolo 2

Tecnologie e strumenti adottati per la realizzazione del sistema

Questo capitolo è dedicato alla descrizione degli strumenti e tecnologie sperimentate ed adottate nel processo di reperimento ed elaborazione del testo proveniente da fonti web. In primo luogo illustreremo le tecnologie impiegate per la fase di crawling dei documenti, successivamente introdurremo il tema del Text Mining. In seguito sarà mostrato nel dettaglio un potente strumento di data mining: SPSS Modeler Premium con particolare attenzione al modulo dedicato all'analisi del testo, SPSS Text Analytics.

2.1 Crawling dei documenti

Il crawling è l'attività che permette di recuperare in maniera automatica documenti provenienti dal web. Un crawler (detto anche spider o robot), è un software che analizza i contenuti di una rete (o di un database) in un modo metodico e automatizzato. Un uso estremamente comune dei crawler è in ambito Web. Generalmente, il crawler si basa su una lista di URL da visitare, fornita dal motore di ricerca. Durante l'analisi di un URL, identifica tutti gli hyperlink presenti nel

documento e li aggiunge alla lista di URL da visitare. Il crawler può essere utilizzato anche per scopi più specifici, per reperire ad esempio solo un certo tipo di documenti (di interesse) da fonti web precompilate. Per fare ciò il crawler, oltre a necessitare della lista degli URL da visitare, richiede di essere parametrizzato in altri aspetti. Sappiamo infatti che non tutto quello che è presente in una pagina web è interessante, alcune parti sono destinate ad annunci pubblicitari, link ad altre sezioni non utili ai fini della ricerca. E' opportuno, quindi, per ambiti specifici configurare il crawler al fine di evitare la raccolta di informazioni non di interesse per il proprio dominio di ascolto, permettendo il reperimento di informazioni "pulite". Non sempre è possibile raccogliere solo ed esclusivamente l'informazione desiderata, ma molto spesso analizzando la struttura delle pagine web è possibile individuare un template comune a tutte le pagine web dello stesso sito. Abitualmente, il contenuto dell'informazione desiderata, è racchiuso in un tag "div" ricorrente (che molto spesso viene nominato "content"). La parametrizzazione del crawler ci consentirà, come vedremo, di escludere buona parte della pagina web e di prelevare solo il testo proveniente da determinati tag html (esempio in figura 2.1).

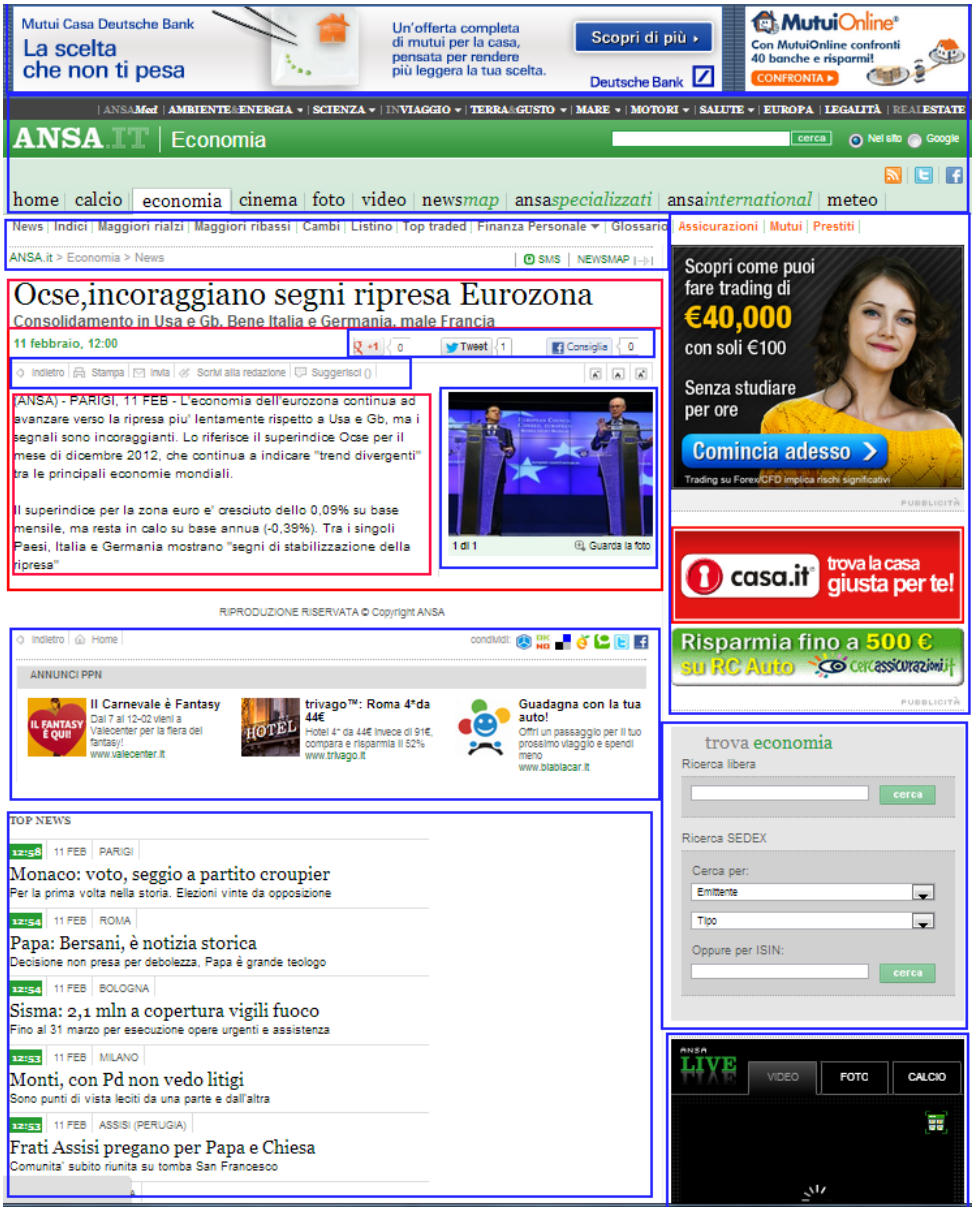

Figura 2.1. Pagina web di ansa.it, il template utilizzato è lo stesso per tutte le pagine dello stesso sito.

Lo strumento preposto alla ricerca ed esplorazione di documenti nel web, utilizzato nel nostro sistema di Social Business Intelligence, è il crawler iSyN-SC, integrato nella piattaforma semantica Synthema. Che cosa ci permette di fare questo potente strumento? Come abbiamo anticipato, il crawling è il processo di scansione delle fonti documentali finalizzato al caricamento dei contenuti testuali nel DB del server

di indicizzazione. La scansione delle fonti può essere pianificata e ripetuta a determinati intervalli di tempo in modo da aggiornare regolarmente l'indice e consentire la ricerca di contenuti variabili. Le fonti documentali sono repository da cui vengono prelevati i contenuti da indicizzare (documenti, e mail, pagine web,…). Possono essere interne (accesso diretto e permanente dal file system del server di indicizzazione) o esterne (accesso tramite collegamento su richiesta ad un altro server mediante un protocollo di comunicazione quale FTP, SMB, HTTP, POP3, IMAP, ecc.). Ciascuna fonte è associata a una specifica area. In questa sezione si descriveranno i passaggi più importanti nella parametrizzazione del crawler. Per iniziare la parametrizzazione è innanzitutto necessario definire le fonti documentali sulle quali eseguire il processo di scansione (crawling). Data la moltitudine di tipi diversi di documenti presenti nel web, il crawler prevede un comportamento diverso a seconda del tipo di fonte a cui vogliamo attingere. Con iSyN-SC è possibile gestire diversi tipi di fonti documentali tra cui : feed RSS, mailbox, database o siti web tramite websearch con motori di ricerca. I nostri sforzi si sono focalizzati principalmente sulle fonti documentali di tipo Web e Feed RSS (quest'ultime le più affidabili nel tempo). Brevemente illustriamo gli aspetti più importanti da considerare che differiscono per le due tipologie di fonti documentali. Per le fonti web, è necessario specificare i filtri sugli indirizzi url. Esistono due tipi di filtri: positivi e negativi. I filtri positivi hanno lo scopo di limitare la ricerca al sito Web senza estenderla a eventuali siti esterni collegati. I filtri negativi consentono invece di escludere sistematicamente determinati indirizzi Web. Inoltre è necessario indicare il livello di profondità che si vuole raggiungere con la ricerca seguendo i vari link compresi nella pagina. Si possono estrarre e classificare anche i documenti accessibili tramite i link presenti in quella pagina web, specificando il livello di profondità desiderato.

Per i feed RSS, indipendentemente dal tipo di contenuto selezionato, è possibile filtrare i documenti estratti in base a parole chiave. Il filtro ha effetto sia sui titoli che sugli abstract e come nel caso delle fonti web può essere positivo, includendo

tutti i documenti che presentano la chiave di ricerca, che negativo, escludendo i documenti. Un aspetto molto importante, comune a tutte e due le tipologie, è la possibilità di segmentare la pagina e prelevare la porzione di documenti desiderata indicando opportunamente il tag div di interesse. Riprendendo l'esempio di figura 2.1, le pagine degli articoli su ansa.it (ma vale solitamente per tutte le i siti web o feed rss) mantengono lo stesso layout, quindi è importante identificare le parti di testo rilevanti ed escludere il resto (ad esempio spazi pubblicitari e link social, si veda in proposito la figura 2.1).

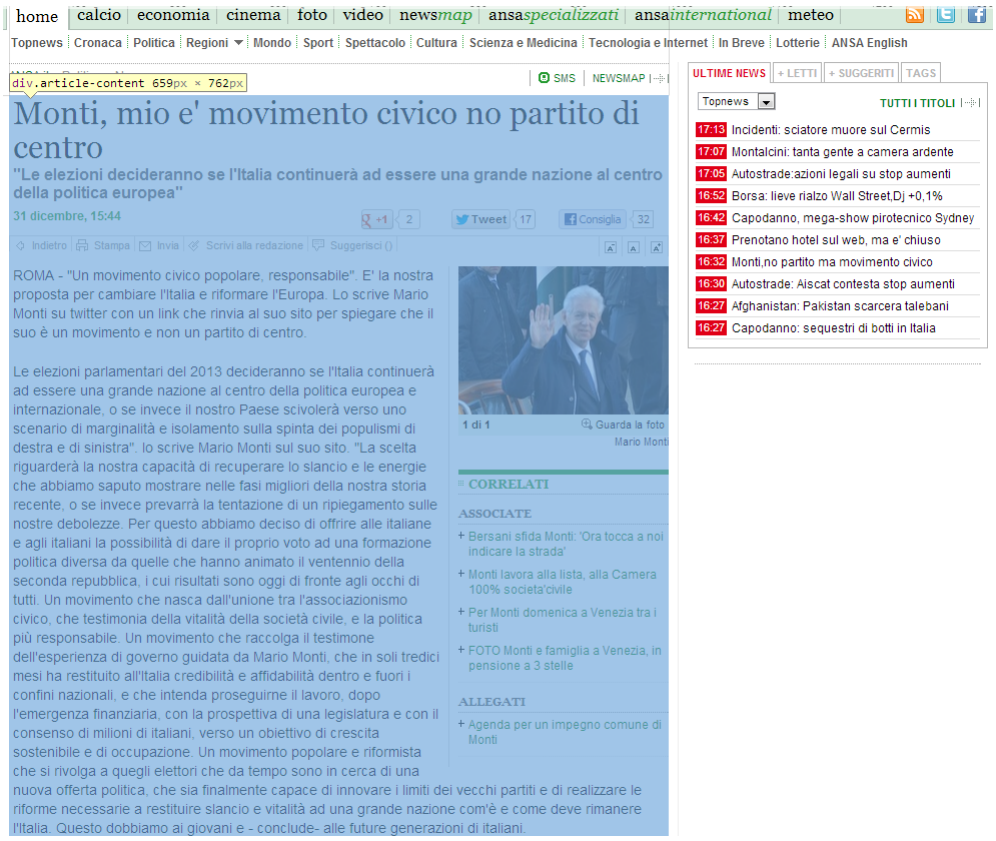

Figura 2.2 Articolo racchiuso in un determinato div

Questo non evita però di recuperare anche parti della pagina che possono non essere interessanti, come alcuni link funzionali o immagini (figura 2.3).

Figura 2.3 Alcune parti all'interno dell'articolo non sono utili.

Allo stesso modo, è possibile escludere questi oggetti non interessanti, individuando relativi i tag div.

Durante il processo di scansione della varie fonti documentali, abbiamo riscontrato alcune difficoltà nel reperimento di documenti dai social network come Facebook e Twitter. Non è stato possibile, infatti, recuperare in maniera automatica i messaggi pubblici degli utenti da questi due popolari network, in quanto dotati al loro interno di particolari tecniche "anti-scansione automatica" delle loro pagine web. Il motivo è facilmente intuibile, i messaggi degli utenti compongono proprio una parte del valore di questi due social network, non è quindi "gradito" che siano liberamente (e gratuitamente) accessibili per ricerche ed analisi. Anticipiamo al lettore, che questo non ha compromesso, in nessun modo, il test del sistema su documenti provenienti dai social network, poiché si è provveduto manualmente al recupero di un campione significativo. Nella metodologia prevista per la realizzazione di un sistema di Social Business Intelligence, la fase di configurazione del crawling viene ripetuta ciclicamente nel tempo. Come vedremo, a valle dell'elaborazione dei documenti recuperati, potrebbero manifestarsi entità con riferimento a topics inaspettati o che inizialmente non avevamo considerato meritevoli di far parte della nostra cerchia di keyword. Per questa ragione la fase di analisi e valutazione di nuove keyword, e di conseguenza della loro configurazione nel crawler, si ripete, sempre più raffinata ed aggiornata delle nuove tendenze. Quest'ultimo aspetto è molto marcato e degno di grande attenzione soprattutto per particolari domini di

ascolto, dove la dinamicità degli argomenti e delle tematiche di interesse varia molto nel tempo (come può essere l'ambito politico, le cui caratteristiche sono analizzate nel dettaglio nel capitolo 4).

2.2 Tecnologia: Text Mining

Oggi, il sempre più crescente volume di informazioni è di tipo non strutturato o semi strutturato. Questa grande abbondanza di informazioni pone il problema a molte organizzazioni, società o imprese di chiedersi: "Come possiamo immagazzinare, comprendere e utilizzare questa mole di dati?" (Witten 2012). Il Text Mining è un processo di analisi di insiemi di documenti testuali, mediante l'impiego di tecniche di Data Mining, per l'elaborazione di dati destrutturati (pagine web, e-mail, ecc.) e più in generale di qualsiasi corpus di documenti, allo scopo di:

- individuare i principali gruppi tematici
- classificare i documenti in categorie predefinite
- scoprire relazioni nascoste (legami tra argomenti, persone, ecc...)
- individuare informazioni specifiche (nomi di prodotti, persone, aziende, tematiche ecc…)
- estrarre concetti per la creazione e sviluppo di ontologie (ontology learning).

Il Text mining può essere genericamente definito come il processo automatizzato di analisi del testo con il fine di estrarre informazioni utili (Cunningham, Humphreys, Gaizauskas & Wilks 1997). Il testo è destrutturato, e non è facile da affrontare dal punto di vista algoritmico. Tuttavia, nella cultura moderna, il testo è il veicolo più comune per lo scambio formale (ma anche informale) di informazioni. I documenti testuali presenti nel web, contengono generalmente, la comunicazione di fatti (oggettivi) e/o opinioni (soggettivi) e l'opportunità di estrarre in maniera automatica queste informazioni appare fin da subito dotata di

grande potenziale, se pensiamo agli innumerevoli ambiti in cui si potrebbe sfruttare questa tecnologia.

Nel resto di questa sezione si discuterà il rapporto tra Text Mining e Data Mining, e tra il Text mining e la scienza che studia l'analisi del linguaggio naturale. Proprio come il Data Mining si pone come obiettivo la ricerca di modelli (pattern) nei dati, il text mining è in cerca di pattern nel testo. Tuttavia, anche se esiste una somiglianza di base ci sono anche alcune differenze. Il Data Mining può essere meglio rappresentato come l'estrazione di informazioni implicite, precedentemente sconosciute, e potenzialmente utili dai dati (Witten 2012). Per il Text Mining, invece, generalmente, le informazioni da estrarre sono chiaramente ed esplicitamente indicate nel testo, se consideriamo il punto di vista umano. Il problema di fondo, è dato dal fatto che le informazioni non sono formulate in un modo facilmente comprensibile da un elaboratore automatizzato. Il Text Mining si pone come obiettivo, rispetto al Data Mining, quello di estrarre dal testo una forma di dato, che sia successivamente adatta per una ulteriore elaborazione automatica, senza la necessità di un intermediario umano. Sebbene ci sia una netta differenza filosofica, dal punto di vista dell'elaboratore i problemi da affrontare sono molto simili. Il testo appare al sistema automatizzato, altrettanto criptico come lo sono i dati grezzi quando si mette in atto un processo di Data Mining (Witten 2012).

Come già accennato l'obiettivo comune è l'estrazione di informazioni "potenzialmente utili". Per il Data Mining il "potenzialmente utile" è dato dalla seguente interpretazione: il punto fondamentale per il successo è che le informazioni estratte, forniscano una chiave di interpretazione per aiutare a comprendere i dati. Nel Text Mining l'applicazione di questo criterio è meno immediata perché, a differenza del Data Mining, lo stesso input come spiegato in precedenza, è già comprensibile. Infatti per le tecniche di Data Mining le prestazioni possono essere misurate mediante l'utilizzo di metodi statistici che ci permettono, di conseguenza di confrontare differenti criteri di estrazione di dati sullo stesso problema. Anche la valutazione e il confronto delle prestazioni è meno

immediato per le tecniche di Text Mining (Witten 2012). Per riassumere un output "potenzialmente utile" di una elaborazione di Text mining equivale alla sintesi delle caratteristiche salienti di un corpo di testo.

L'elaborazione del linguaggio naturale, detta anche NLP (Natural Language Processing), è il processo, di analisi, automatico per mezzo di un calcolatore elettronico, delle informazioni scritte o parlate nel linguaggio umano o naturale (Cunningham, Humphreys, Gaizauskas & Wilks 1997). Questo processo è reso particolarmente complesso è laborioso a causa delle caratteristiche intrinseche del linguaggio umano, ricche di ambiguità. Per questo motivo il processo di elaborazione viene suddiviso in più fasi:

- *analisi lessicale*: scomposizione di un'espressione linguistica in token (in questo caso le parole)
- *analisi grammaticale*: associazione delle parti del discorso a ciascuna parola nel testo
- *analisi sintattica*: arrangiamento dei token in una struttura sintattica (ad albero: parse tree)
- *analisi semantica*: assegnazione di un significato (semantica) alla struttura sintattica e, di conseguenza, all'espressione linguistica.

Le tecniche di Text Mining sembrano condividere gli interi principi alla base dell'elaborazione automatica del linguaggio naturale. Ma, in realtà, anche se le tecniche di Text Mining utilizzano alcune tecniche di NLP la maggior parte degli sforzi profusi non sono orientati agli aspetti più profondi, cognitivi, della elaborazione del linguaggio naturale, bensì sono orientati agli aspetti più pratici e concreti, da questo punto di vista più vicini ai principi dell'Information Retrievel (Witten 2012). La ragione è da ricercare nelle dinamiche cha hanno portato nel tempo allo sviluppo di questa scienza chiamata Natural Language Processing. I primi studiosi, attorno agli anni '50 ipotizzarono che metodi basati su semplici traduzioni "parola per parola" potessero fornire buoni risultati e utili "traduzioni

grezze", successivamente raffinabili mediante tecniche basate su elementari analisi sintattiche.

L'unico risultato ottenuto da questi studi, è stato capire come le problematiche da affrontare fossero ben più complesse. Si è capito che il linguaggio naturale, anche quello di un bambino analfabeta, è un mezzo incredibilmente sofisticato ed eterogeneo che non è correttamente interpretabile da tecniche non altrettanto complesse. Questa eterogeneità è data, in sostanza, da ciò che consideriamo come conoscenza di "senso comune", che a causa della sua natura è estremamente difficile da codificare e utilizzare in forma algoritmica (Witten 2012). Le tecnologie alla base del Text Mining possono essere viste come è una conseguenza delle esperienze maturate nello studio delle tecniche di NLP, applicate in un contesto più delimitato con un approccio meno "estremo" e più vicino agli aspetti pratici e indirizzati al risultato finale.

2.3 Strumento: IBM SPSS Modeler Premium

Uno dei software più conosciuti nell'ambito del Text Mining (e non solo) è senza dubbio IBM SPSS Modeler Premium. Questo applicativo è un workbench di data mining completo che consente di sviluppare modelli predittivi, identificare pattern e trend in dati strutturati e non strutturati utilizzando un'unica interfaccia visuale supportata da una analitica avanzata. Permette inoltre di modellare i risultati e comprendere quali fattori li influenzano (IBM, 2011).

Principali caratteristiche dello strumento:

- Creazione di modelli predittivi avanzati.
- Estrazione di concetti, opinioni e relazioni chiave.
- Identificazione delle relazioni tra concetti, attitudini, persone, aziende ed eventi
- Compatibilità con una vasta gamma di piattaforme, database, fogli di calcolo e file di testo.
- Workbench di analitica dei testi (Text Link Analysis) completamente integrato per analizzare i testi di documenti, e-mail, blog, feed RSS e altre sorgenti di testo. Costituisce il cuore del motore di Text Mining, questo modulo è noto come IBM® SPSS® Text Analytics.

2.3.1 IBM SPSS Text Analytics

Questo modulo è in grado di offrire potenti mezzi di analisi del testo, sfruttando avanzate tecnologie linguistiche di Natural Language Processing (NLP) per processare efficientemente grandi varietà di dati testuali non strutturati, e da questi, estrarre concetti chiave ed individuare relazioni tra di essi.

Quasi l'80% di dati posseduti da una qualsiasi organizzazione sono in formato documento testuale, ad esempio report, pagine web, email ecc (IBM, 2011). Il testo è anche un fattore chiave per permettere ad una organizzazione di ottenere

informazioni importanti dai propri clienti. Un sistema che incorpora tecnologie NLP è in grado di estrarre in maniera intelligente concetti singoli e composti. Inoltre la conoscenza del linguaggio permette la classificazione dei termini nei relativi gruppi, come possono essere prodotti, organizzazioni, persone o più generalmente topic di interesse, utilizzando il significato e il contesto (IBM, 2011). I sistemi linguistici di questo strumento sono knowledge-sensitive, ciò significa che maggiori e migliori sono contenuti nei loro dizionari, più alto sarà il livello del risultato dell'analisi. SPSS Text Analytics è costituito da una serie di risorse linguistiche, come dizionari per termini e sinonimi, librerie e template. Questo strumento permette inoltre di sviluppare e raffinare queste risorse per adattarle al proprio contesto o dominio di ascolto. E' possibile quindi realizzare iterativamente la verticalizzazione del sistema, necessaria per migliorare l'accuratezza dei concetti recuperati e la categorizzazione di essi. L'approccio adottato da SPSS è sostanzialmente di tipo statistico, di fatto le tecniche utilizzate sono orientate al pattern recognition, e non è prevista una vera e propria analisi sintattica per capirne il significato. Sebbene includa, al suo interno, alcune tecnologie sintattiche possiamo ritenere SPSS un membro della famiglia dei metodi cosiddetti empirici. I metodi empirici infatti si basano su tecniche di "pattern matching", e sono in grado di rilevare e comprendere concetti e relazioni se la struttura della frase rispetta un pattern (un modello) noto. Tale caratteristica rende questi metodi ideali, ad esempio, per testi che non rispettano interamente le regole di una lingua (dove ad esempio un metodo prettamente sintattico andrebbe in grossa difficoltà). Questo aspetto rappresenta un grande vantaggio se pensiamo anche alle fonti di riferimento dell'input testuale, come i Social Network, dove spesso la correttezza della grammatica e della sintassi viene in secondo piano, rispetto all'estrema ricerca dell'immediatezza di un concetto espresso in pochi caratteri e con un gergo informale. L'idea che sta alla base a tutti i metodi di questa famiglia, è quella di definire non, un linguaggio pienamente corretto, bensì il linguaggio abitualmente utilizzato.

In questa sezione si illustreranno tutte le tappe che compongono l'elaborazione del testo mediante il workbench di analitica dei testi SPSS Text Analytics. Per il momento non siamo interessati a conoscere le origini di questi dati testuali, in quanto il tipo di sorgente o fonte, non ha nessuna influenza sul processo di Text Analisys. Anche se non sono esplicitamente dichiarati è possibile individuare tre passaggi fondamentali (figura 2.4) che compongono l'analisi del testo e sono:

1. Pre-analisi dell'input testuale
2. Identificazione dei concetti
3. Analisi delle relazioni tra concetti (TLA)

Alcune attività che compongono questi step non sono del tutto trasparenti all'utente, in quanto come vedremo non tutte le parti che compongono questo processo sono modificabili da parte dell'utilizzatore. Si è preferito evidenziare anche queste fasi "nascoste" per fornire una descrizione quanto più possibile completa dell'articolato processo di analisi del testo.

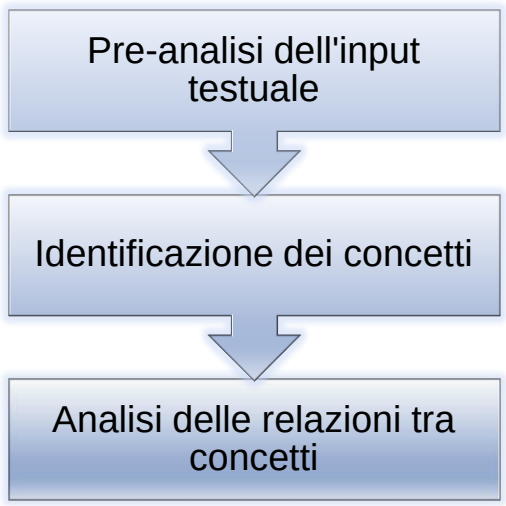

Figura. 2.4 Sequenza fasi dell'elaborazione testuale

La fase iniziale è importante per permettere al motore di analisi vero e proprio di elaborare il testo senza intoppi. Data la compatibilità, con una grande varietà di tipologie di input, che questo strumento mette a disposizione, la prima attività di pre-analisi serve per uniformare il testo in modo tale che l'analisi sia totalmente indipendente dal tipo di formato. Per il caso specifico, come vedremo in seguito, il nostro sistema anticipa questa attività nella fase di crawling dei documenti, che si pone a monte dell'elaborazione del testo, e si occupa di recuperare dal web i documenti testuali di interesse con particolari criteri che evidenzieremo successivamente. Una seconda attività di pre-analisi, sicuramente importante è la frammentazione del testo. Per frammentazione del testo si intende la separazione dei caratteri speciali presenti nel testo (come quelli di punteggiatura, spazi, virgolette ecc) dalle singole parole, in modo tale da implementare una prima disambiguazione. Lo strumento è in grado di distinguere anche i segni di punteggiatura utilizzati per le abbreviazioni, sigle, valute ecc classificando correttamente quindi quelle voci che li contengono. Questa fase è una di quelle non trasparenti all'utente, in altre parole lo strumento per come è stato realizzato, senza dare evidenza di ciò, si occupa automaticamente della pre-analisi senza offrire la possibilità di modifica di nessun aspetto. Nella fase successiva, si effettua l'estrazione dei concetti presenti nel testo. Lo strumento possiede una base di conoscenza che comprende una serie di risorse linguistiche che ci aiuteranno in questa difficile operazione. Le tipologie di risorse messe a disposizione sono due: le cosiddette risorse precompilate (preimpostate nel sistema fornendone una base di partenza) ed altre risorse che sono invece totalmente a carico dell'utente come l'inserimento di keyword e la tipizzazione. Le risorse precompialate costituiscono la base di conoscenza iniziale per avviare un processo di Text Analysis. Prima di esaminare nello specifico queste risorse occorre ricordare al lettore la natura dello strumento che si sta illustrando. SPSS Text Analytics è fondamentalmente uno strumento di analisi del testo con un approccio di tipo statistico (più orientato cioè alla ricerca di pattern nel testo che di vera comprensione di quest'ultimo). Quando si descrive uno approccio come non prettamente linguistico (come nel caso in

esame) non significa, tuttavia, che il sistema non si serva al suo interno di alcune tecnologie linguistiche (non evidenti all'utente) per effettuare l'analisi del testo. Le tecnologie linguistiche a cui si fa riferimento sono ad esempio l'analisi grammaticale e sintattica del testo. Mentre la prima è sostanzialmente indipendente dalla lingua (almeno in quelle di stampo anglosassone o latino) la seconda è molto influenzata dalla linguaggio utilizzato. Questo aspetto porta alla luce purtroppo alcune problematicità della base di conoscenza della lingua italiana, fornita da questo strumento. La risorse linguistiche di base per l'italiano non si sono dimostrare all'altezza della controparte inglese (data la natura anglosassone dello strumento è abbastanza comprensibile). Tuttavia, i problemi principali riscontrati hanno riguardato:

- La mancata lemmatizzazione delle parole, difficoltà quindi nel riportare ogni parola al lemma base (problema a cui si è dovuto ovviare autonomamente come si illustrerà in seguito)
- La mancata rilevazione delle forme verbali utili per l'analisi sintattica, che viene sostanzialmente nascosta all'utente, se non nella veste dei cosiddetti "schemi di estrazione" che verranno illustrati successivamente.

Nonostante ciò, è utile evidenziare la presenza di queste risorse precompilate, note anche come dizionari. Alcuni di essi, sono raffinabili da parte dell'utilizzatore (anche se occorre dire che in alcuni casi sono consigliate competenze specifiche di tipo linguistico), in altri invece non è possibile apportare nessuna variazione. I dizionari di cui si parla si dividono in:

- dizionari per l'estrazione di concetti,
- dizionari per le entità non linguistiche

I dizionari per l'estrazione sono composti da due parti: le part of speech e gli schemi di estrazione. Le part of speech non sono altro che la rappresentazione delle varie categorie della grammatica italiana come il nome, verbo, aggettivo e le altre forme grammaticali. Nella figura 2.5 si riportano nello specifico tutte le categorie

presenti. Si evidenzia il fatto, come anticipato, che alcune di esse non sono previste per il template italiano e quindi non sono riconosciute effettivamente nel testo.

```
#------------------------------------------------------------------------------#
# Part-of-speech codes used/usable for Italian Basic Resources
#------------------------------------------------------------------------------#


# a = adjective
# b = adverb (not used in the Italian Basic template)
# c = preposition ("dei", "dello", "com", ...)
# C = misspelling  (not used in the Italian Basic template)
# d = determiner ("el", "gli", "l", ...)
# f = first name
# i = middle initial in person name
# m = noun (n) or unknown (u)
# n = noun
# p = past participle ("suggestionato", ...)
# s = stop word (not used in the Italian Basic template)
# t = title (not used in the Italian Basic template)
# u = unknown
# v = verb (any verb)  (not used in the Italian Basic template)
# V = verb (infinitive)  (not used in the Italian Basic template)
# x = auxiliary verb ("siamo", "avete", ...) (not used in the Italian Basic template)
# y = particle ("van")
```

Figura 2.5 Part of speech per l'italiano

Per ovvie ragioni le part of speech non sono modificabili dall'utente. Gli schemi di estrazione invece servono per identificare grazie a dei modelli (o anche pattern) delle entità (composizione di termini) nel testo. Infatti il motore linguistico sottostante mediante l'uso di tecniche di Natural Language Processing (nascoste all'utente) può individuare concetti non noti (verrà illustrato in seguito il significato di noto) nella frase, ma solo se quest'ultimi rispettano una struttura compresa negli schemi di estrazione. Questi "pattern" sono utili per indicare al motore anche possibili entità composte (o cosiddette multi-word). Inoltre attraverso tali schermi è possibile identificare nomi propri, sia che riguardino persone, sia organizzazioni.

Gli schermi presenti nella base di conoscenza iniziale italiana sono i seguenti, per i nomi comuni:

Schema di estrazione	Esempio di entità
am	grande arteria
maa	sindrome respiratoria acuta
macma	soluzione pacifica del dramma iracheno
macm	quartier generale di Falluja
ma	spiriti deboli
mcmm	volo della singapore airlines
mcma	cura degli animali abbandonatti
mcmcm	spiegazione del movimento del cuore
mcm	festa della madonna

Tabella 2.1 Schemi di estrazione per estrazione concetti

Per i nomi propri di persone:

Schema di estrazione	Esempio di entità
ffm	John Edmund Doe
fim	John M. Doe
fym	John van Doe
fmu	John Smith Doe
fm	John Doe
ff	John Mary

Tabella 2.2 Schemi di estrazione per nomi propri

Come è possibile notare uno schema di estrazione può essere formato da più categorie grammaticali, formando semplici o complesse multi word. Gli schermi sono, teoricamente, modificabili, è quindi possibile aggiungerne di nuovi e/o toglierne qualcuno dei presenti. Per effettuare con successo questo tipo di

modifiche si rendono necessarie competenze linguistiche non banali, al fine di non intaccare le prestazioni del sistema finale si è preferito lasciare come pre-impostato dal sistema. Per quanto riguarda i dizionari delle cosiddette entità non linguistiche, essi sono molto utili per identificare quelle entità che non ricoprono un ruolo all'interno della frase come date, url web, indirizzi, numeri di telefono, valute, misure di ogni tipo e altre ancora. Per tale operazione sono definite nel sistema delle espressioni regolari avanzate per il riconoscimento di queste particolari entità. Data la quantità considerevole delle espressioni presenti nella base di conoscenza ne forniamo solamente alcuni esempi, in figura 2.6.

```
[PhoneNumber]
Reg_exp1=\(?[0-9]{3}\)?[ -][0-9]{3}[ -][0-9]{4}
      Es: (703) 740-2440, 866-854-2507
Reg_exp2=[+]?[0-9]{2}[ -][0-9]{3}[ -][0-9]{3}[ -][0-9]{4}
      Es: +44 208 288 4433
Reg_exp3=[+]?[0-9]{2}[ -]\(0\)[0-9]{4}[ -][0-9]{6}
      Es: +44 (0)7785 444650
Reg_exp4=[+]?[0-9][ .-][0-9]{3}[ .-]?[0-9]{3}[ .-]?[0-9]{4}
      Es: +1 3343233033 or +1.334.323.3033
[italian/PhoneNumber]
Reg_exp1=\(?[0-9]{3}\)?[ -][0-9]{3}[ -][0-9]{4}
            Es: (703) 740-2440, 866-854-2507
Reg_exp2=[+]?[0-9]{2}[ -][0-9]{3}[ -][0-9]{3}[ -][0-9]{4}
            Es: +44 208 288 4433
Reg_exp3=[+]?[0-9]{2}[ -]\(0\)[0-9]{4}[ -][0-9]{6}
            Es: +44 (0)7785 444650
Reg_exp4=[+]?[0-9][ .-][0-9]{3}[ .-]?[0-9]{3}[ .-]?[0-9]{4}
            Es: +1 3343233033 or +1.334.323.3033
```

Figura 2.6 Esempi di espressioni regolari, utili per individuare espressioni non linguistiche

Le keyword e i tipi fanno parte, invece, di quelle risorse a carico esclusivamente dell'utente e possiedono un ruolo fondamentale negli interventi di verticalizzazione del sistema per l'ambito di ascolto. Dovremo obbligatoriamente agire su queste risorse per migliorare le performance dell'analisi del testo. L'inserimento delle keyword e dei tipi riguardano sostanzialmente la parte di conoscenza già citata come knowledge-sensitive. E' con la tipizzazione e inserimento di keyword che si inizia a parlare di analisi semantica del motore. Aggiungendo tipi di concetti di interesse e keyword al loro interno, con criteri ben ponderati, è possibile migliorare l'interpretazione semantica del motore, che sarà così in grado di associare ai concetti di interesse estratti il tipo a cui fanno riferimento, fornendo in questo modo il loro significato. Ogni keyword deve appartenere ad almeno un tipo. Possiamo inserire la stessa keyword in più tipi, ma si dovrà risolvere il conflitto determinandone un unico riferimento. Per raggruppare ulteriormente le categorie di tipi è possibile definire delle librerie. La struttura gerarchica di queste risorse sarà quindi la seguente:

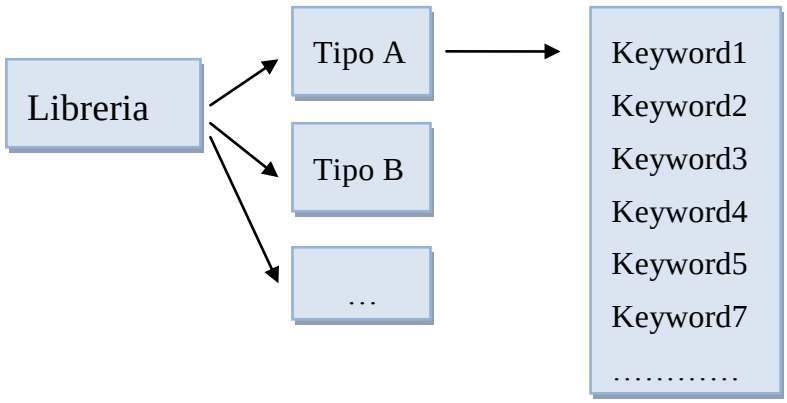

Figura 2.7 Struttura delle risorse per la tipizzazione

Le modifiche a queste risorse si applicano attraverso il pannello "Risorse Libreria" dell'interfaccia grafica del Modeler di Text Analytics (figura 2.8).

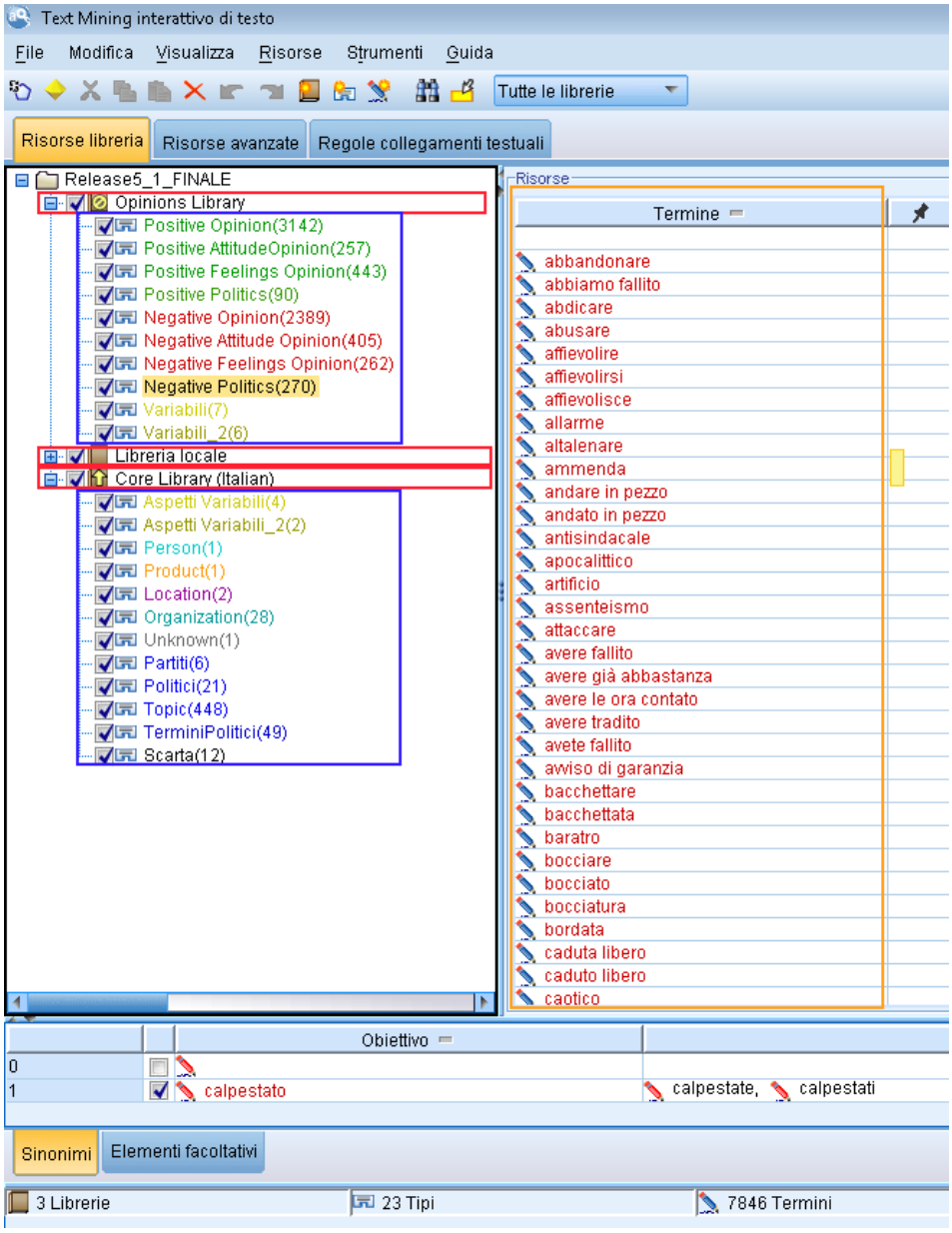

Figura 2.8 Rappresentazione pannello di inserimento keyword (riquadro arancione), tipi (riquadro blu) e librerie (riquadro rosso)

La determinazione dei tipi, e la giusta tipizzazione dipende chiaramente dell'obiettivo che ci si pone per la fase di analisi del testo. Per un progetto di sentiment analysis, in cui si ha come obiettivo quello di rilevare opinioni su

determinati argomenti espressi nel testo, è bene definire innanzitutto almeno due librerie distinte, una per i qualificatori e una per gli argomenti di interesse.

- La libreria dei qualificatori conterrà tutti le keyword (raggruppate in tipi) che per il caso di studio, svolgono la funzione di opinione (polarizzata positivamente o negativamente). Parliamo quindi di aggettivi, ma anche di composizioni di parole che riconosciamo essere utilizzate nel gergo comune o specifico, per esprimere un giudizio su un determinato concetto.
- La libreria degli argomenti di interesse riflette, a maggior ragione, la tassonomia dell'ambito di ascolto che ci interessa catturare. Per non perdere di generalità, è buona norma definire una serie di tipi che includano keyword di interesse per l'analisi, i cosiddetti "topic".

Per permettere una maggiore organizzazione delle nostre librerie è possibile definire per ogni keyword inserita dei sinonimi. In questo modo avremo un termine di riferimento e tanti altri saranno ricondotti a quella keyword. L'utilizzo o meno di questa possibilità dipende anche dalla granularità che preferiamo per la nostra analisi e soprattutto se desideriamo o meno avere questo livello di aggregazione già in questa fase.

Dopo l'estrazione dei concetti è possibile individuare le relazioni che esistono tra di essi. Questa attività viene denominata anche Text Link Analysis o più brevemente con TLA. E' il passo fondamentale per una completa analisi del testo. Senza di essa avremmo solo una serie di concetti rilevati nel testo a seguito di certe logiche di linguistica e semantica, ma non saremmo di certo in grado di associare delle opinioni espresse ad un particolare concetto presente. La situazione in cui ci troviamo prima di avviare l'analisi delle relazioni, può essere rappresentata dall'esempio in figura 2.9.

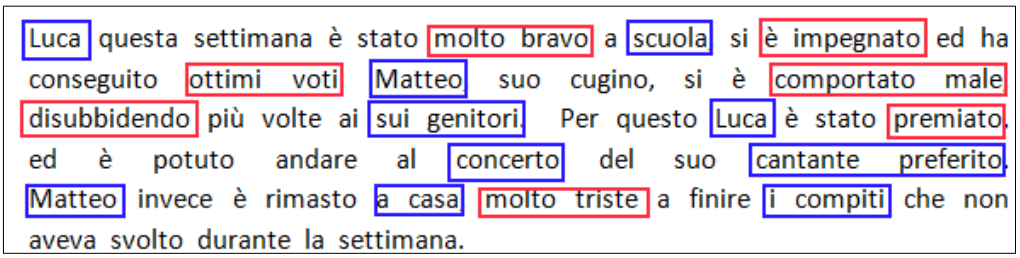

Figura 2.9 Rappresentazione di un possibile output della fase di estrazione concetti

Grazie alle keyword inserite e alle risorse linguistiche precompilate sono stati rilevati una serie di concetti all'interno del testo. Sempre in riferimento alla figura 2.9 in blu sono evidenziati i concetti estratti e, ipotizziamo, facenti parte delle keyword della libreria dei topic, mentre in rosso quelli della libreria dei qualificatori. Chiaramente la bontà dell'estrazione dei concetti dipende dalla qualità delle keyword e della tipizzazione progettate nella fase precedente. Quello che rimane ora, è associare tra loro i vari concetti, in modo tale da collegare i concetti di tipo topic con i qualificatori che ne esprimono un'opinione, o quanto meno nel nostro caso una attribuzione positiva, negativa o neutra. Per fare ciò SPSS Text Analitycs, mette a disposizione un pannello denominato "Text Link Rules" (in figura 2.10) in cui poter definire appunto le regole per la creazione di strutture (o pattern) di frase che possano individuare le relazioni tra i concetti (IBM, 2011). Vista in questo modo potrebbe sembrare un'operazione molto complicata, in realtà il sistema, di per se, è molto semplice e segue logiche di tipo pattern matching per confrontare la struttura della frase con quelle presenti nelle varie regole progettate.

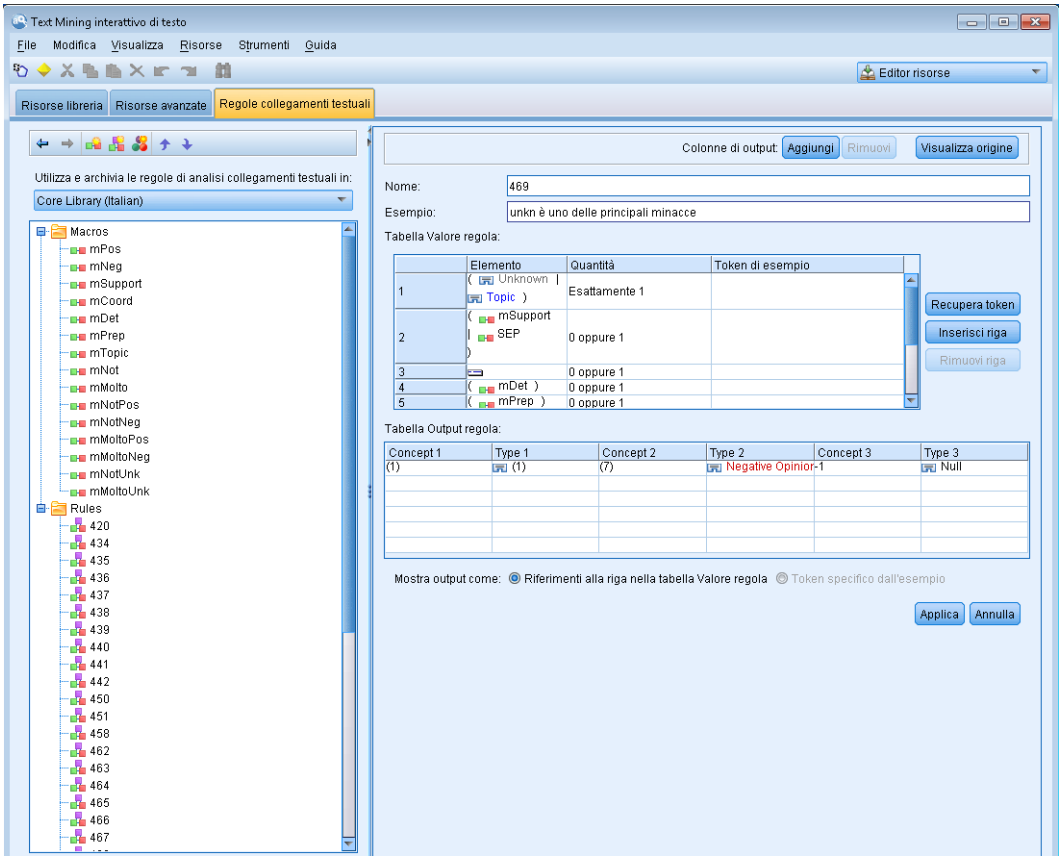

Figura 2.10 Interfaccia grafica del pannello di modifica TLA

Per definire una regola in grado di catturare pattern significativi nel testo, SPSS Text Analytics mette a disposizione una serie di strumenti. Ad esempio le regole possono includere al loro interno macro, tipi e spazi. Gli spazi sono considerati come una serie di parole di intervallo, non classificate in nessun tipo o in nessuna macro. Le macro invece, fungono da contenitori, nei quali possiamo mettere sia liste di termini, sia dei tipi di keyword (definiti in precedenza) formando delle vere e proprie composizioni di insiemi di parole. Senza anticipare nulla al lettore sulle varie modifiche ed integrazioni introdotte nelle operazioni di verticalizzazione del motore (che tratteremo nel capitolo 6), vediamo come sono fatte le macro, osservando alcuni esempi presenti nel nostro sistema in figura 2.11.

```
name=mPos
value = ($Positive Opinion|$Positive AttitudeOpinion|
$Positive Feelings Opinion|$Positive Politics|$Positive
Social Politics)

name=mNot
value=(nn|poco|poca|poche|pochi|non|mai|nessun|nessu
a|meno|priva di|senza|scarso|scarsa|scarsi|scarse|no
|bassa|andrebbe|andrebbero|nulla)

 name=mNotPos
value=(perdere|nn|poco|poca|poche|pochi|non|mai|nessun|
nessuna|meno|priva di|senza| scarso| scarsa|scarsi|
scarse|no| bassa|andrebbe|andrebbero|dovrebbero
essere|dovrebbe essere|nulla di|limiti del|limiti
della|limite della|limite del|niente|non essere|non
abbastanza) $mDet? @{0,2} ($mPos)
```

Figura 2.11 Esempi di macro, utili per la costruzione delle regole

La prima macro in figura, *mPos* (modificabile anche da interfaccia grafica come da figura 2.12), funge da agglomerato per tutte quelle categorie di qualificatori positivi (descritti in precedenza). La seconda macro *mNot*, invece, viene utilizzata invece come agglomerato di termini che usualmente ricoprono il significato di negazione all'interno della frase. Mentre la terza macro, *mNotPos*, presenta una particolarità ulteriore. La sua funzione è quella di contenere tutti i termini che negano un aspetto positivo. Se osserviamo infatti al suo interno include due ulteriori macro: *mPos* (già vista in precedenza) e *mDet* che include gli articoli determinativi, indeterminativi e pronomi. La codifica della macro ripropone la modalità di composizione delle regole che vedremo in seguito.

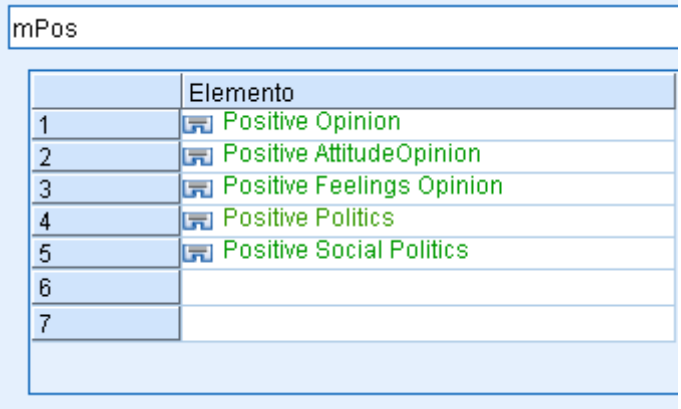

Figura 2.12 Esempio dell'interfaccia grafica della modifica di una macro

In tal modo, si possono elaborare regole complesse ma allo stesso tempo comprensibili. Si tratta chiaramente di un vantaggio importante per rendere facilmente manutenibile il nostro sistema nel corso del tempo, a fronte delle continue modifiche che si apporteranno.

Per progettare delle regole che diano risultati soddisfacenti, è bene pensare prima all'obiettivo che si vuole raggiungere. Per individuare una relazione tra concetti, è necessario avere regole contenenti pattern che combacino esattamente con la struttura della frase. Ora, come vedremo i pattern che si possono creare mediante le regole possiedono una limitata dose di flessibilità, infatti l'enorme varietà del gergo umano e la molteplicità dei modi che esistono per esprimere uno stesso pensiero fanno si che l'attività di progettazione delle regole sia molto laboriosa e richieda tempo. Per comporre una regola possiamo combinare a nostro piacimento macro, tipi ed i intervalli di parole. Le possibilità sono praticamente infinite, dobbiamo pertanto legare questi meccanismi, in modo tale da identificare strutture di frasi ricorrenti. Il nostro obiettivo è indirizzato alla sentiment analysis, cioè alla ricerca di opinioni (polarizzate e non) collegate a topic di interesse (un argomento, una persona, un organizzazione, un prodotto ecc). Tutte le regole che sono state implementate sono state progettate seguendo questo obiettivo.

La descrizione delle regole implementate per il nostro progetto sarà dibattuta insieme agli altri aspetti specifici al caso di studio nel capitolo 6, perciò in questa sezione ci limiteremo a fornire un esempio di regola e del suo funzionamento di base. Si ipotizzi di analizzare la frase:

"L'iPhone è uno smartphone valido ma molto costoso"

Supponiamo di avere nei nostri dizionari le keyword *"iPhone"* (di tipo Topic), *"valido"* e "*costoso*" (probabilmente una del tipo positivo e l'altra del tipo negativo). L'estrazione dei concetti riuscirà a rilevare i concetti per cui avremo:

Oggetto	**Tipo**	**Macro**
L'	-	$mDet
iPhone	$Product	$mTopic
è	-	$mSupport
uno	-	$mDet
smartphone	$Unknown	-
valido	$Positive	-
ma	-	$mCoord
molto	-	$mMoltoNeg
costoso	-	-

Tabella 2.3 Esempio di estrazione di concetti

Per catturare tutte le opinioni espresse in questa frase e legarle al concetto giusto dovremmo avere una regola del tipo:

```
mTopic + mSupport + @(0,2) + $mPos + $mCoord +
($mMoltoNeg| $mNeg)
```

Più precisamente la regola sarà così strutturata:

Ind.	**Macro**	**Lunghezza**
1	mTopic	Esattamente 1
2	mSupport	Esattamente 1
3	*"Intervallo di parole"*	Tra 0 e 2
4	mPos	Esattamente 1
5	mCoord	Esattamente 1
6	($mMoltoNeg | $mNeg)	Esattamente 1

Tabella 2.4 Esempio di struttura di una regola

E nella tabella output dovremo mettere:

Concetto1	**Tipo1**	**Concetto2**	**Tipo2**
(1)	(1)	(4)	OPINIONE POSITIVA
(1)	(1)	(6)	OPINIONE NEGATIVA

Tabella 2.5 Esempio di output di una regola

In questo modo avremmo associato un'opinione positiva ad iPhone ("valido") e un'opinione negativa ("molto costoso"). Con una regola così formulata, d'ora in avanti, saremo in grado di rilevare le relazioni tra concetti che presentano questo tipo di struttura. A seconda di come e di quali regole progettiamo, un pattern presente nella frase potrebbe essere compatibile potenzialmente con più di una regola. L'ordine di esecuzione dipende da come che abbiamo definito la disposizione delle regole nella lista. Quelle più in alto avranno la precedenza su quelle più in basso. In sostanza la prima che sarà compatibile (farà match con il pattern) prevarrà sulle altre. E' buona norma per questo motivo, mettere nelle posizioni più alte della lista regole con pattern molto complessi e nelle parti più basse quelli con pattern meno complessi. Questo ci dà la possibilità di avere

maggiore robustezza, evitando che pattern semplici possano inibire la rilevazione di pattern più complessi.

Capitolo 3

Architettura e implementazione del sistema

Nei precedenti capitoli abbiamo presentato lo scenario della Social Business Intelligence, e l'insieme di tecnologie e strumenti adottati per la realizzazione del sistema. In questo capitolo si analizzerà dettagliatamente l'architettura di un sistema di Social BI nel suo complesso e l'implementazione, in tutti i suoi aspetti, del prototipo realizzato.

3.1 Architettura funzionale

Per un sistema di Social BI, i social media, rappresentano la principale fonte da cui attingere, per il reperimento dei dati di input. Questi dati, provenienti dal mondo web, (per lo più di tipo destrutturato) devono convergere insieme ai dati (di tipo strutturato) provenienti dai sistemi di BI già presenti in azienda, in un processo di memorizzazione finalizzato al popolamento del database operazionale. Dai dati grezzi, avviando un processo di analisi del testo, estraiamo i dati elaborati, contenenti al loro interno valore informativo non ancora completamente espresso. Il database alimenterà i sistemi di sintesi, come un data warehouse dedicato al social e/o il dw aziendale, e attraverso tool Olap potremo rendere disponibili informazioni utili per il "decision making" a diversi livelli aziendali e per diversi

settori e processi dell'azienda. L'architettura mostrata nella figura 3.1 può rappresentare fedelmente il sistema nel suo complesso, così come è stato descritto.

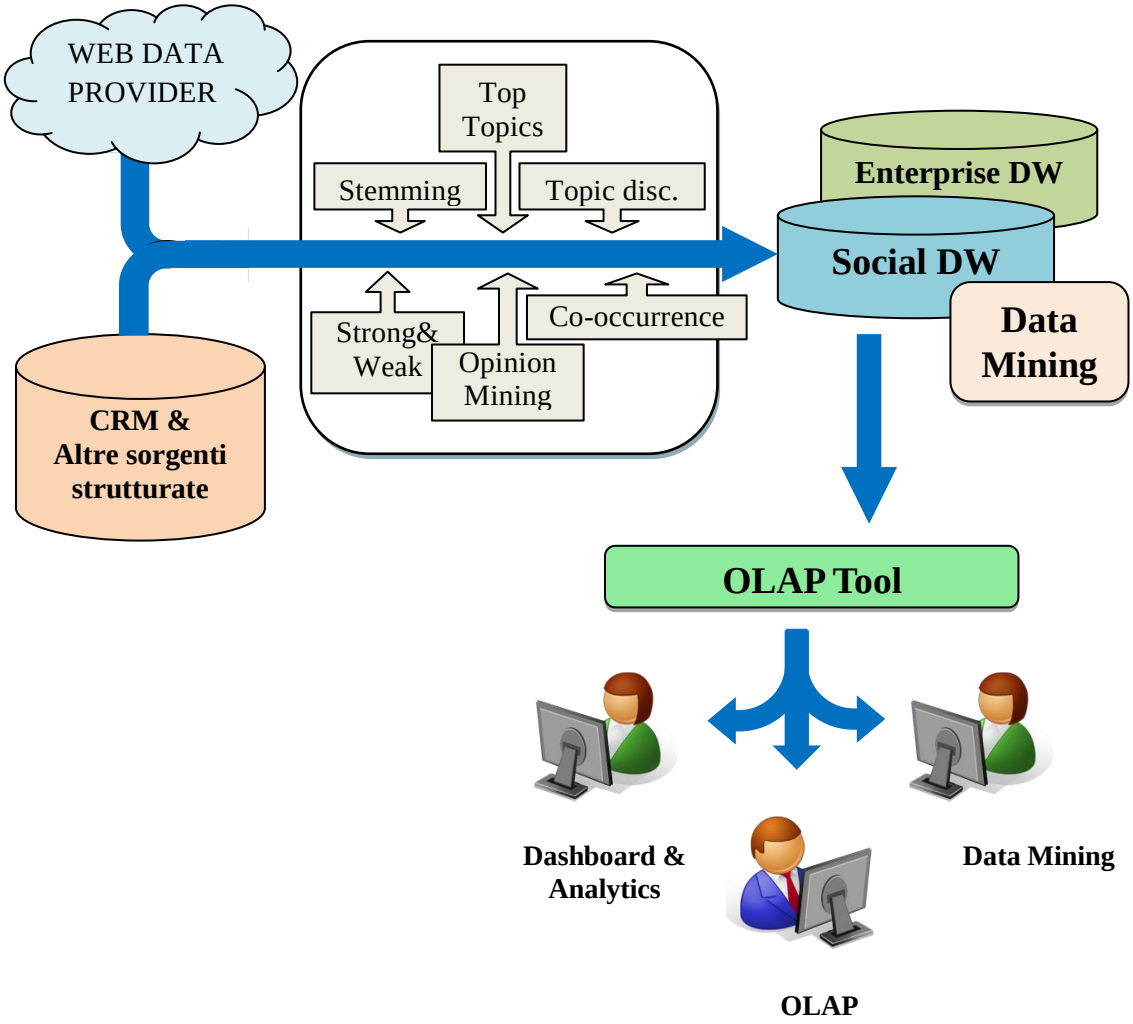

Figura 3.1 Architettura di un sistema di Social BI nel suo complesso.

L'architettura del sistema di SBI, è così ideata nel suo complesso, per mettere a disposizione dell'utente finale una serie di macro funzionalità, che possiamo riassumere in:

- *Conteggio*: identifica la capacità di tenere traccia di tutte le occorrenze di un termine o di un concetto di interesse (dove per concetto si intende un argomento al quale si può fare riferimento utilizzando parole diverse).
- *Co-Occorrenza*: fa riferimento all'individuazione delle occorrenze nelle quali due termini distinti compaiono all'interno della stessa frase. Identificano un legame tra questi due termini e si chiamano per questo co-occorrenze.
- *Topic Discovery*, ossia la catalogazione di documenti estratti dal web a seconda degli argomenti trattati ed ai concetti contenuti in essi. Generalmente è richiesto l'utilizzo di tecniche di clustering.
- *Opinion Mining*: (o sentiment analysis) è la capacità di riconoscere opinioni all'interno di un testo, attraverso l'individuazione di frasi e termini aventi una accezione polarizzata (negativa o positiva).

Inoltre, dalle sopraccitate funzionalità possiamo ottenere altre funzioni di dettaglio, che forniscono un'elaborazione più specifica del dato, precisamente:

- Dal conteggio possiamo ottenere le seguenti sotto-funzionalità:
    - *Conteggio* delle occorrenze dei termini o di un insieme di *topic* noti. Per poter funzionare efficacemente richiede un'operazione di "normalizzazione semantica" in cui tutti i termini usati per identificare uno specifico topic vengono riconosciuti come equivalenti al topic stesso potendo così, conteggiare le reali occorrenze. Ad esempio, nell'ambito politico, ci si potrebbe riferire a Silvio Berlusconi dicendo "Berlusconi", "Silvio", "il Cavaliere" o usando altri appellativi. In tal modo l'informazione estratta, può

essere ulteriormente elaborata per evidenziarne specifici aspetti come:

- i *topic più discussi* (top topics) in un certo intervallo temporale definito in precedenza.
- i *nuovi topic* (new topics) argomenti menzionati nei nuovi testi recuperati, che non erano stati mai trattati in precedenza.
- i *topic di tendenza*, cioè quei topic che, in particolari momenti storici, diventano *trendy* ovvero particolarmente discussi e dibattuti (potrebbero essere argomenti del tutto nuovi o argomenti che non avevano mai attirato l'attenzione del pubblico, come in quel momento).

- L'utilizzo della funzione di co-occorrenza permette di riconoscere un legame, sintattico, semantico o più semplicemente di vicinanza, tra due concetti. Questo risultato ci può guidare nel comprendere cosa si dice di (o più obiettivamente, cosa viene legato a) un determinato concetto di interesse.

- Nella macro categoria dell'Opinion mining sono state determinate le seguenti funzionalità:
  - *Opinion Mining di base* consente l'individuazione nel testo di opinioni (positive o negative) legate ad un concetto, potendo discriminare eventualmente sulla base della sorgente dell'informazione o sul periodo temporale o su un qualsiasi altro attributo appartenente al testo.
  - *Opinion Mining l*egato con la *Co-Occorrenza* ***p***ermette di riconoscere i punti deboli e punti di forza di un singolo concetto rilevando tutte le sue relazioni (sintattiche e semantiche) presenti nei testi. In questo caso l'opinione su un concetto è differenziata rispetto al legame con altri concetti.

L’ultima funzionalità citata nell’elenco fa riferimento ad un tipo particolare di elaborazione che ha come obiettivo quello di “condensare” in un solo report tutto il valore informativo estratto dai testi, relativamente ad un topic di interesse. Per questo scopo il processo di cui ci si serve viene denominato “riduzione di dimensionalità” e consente di raffigurare su un piano cartesiano (bi-dimensionale) tutti i metadati (dimensioni) dei concetti individuati, in relazione tra loro. Sarà quindi possibile comprendere in una sola dashboard la distanza tra i topic rappresentati, dalla sintesi delle loro caratteristiche rilevate (rilevanza, numero di occorrenze, polarizzazione dell’opinione ecc).

Il fulcro di questa macro-architettura (figura 3.1) è senza dubbio la parte di manipolazione dei dati in input, che prima di essere immagazzinati nel database operazionale in forma strutturata, subiscono un processo di elaborazione finalizzato all’elargizione delle funzionalità sopradescritte. Per il progetto di tesi, è stata progettata e realizzata l’architettura funzionale (figura 3.2) che prevede l’impiego di diverse tecnologie e strumenti. In primo luogo, si osservi il reperimento dei dati dal web, attraverso il crawler Synthema (analizzato nel capitolo 2) e il loro l’immagazzinamento in un database operazionale, appositamente implementato. Successivamente l’architettura prevede l’elaborazione dei dati grezzi mediante un motore di analisi del testo, basato sullo strumento SPSS Text Analytics (approfondito nel capitolo 2), e la conseguente memorizzazione dei dati ottenuti.

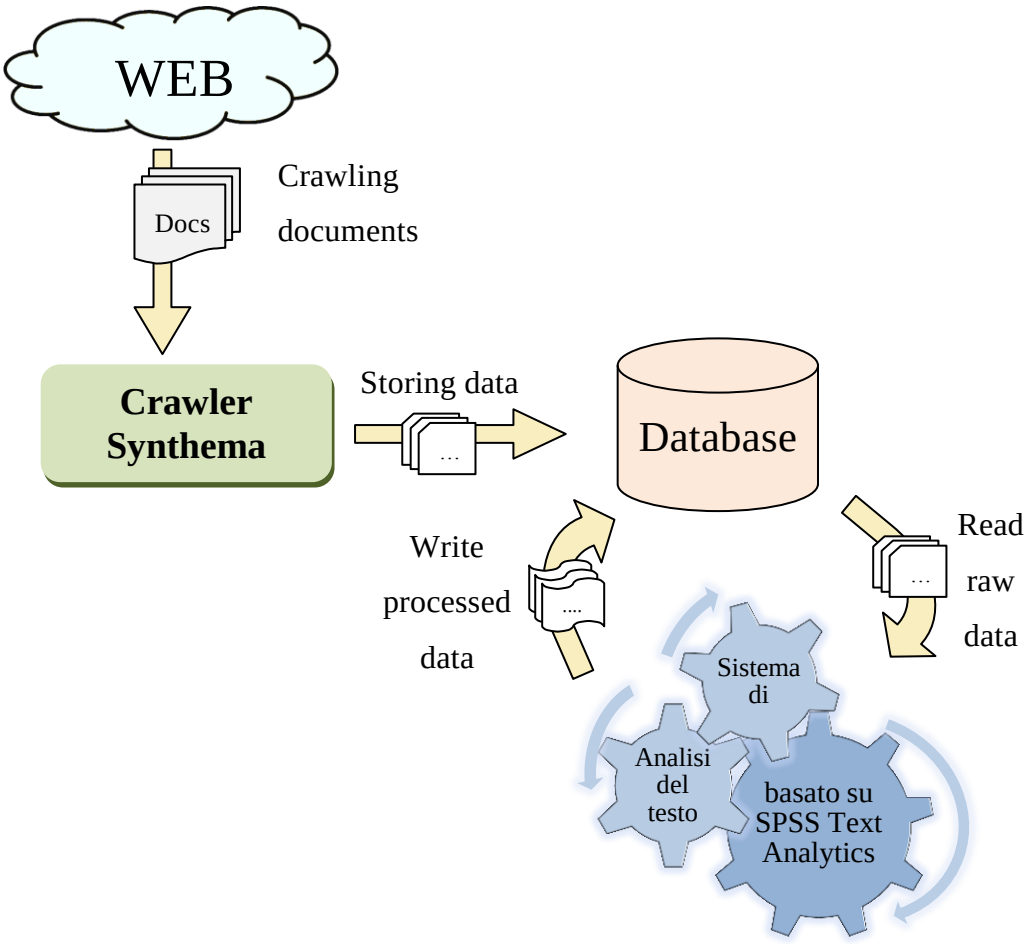

Figura 3.2 Architettura funzionale realizzata per il sistema di Social BI

L'architettura funzionale così rappresentata è determinata, come possiamo osservare, da quattro flussi principali:

- *Crawling dei documenti*: flusso che permette di recuperare (nelle modalità descritte nel capitolo precedente) i documenti dal mondo web attraverso un crawler (nel nostro caso, integrato nella piattaforma Synthema)
- *Immagazzinamento dei dati*: flusso che effettua la memorizzazione attraverso appositi job (che approfondiremo in seguito) i testi prelevati dal crawler, nel database operazionale.

- *Lettura dei dati grezzi dal database*: flusso che ci permette di leggere i dati grezzi memorizzati nel database e di uniformarli all'input richiesto dal sistema di elaborazione del testo (flusso compreso nell'implementazione di quest'ultimo che approfondiremo in seguito).
- *Scrittura dei dati processati*: flusso che permette di riscrivere sul database operazionale, i dati elaborati (concetti, relazioni, concetti e relazioni polarizzate) da parte del sistema di Text Analytics (anch'esso compreso nell'implementazione di quest'ultimo, che approfondiremo in seguito).

Il motore di elaborazione del testo, si basa sulle tecnologie di analisi dello strumento SPSS Text Analytics. Analizzando in dettaglio questa parte importante dell'architettura, è possibile identificare un processo di elaborazione che, per quanto non sempre evidente in tutte le sue parti, è definito da una sequenza di attività, di sotto elaborazione.

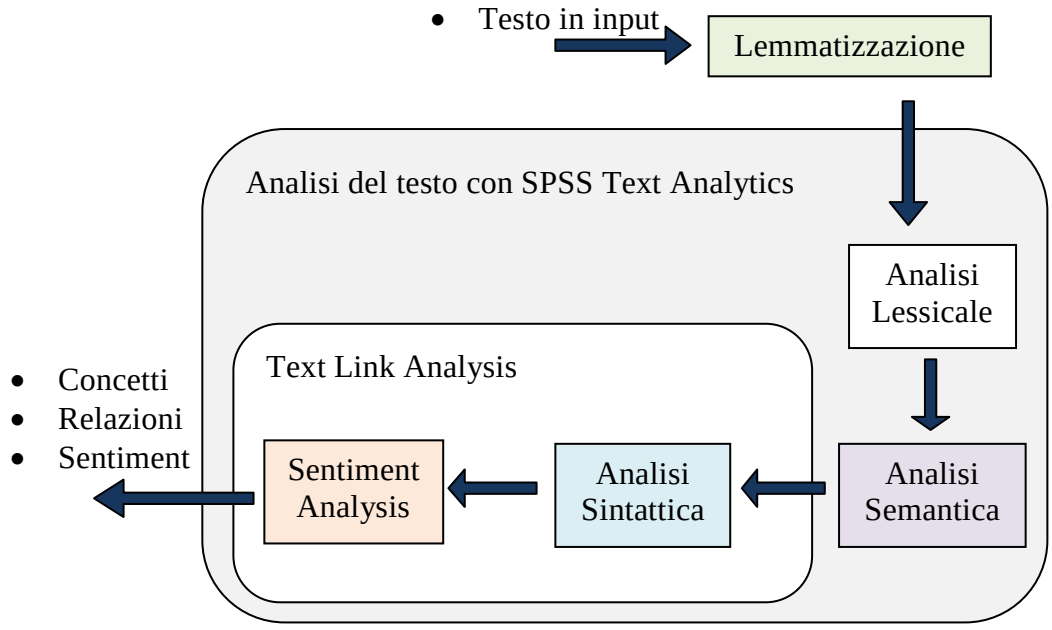

Figura 3.3 Sistema di elaborazione del testo

Il sistema di elaborazione del testo, rappresentato in figura 3.3, è costituito da una serie di attività di analisi, non tutte relative al motore basato su SPSS Text Analytics. La lemmatizzazione, di fondamentale importanza per un processo di analisi del testo, in riferimento alle problematiche riportate nel capitolo due, è stata implementata attraverso una procedura esterna mediante l'impiego di un dizionario di termini e lemmi corrispondenti. Tralasciando per il momento questo aspetto (che tratteremo nel paragrafo relativo all'implementazione), esaminiamo brevemente le attività svolte dal motore:

- *analisi lessicale*: individua i concetti presenti nelle librerie dei termini, e quelli cosiddetti "unknown" che rispettano le strutture determinate dagli schemi di estrazione.
- *analisi semantica*: si attribuisce il significato ai concetti individuati nella fase precedente, considerando la tipizzazione formulata nelle librerie.
- *analisi sintattica*: si individuano le relazioni tra i concetti attraverso la determinazione di pattern di TLA (Text Link Analysis).
- *sentiment analysis*: con la polarizzazione dei concetti (data dall'analisi semantica) e le relazioni identificate nell'attività precedente otteniamo l'analisi del sentiment.

3.2 Implementazione

Analizzata l'architettura prevista per la realizzazione del sistema di Social Business Intelligence, passiamo a descrivere tutti i dettagli relativi all'implementazione. La prima parte sarà dedicata alla descrizione dell'implementazione della base dati, poi verrà illustrata l'implementazione dei processi di lettura e scrittura del database ed infine la creazione delle regole, che costituiscono il cuore del motore di analisi del testo.

In primo luogo, osserviamo l'implementazione del database, che fa da supporto per l'intera architettura del sistema di Social BI (figura 3.2). La basi dati è stata progettata per supportare appieno le funzionalità descritte nel paragrafo precedente, fungendo da contenitore sia per i dati grezzi di input (semplici testi), che per i dati di output del processo di elaborazione, contenenti l'intero valore informativo presente nell'input testuale.

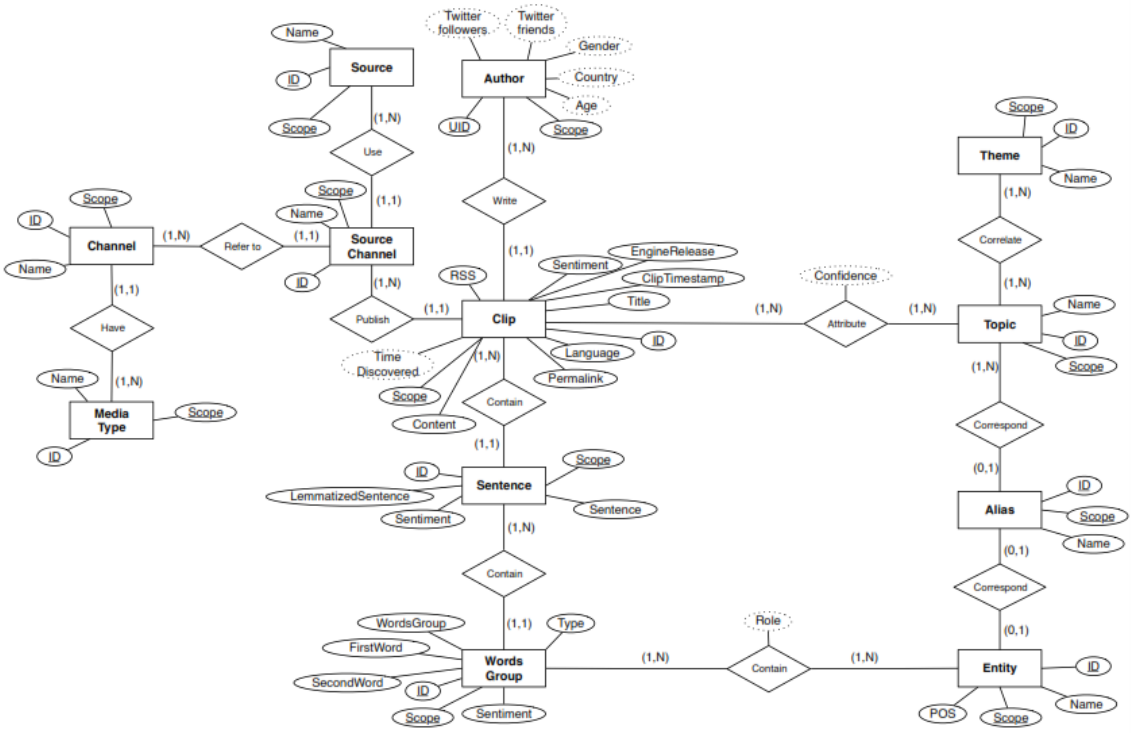

Figura 3.4. Schema ER del database operazionale

Lo schema ER prevede una serie di entità, che descriviamo brevemente:

- Clip: sono i documenti testuali recuperati nella fase di crawling. Presenta tutta una serie di attributi (language,title ecc) relativi al documento recuperato.
- Sentence: sono le frasi segmentate delle clip. Ogni sentence è individuata dalla clip a cui appartiene.
- WordsGroup: sono i gruppi di parole (concetti) legati da una relazione (con *sentiment* positivo, negativo o neutro) individuati dal motore di analisi del testo, in una sentence. I concetti sono *firstword* e *secondword* e il tipo (*type*) di relazione è identificata dalla regola che ne ha permesso l'individuazione.
- Entity: le entità sono i concetti rilevati nei wordsgroup. Data la formulazione delle regole di SPSS (che prevedono solamente relazioni a due concetti) nel nostro caso possono riferirsi al *firstword* o *secondword*.
- Alias: comprendono tutti i termini e sinonimi con cui si fa riferimento ad un determinato topic (soprannomi, sigle, appellativi). Vengono istanziati staticamente.
- Topic: sono gli argomenti di interesse per l'analisi. Rappresentano il candidato di riferimento degli alias. Vengono istanziati staticamente.
- Theme: sono i temi di interesse per l'analisi.
- Author: contiene le informazioni sugli autori delle clip recuperate.
- Source: contiene le informazioni sulla sorgente da cui si è recuperata la clip.
- Channel: contiene le informazioni sul canale da cui si è recuperata la clip.
- MediaType: contiene le informazioni dei diversi tipi di social media.

e due associazioni di tipo molti a molti, che sono:

- Contain (WordsGroup,Entity): contiene tutte le associazioni in cui una *entity* compare in un *wordsgroup*.
- Attribute(Clip, Topic): contiene tutte le associazioni in cui un *topic* compare in una *clip*.

La descrizione dell'implementazione dell'architettura prosegue analizzando più dettagliatamente i flussi descritti nel paragrafo precedente. Una volta recuperati i documenti dal web mediante il crawler Synthema, i documenti rimangono archiviati nella piattaforma di Synthema, in attesa di essere trasferiti nel nostro database operazionale. L'immagazzinamento dei dati nel nostro db (flusso "storing data" di figura 3.2) avviene mediante job realizzati tramite il potente software open source Talend, che fornisce tools per l'integrazione e gestione dei dati, e per l'interfacciamento delle applicazioni. Grazie al linguaggio multipiattaforma Java, con cui è stato realizzato, Talend permette la compatibilità tra diverse applicazioni e sistemi coesistenti in un'architettura complessa (Talend, 2006). L'interfacciamento con la piattaforma Synthema avviene mediante un web service messo a disposizione appositamente per il reperimento automatico dei documenti recuperati dal crawler. Il job di talend (raffigurato in figura 3.5) si occupa di effettuare periodicamente una chiamata al server (nodo *tHttpRequest_1*) di Synthema, e recuperare i testi (clips) strutturati in file xml (nodo *tFileInputXML_1*) comprensivi di una serie di metadati, come ad esempio la data di reperimento, l'autore (se presente), la fonte (se identificabile), il titolo della clip e altri ancora. Inoltre la piattaforma fornisce per ogni clip anche la suddivisione in sentences (frasi). Successivamente il job si occupa di scrivere nelle apposite tabelle del database (nello specifico nelle tabelle clip e sentence) tutti i dati reperiti dal server. Una volta trasferiti i dati sul nostro database, nelle modalità sopra descritte, occorre eliminare i corrispettivi documenti nel server, per evitare l'esaurimento dello spazio concessoci dalla piattaforma Synthema.

Prima di descrivere il flusso di lettura e quello dei dati ottenuti dall'elaborazione del testo occorre approfondire l'argomento, più volte citato, ma fin'ora non ancora dibattuto, della lemmatizzazione. La lemmatizzazione è un processo di elaborazione del testo, che effettua la riduzione di una forma flessa di una parola alla sua forma canonica (non marcata), detta lemma.

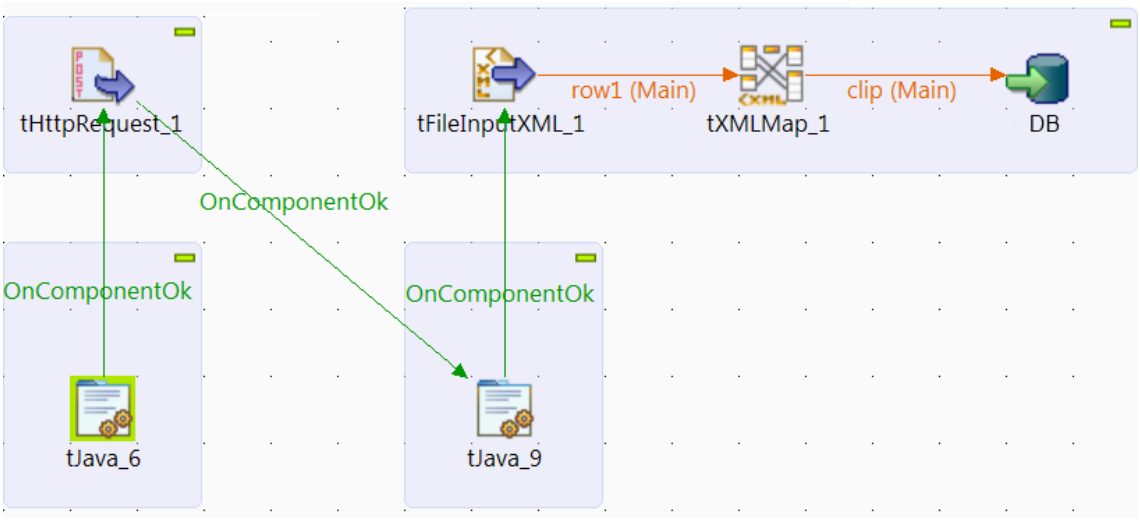

Figura 3.5 Processo di immagazzinamento (storing) dei documenti nel database operazionale

Nell'elaborazione del linguaggio naturale, la lemmatizzazione è il processo algoritmico che determina automaticamente il lemma di una data parola. Il processo può coinvolgere altre attività di elaborazione del linguaggio, quali ad esempio l'analisi morfologica e grammaticale. Nell'italiano così come in molte altre lingue, le parole appaiono in diverse forme flesse. Ad esempio il verbo *mangiare*, può apparire come *mangia*, *mangiò*, *mangiando* e cosi via. La forma canonica, *mangiare*, è il lemma della parola ed è la forma di riferimento per cercare la parola all'interno di un dizionario. Per un processo di analisi del testo la lemmatizzazione è di fondamentale importanza per diversi aspetti, sia di efficacia dell'analisi, che di efficienza dell'implementazione. Prima di tutto occorre ricordare, che per la corretta individuazione dei termini è necessario che la parola o multi word sia presente nelle librerie dei concetti, e che la conseguente tipizzazione ne definirà il significato. Senza la lemmatizzazione, l'inserimento di un concetto nella base di conoscenza comporterebbe il dover immettere (e quindi prima pensare) tutte le forme non canoniche che godono dello stesso significato di quel concetto. Questa eventualità, non è praticabile se pensiamo alle forme verbali che,

soprattutto nella lingua italiana, possiamo trovare in innumerevoli conformazioni. Questo aspetto condiziona fortemente sia l'efficienza dell'implementazione ma anche l'efficacia dell'analisi, qual'ora non individuassimo precisamente tutte le forme associabili al concetto che vogliamo introdurre nella nostra base di conoscenza. Data l'importanza della sua funzione e considerate le già citate difficoltà del motore in questa operazione, è stata realizzata una procedura in linguaggio c#. La procedura permette, consultando una tabella di look-up, basata su un dizionario dei lemmi della lingua italiana (morph-it_048) (Eros Zanchetta & Marco Baroni, 2005), di sostituire ogni termine presente nel testo, con il lemma base corrispondente, presente nel dizionario (esempio in figura 3.6).

Prima:

"Il Consiglio comunale di Agrigento ha approvato gli aumenti, al massimo previsto dalla legge, per Imu e addizionale Irpef."

Dopo:

"Il Consiglio comunale di Agrigento ***avere approvare il aumento****, al massimo previsto dalla legge, per Imu e addizionale Irpef".*

Figura 3.6 Esempio di lemmatizzazione

Formalmente non si tratta di una vera e propria lemmatizzazione, in quanto non viene effettuata un'analisi morfologica o grammaticale, bensì una sostituzione uno a uno con il lemma di riferimento. Questo può comportare, a volte, alcuni inconvenienti in quanto le parole prese una ad una senza conoscerne il contesto non hanno un'interpretazione sempre univoca (ad esempio il "colle", inteso come il colle del Viminale, senza modificare il dizionario diverrebbe, erroneamente nel il nostro contesto, "colla"). L'importanza che il ruolo della lemmatizzazione occupa all'interno dell'analisi ha prevalso su questi inconvenienti.

Affrontato il tema della lemmatizzazione, possiamo ora concentrarci sull'implementazione dei flussi di lettura dal database e di elaborazione e scrittura dei dati estratti, nella base dati. Questo anello, viene svolto interamente mediante lo strumento SPSS Modeler (con il quale si effettua anche l'analisi del testo, grazie al modulo dedicato SPSS Text Analytics), che ci permette di implementare i flussi di lettura e scrittura su un database operazionale. Come possiamo osservare dalla figura 3.7, il processo prevede la lettura della tabella *sentence* del database (riquadro piccolo rosso in figura 3.7), e successivamente l'elaborazione dei dati (riquadro grande blu in figura 3.7), che è suddivisa in due processi di analisi paralleli:

- Il primo effettua l'analisi del testo orientata a fornire il sentiment (opinione) espresso nella frase (un risultato aggregato).
- Il secondo effettua l'analisi del testo orientato a fornire il sentiment a livello dei singoli concetti, analizzando le relazioni che intercorrono tra di essi (potenzialmente possono esistere più opinioni espresse su più concetti nella stessa frase).

In seguito si effettua la scrittura dei dati (riquadri piccoli in verde in figura 3.7) risultanti dall'elaborazione nelle tabelle *sentence* (con la definizione del campo *sentiment*) e *wordsgroup* (con la definizione dei campi *firstword*, *secondword*, *sentiment* ecc, a seconda del pattern individuato).

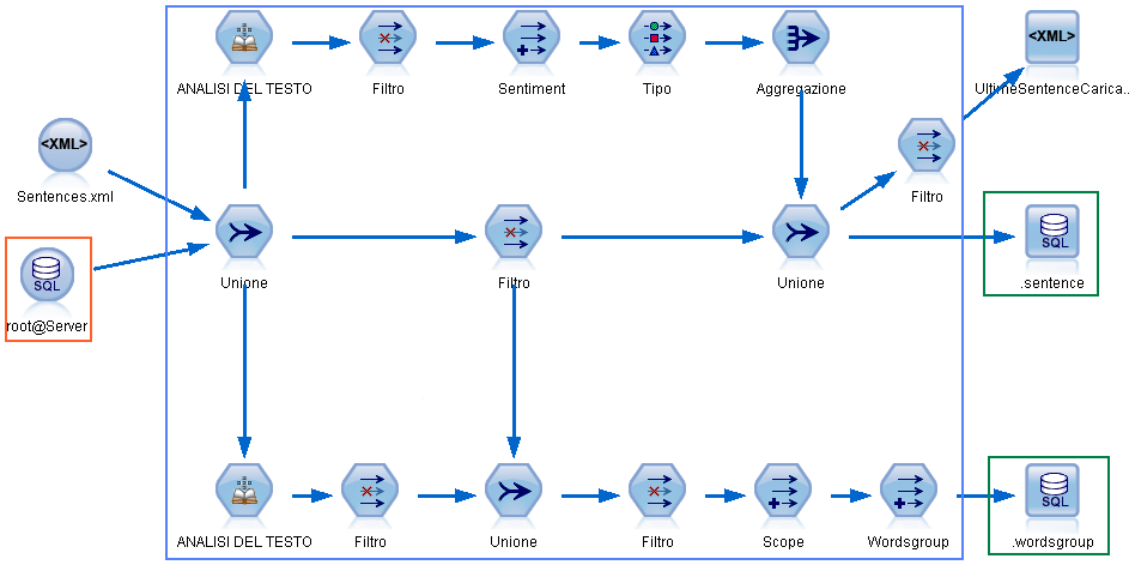

Figura 3.7 Anello di lettura, elaborazione e scrittura.

Dopo il processo di elaborazione, si propaga la scrittura dei dati risultanti nelle altre tabelle nello specifico di:

- *Entity*: con l'inserimento dei nuovi concetti individuati (che si riferiscono ai campi *firstword* e *secondword*), solo se non già presenti nella tabella.
- *Contain*: con la definizione delle associazioni tra i record della tabella *entity* e quelli della tabella *wordsgroup*.

Per concludere questa analisi minuziosa di tutti i dettagli implementativi del sistema, ricordiamo al lettore, che l'analisi del testo prevede l'esecuzione di attività di TLA (rappresentate in figura 3.3), che necessitano dell'implementazione di regole (si veda il capitolo due per maggiori dettagli) per funzionare correttamente. Riprendendo quanto detto nel capitolo due, le regole sono fondamentali per poter individuare, all'interno dei testi, pattern che riproducano strutture normalmente utilizzate nel linguaggio comune. E' in questo modo che lo strumento effettua, per così dire, l'analisi sintattica del testo, e permette di individuare le relazioni tra i concetti espressi. L'implementazione delle regole è stata incentrata sulle

prerogative imposte dal sistema di Social BI, ed in primo luogo, sulla sentiment analysis, che rappresenta la funzionalità più importante del sistema di Social BI, ma non l'unica. Nello specifico i requisiti considerati nella costruzione delle regole sono stati:

- rilevazione dei termini polarizzati (con accezione negativa o positiva) all'interno del testo, e dove possibile identificazione del concetto espresso a cui si faceva riferimento.
- rilevazione di tutte quelle strutture di frase, che seppur prive di sentiment (sentiment neutro), contenessero concetti di interesse per l'analisi.
- rilevazione di tutte quelle strutture di frase, che potessero evidenziare concetti cosiddetti "unknown", non di interesse, ma potenzialmente rilevanti per il dominio di ascolto.

Come già ampiamente sottolineato, le possibilità nella creazione delle regole sono praticamente infinite. Per contenere questa complessità abbiamo ideato un approccio di tipo "conservativo". Per approccio conservativo si intende la costruzione di regole a diversi livelli di "specificità", in modo tale da individuare pattern molto particolari e fini (meno ricorrenti ma più precisi) e allo stesso tempo, mediante altre regole, rilevare anche pattern più semplici (più ricorrenti ma meno precisi). Infatti, fin da subito è apparso chiaro, come individuare tutte le possibili strutture di frasi, in grado di rispettare i requisiti sopra elencati, fosse veramente arduo, ma allo stesso si percepiva la possibilità di ottenere comunque ottimi risultati con una combinazione intelligente di regole più o meno specifiche. In tutto sono state implementate più di 40 regole, di cui 32 per pattern specifici, tra le quali possiamo osservare, a titolo di esempio esplicativo, quella rappresentata in figura 3.8, e altre di tipo "nearto", cioè che si basavano sul concetto di distanza tra topic e opinione (figura 3.9).

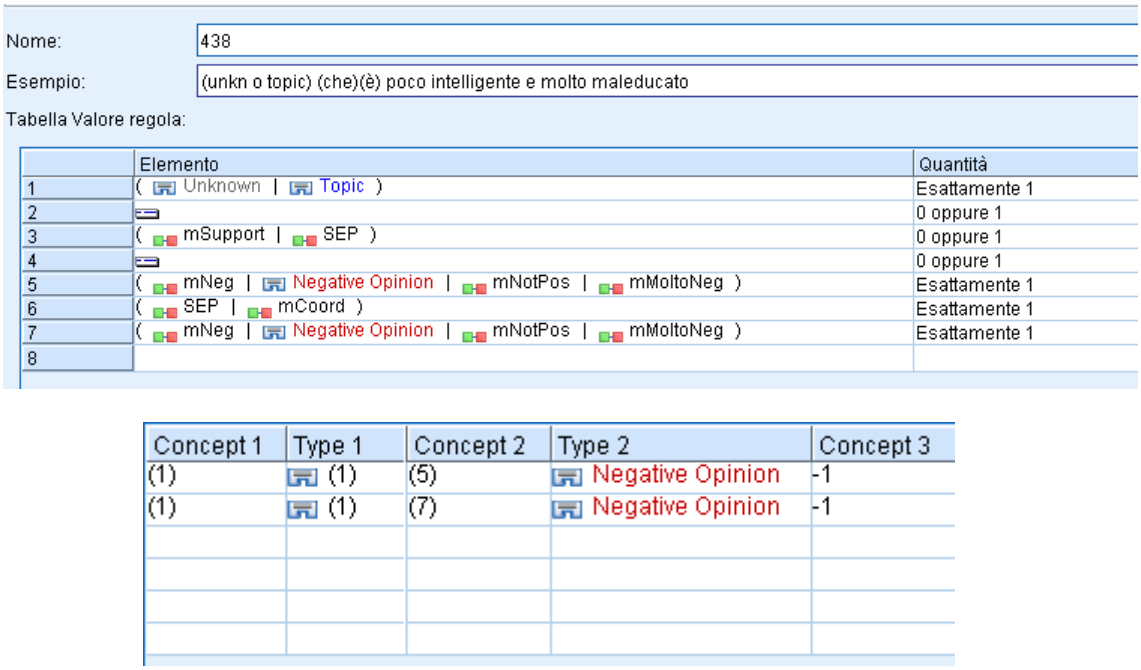

Figura 3.8 Esempio di regola per pattern specifico

Nella regola sopra illustrata, si ricercano tutti quei pattern in cui viene espressa una duplice opinione negativa ravvicinata, ma dove prima si è citato un concetto di interesse (topic) o un concetto potenzialmente rilevante per l'analisi (unknown) a cui le opinioni fanno riferimento. Questo pattern, così come altri di analoga tipologia, potrebbe risultare abbastanza infrequente, quindi è opportuno creare regole (di "salvataggio" o sicure) come quella raffigurata in figura 3.9, che ci permettano comunque di rilevare un sentiment nella struttura, tramite una regola "nearto" (di vicinanza) quindi meno precisa. Le regole di vicinanza sono molto semplici, in quanto sono costituite da un pattern composto da un concetto di tipo *topic* o *unknown* e a breve distanza (inframmezzo di 0 fino a 4 parole) un concetto polarizzato (nel caso in esame, negativo, molto negativo e non positivo). Infine, in linea con i requisiti del sistema, sono state implementate delle semplicissime regole per il riconoscimento all'interno della frase dei concetti di interesse (topic) e

concetti potenzialmente rilevanti ma non appartenenti (almeno non ancora) alla tassonomia (unknown), espressi singolarmente e privi di sentiment.

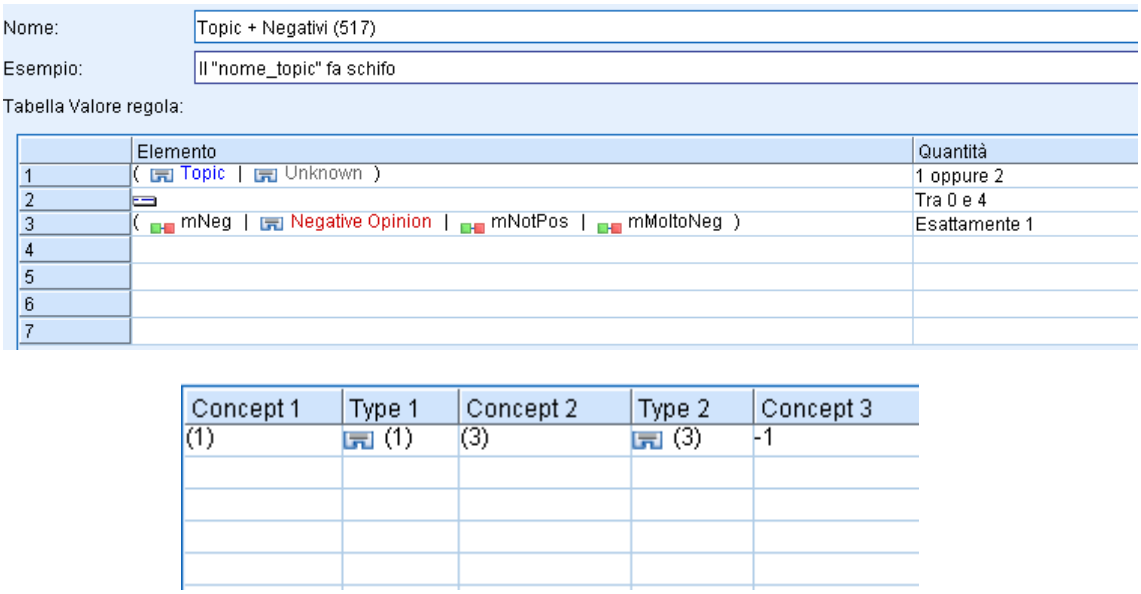

Figura 3.9 Esempio di regola per pattern generico

Conclusa l'implementazione delle regole secondo l'approccio sopra descritto, è di fondamentale importanza definire un ordinamento delle regole appropriato. Infatti, in alcuni casi può accadere, che alcune regole si attivino congiuntamente su uno stesso pattern contenuto in una frase, causando conflitti di attivazione. Diventa quindi importante definire un ordine di precedenza, con il quale decidere in caso di conflitto, quale regola, in maniera esclusiva, ha diritto di attivarsi e produrre il suo output. Per sfruttare appieno l'approccio di creazione utilizzato, occorre ordinare l'attivazione delle regole, in modo tale, da avere le regole orientate all'individuazione dei pattern più specifici, nelle prime posizioni della lista ordinata, mentre quelle più generiche nelle posizioni in fondo alla lista. Questo metodo ci permette di non pregiudicare la rilevazione di strutture particolari e molto precise, e allo stesso tempo garantisce l'individuazione (ovviamente se presente) del sentiment nella frase.

Capitolo 4

Caso di studio: la politica italiana

Questo capitolo è dedicato all'approfondimento del caso di studio affrontato per il progetto di tesi. Il caso scelto è l'ambito politico, più precisamente la politica nazionale italiana. In questa sezione si descriveranno, quindi, le peculiarità proprie del tema politico, alcune caratteristiche constatate durante lo sviluppo del sistema di Social Business Intelligence ed infine si illustreranno nel dettaglio le fasi di creazione della tassonomia, ontologia e la selezione delle fonti da cui sono stati estratti i documenti per l'analisi.

4.1 Introduzione al caso di studio

I casi di studio più naturali per progetti di Social Business Intelligence sono solitamente considerati quelli legati all'ambito commerciale e come già osservato legati al "brand/reputation management" o più ampiamente al settore marketing di una azienda. Ma la trasversalità della rete, consente analisi molto interessanti anche per tanti altri ambiti d'ascolto, non ultimo quello politico, i cui temi sono molto dibattuti sul web. Inutile nascondere come le incombenti elezioni politiche di fine febbraio 2013, facciano da traino a tutto questo, generando grande interesse in termini di numero pagine web dedicate a temi politici, ai partiti e agli stessi politici che si trovano in piena campagna elettorale. L'esplosione del web 2.0 prima e delle

piattaforme social poi, incidono su tantissimi aspetti della nostra vita. La rete, rispetto ai canonici canali come tv e radio, ha permesso una comunicazione bidirezionale, permettendo la partecipazione sociale e lo scambio di idee, opinioni e informazioni, per tutti i temi di interesse pubblico, compreso quello politico. E' ormai evidente come i social network siano diventati strumento di comunicazione fondamentale anche per i politici, resosi conto delle potenzialità di questo canale comunicativo (alcuni dicono che Obama abbia vinto le elezioni del 2009 grazie anche ai social media). Infatti tutti i principali politici e partiti possiedono oramai una propria pagina ufficiale in tutti i principali social network, senza considerare come alcuni movimenti politici siano addirittura nati proprio dal web. Durante il periodo di lavoro sul progetto di tesi, abbiamo potuto osservare come la partecipazione del pubblico sui temi politici, risulti molto viva sui social media. Se uniamo questa presenza sociale alle "tradizionali" fonti di informazione della rete, come quotidiani on-line,blog e portali web, ci troviamo di fronte un enorme valore informativo che possiamo e dobbiamo sfruttare.

4.2 Caratteristiche del dominio di ascolto

Innanzitutto è bene ricordare come l'analisi e lo studio approfondito del dominio di ascolto, sia fondamentale per poter operare con profitto in tutte le fasi della realizzazione del sistema, dalla configurazione del crawler (capitolo 2) alla verticalizzazione del sistema (capitolo 5).

Il dominio di ascolto politico, rispetto ad altri ambiti presenta alcune peculiarità degne di interesse. Prima di tutto la forte dinamicità delle tematiche trattate, che rispetto ad altri domini più statici, variano considerevolmente nel tempo in quanto seguono puntualmente le "tendenze" del periodo. I "trend topic" politici sono influenzati dagli avvenimenti della cronaca quotidiana (scandali, casi e notizie del momento) e per questo motivo sono incostanti nel tempo. Noi stessi ci siamo accorti, come siano cambiati i temi d'interesse dai momenti iniziali del nostro

progetto di tesi (settembre 2012) ad oggi. Abbiamo osservato come alcuni argomenti che avevano picchi di attenzione altissimi per determinati periodi, scomparivano in poco tempo dal "parlato" del web. Questa caratteristica non è comune ad esempio al dominio di digital marketing dove i vari prodotti e competitor (classici argomenti di interesse del dominio marketing) sono molto meno variabili, e di conseguenza i relativi topic sono molto più statici nel tempo. Un'altra caratteristica è l'eterogeneità delle tematiche politiche. E' indubbio come le tematiche riguardanti la politica siano spesso legate a quelle economiche e di giustizia, ampliando notevolmente la complessità del dominio di ascolto. Questi aspetti ci fanno intuire come l'ambito di ascolto politico racchiuda dentro di sé complessità non banali, rappresentando quindi un banco di prova importante per un sistema di Social Business Intelligence.

4.3 Creazione della tassonomia

Le prime fasi della metodologia sono indirizzate allo studio approfondito dell'ambito di ascolto, e hanno come obiettivo la formalizzazione della comprensione sul dominio mediante la realizzazione della cosiddetta tassonomia di dominio. La realizzazione della tassonomia, prevede una prima fase di macro-analisi, nella quale si prende conoscenza del dominio applicativo, si formalizzano i bisogni informativi (*inquiries*) del cliente e il cosiddetto scope, comprensivo dei topic e temi di interesse per l'analisi. Quando si parla di *inquiries* ci si riferisce alle interrogazioni e alle domande ad alto livello che il cliente intende porre al sistema una volta pronto, che sottintendono quindi i suoi bisogni in termini di informazioni e di conoscenza sul dominio. Le domande sono volutamente di carattere generico, destrutturare e non legate quindi a nessuna tecnologia. Nel caso in esame, quello politico, abbiamo immaginato la figura di un ipotetico giornalista (o comunque un soggetto interessato ai vari aspetti legati alla politica) che, ad esempio, per la redazione di un articolo è interessato a conoscere "l'opinione del web" su varie

tematiche, argomenti e/o soggetti politici. Ad esempio alcune domande e interrogazioni che il giornalista potrebbe avere l'esigenza di porre al sistema sono:

- Che cosa si dice del politico A?
- Che cosa si dice del partito B?
- Qual è l'opinione su tema C?
- Come è cambiata, in un intervallo di tempo (t_n,t_{n+1}), l'opinione su un politico D?
- Come è cambiata, in un intervallo di tempo (t_n,t_{n+1}), l'opinione su un tema E?
- Principali parole chiave associate al politico F?
- Principali parole chiave associate al tema G?

Inoltre nella creazione della tassonomia bisogna definire il concetto di *scope*, (obiettivo o scopo) cioè la serie di tematiche e soggetti, di interesse per il dominio in cui si opera. Lo scope è composto da macro-aree, che possiamo identificare come *Themes* (temi), che rappresentano le nostre macrotematiche. Nello specifico, per la politica italiana abbiamo individuato le macro regioni di interesse, raffigurate nella tabella 4.1.

Politici e Partiti	Politica Interna	Istruzione	Economia
Sanità	Lavoro	Sicurezza	Ambiente
Trasporti	Giustizia	Politica Sociale	

Tabella 4.1 Themes per la politica italiana

Lo *Scope* è composto anche da *Topics*, che sono l'unità elementare della struttura di classificazione e che corrispondono a vari argomenti specifici e di più basso livello rispetto ai *Themes*. Per il nostro dominio abbiamo individuato circa cento topics di interesse, di cui ne mostriamo un estratto nella tabella 4.2.

Popolo della Libertà	riforma dell'istruzione	legge elettorale
Silvio Berlusconi	crescita	spending review
Angelino Alfano	crisi	finanziamento ai partiti
Roberto Formigoni	debito pubblico	conflitto di interessi
Partito Democratico	campagna elettorale	Giorgio Napolitano
Pierluigi Bersani	spread	TAV
Matteo Renzi	liberalizzazioni	intercettazioni
Unione di Centro	Mario Monti	immigrazione
Movimento 5 Stelle	imu	pareggio di bilancio
Beppe Grillo	carico fiscale	corruzione
Sinistra Ecologia e Libertà	disoccupazione	criminalità organizzata
Niki Vendola	patrimoniale	ammortizzatori sociali
Futuro e Libertà	Roberto Maroni	occupazione

Tabella 4.2 Topics per la politica italiana

Ad un *Theme* possono fare riferimento *N Topics*, ma è altrettanto vero che un *Topic* può appartenere a più *Themes* (esempio in figura 4.1).

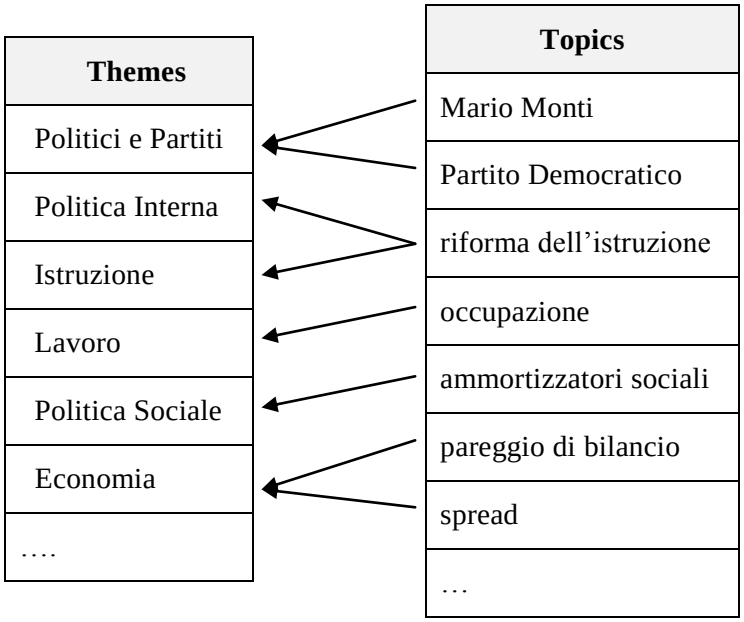

Figura 4.1 Correlazione tra Themes e Topic

I topics nel nostro dominio, sono ad esempio: i vari partiti politici, i politici di punta dei vari partiti, argomenti di interesse come riforme, tematiche correnti (di attualità) e ricorrenti (problema dell'occupazione, precariato) legate ad aspetti politici. Come è possibile intuire il dominio applicativo prevede temi ricorrenti, sempre più o meno attuali, ma anche temi legati a circostanze di attualità che hanno un picco di interesse in un determinato periodo ma che successivamente perdono importanza. Si rende noto come gli argomenti scelti in questa fase della creazione della tassonomia risalgano al periodo di settembre-ottobre 2012. Si noti ad esempio, la mancanza dei topics relativi alle nuove formazioni politiche formatesi nei mesi successivi.

La tassonomia così creata, non è però completa, va attuato un processo di arricchimento nel quale determinare i cosiddetti *alias* (se ricordiamo la struttura del database operazionale). Dobbiamo, cioè, identificare per ogni topic tutti i sinonimi che possono essere utilizzati per riferirsi al quel determinato argomento. L'utilità di questa operazione è piuttosto chiara: arricchire la base di conoscenza permette al motore di analisi di identificare tutte quelle strutture di frasi in cui compare un concetto connesso ad un topic di interesse e non solo il topic stesso. Per l'ambito politico, gli *alias* sono ad esempio:

- i soprannomi e i diversi modi utilizzati nel web per riferirsi ai politici (es: *Monti, Mario Monti, Professore* ecc.)
- tutte le sigle, abbreviazioni e nominativi estesi dei partiti politici (es: *lega, lega nord, carroccio*), delle sigle sindacali, dei ruoli del parlamento, istituzioni ecc (es: *ddl* e *disegno di legge, csm* e *consiglio superiore della magistratura*);
- diverse terminologie utilizzate per una stessa tematica (*patto di bilancio* e *fiscal compact*)

Al termine dell'arricchimento, l'intera tassonomia generata mediante questo processo di comprensione del dominio applicativo, va introdotta nella base di conoscenza del motore di analisi.

La tassonomia non è immutabile nel tempo. Periodicamente è necessario attivare un processo di aggiornamento dei topic d'interesse a fronte dell'evoluzione del dominio di ascolto. Questo scenario può essere dettato principalmente da due situazioni: dall'esigenza di inserire autonomamente un nuovo topic di interesse (ad esempio un nuovo politico o un nuovo partito) o dal rendersi conto dai dati provenienti dall'analisi dei testi, del palesamento di un argomento fortemente dibattuto non ancora compreso nella tassonomia. Quest'ultimo aspetto è molto interessante, perché grazie alla capacità del motore di recepire anche concetti non previsti dalla tassonomia ("unknown"), possiamo comunque avere ricoscerli. In tal modo siamo in grado di valutare, considerando anche il volume del "parlato", se introdurre o meno questi concetti nella tassonomia (diventando di interesse per l'analisi) e di conseguenza nella base di conoscenza del motore, mantenendo il pieno controllo sul dominio applicativo.

La tassonomia di dominio, svolge un ruolo importante anche per la definizione delle keyword di ricerca del crawler (si veda il capitolo 2 per maggiori dettagli). Infatti le keyword che serviranno per la ricerca e il reperimento dei documenti dal web, sono selezionate dai topic di interesse precedentemente definiti.

4.4 Selezione delle fonti

Determinati i themes e topics, dobbiamo valutare e selezionare le fonti web da cui estrarre i documenti che analizzeremo, andando alla ricerca di opinioni (sentiment) sui vari temi e soggetti di interesse. Considerata la vastità del mondo web e l'ambito di applicazione, abbiamo anzitutto suddiviso le fonti in due tipologie:

- Fonti Standard (siti di informazione, testate giornalistiche on-line, blog politici)
- Fonti Social (commenti e post sui social network Facebook e Twitter, blog e forum)

Come è facile immaginare le due tipologie di fonti hanno caratteristiche differenti. Le fonti standard, sono generalmente scritte con un lessico e sintassi più articolati e corretti, i contenuti sono più argomentati e la qualità generale del testo è alta. I testi sono più lunghi e contengono proporzionalmente un numero inferiore di opinioni. Le fonti social, soprattutto quelle provenienti da Facebook e Twitter, sono generalmente scritte con un gergo più povero e la sintassi della frase risulta molto spesso scadente e l'ortografia poco curata. Inoltre lo slang proprio di internet è molto presente con forme tipiche come: abbreviazioni, emoticons ecc. I testi sono più corti (per Twitter addirittura limitati a 140 caratteri) ma proporzionalmente contengono più opinioni e più termini polarizzati (negativamente/positivamente). La distinzione delle due tipologie è doverosa, in quanto gli approcci all'analisi del testo sono influenzati dalle diverse caratteristiche delle fonti. Semplificando, le prime considerazioni ci portano a pensare che testi meglio scritti (fonti standard) siano meglio compresi da metodi più linguistici mentre quelli più approssimativi (fonti social) siano meglio gestiti dai metodi più statistici. I test che mostreremo nel capitolo 6 ci diranno se le nostre previsioni sono state rispettate. Per ognuna delle due tipologie, esistono varie fonti che trattano i temi della politica, basti pensare alla quantità di testate giornaliste on-line, blog indipendenti, siti di informazione che (dei tanti contenuti, riportato anche fatti di politica nazionale), senza dimenticare i vari social network, compresi forum di discussione politica. Questo ci ha costretto a restringere il numero di fonti da prendere in considerazione per la fase di crawling dei documenti. Si è reso quindi necessario stilare una sorta di classifica delle migliori fonti. Il criterio di valutazione considerato è stato il numero di risultati restituiti da opportune "web search" con il motore di ricerca Google. Per questa attività abbiamo preselezionato un ristretto gruppo di parole chiave (generalmente coincidenti o sinonimi di topics precedentemente definiti) che facessero riferimento sia a temi politici di stretta attualità che ricorrenti. Ad esempio, due ricerche web effettuate si sono basate su argomenti di politica di stretta attualità nel periodo di ottobre 2012. I risultati, di una di queste query, mostrati nella tabella 4.3 evidenziano la superiorità, in termini di numero di

documenti trovati, delle testate on-line sui portali web, per le fonti standard e una netta preminenza dei social network sui forum e blog pubblici, per quanto riguarda le fonti social.

intercettazioni napolitano -video site:		
Tipologia	**Fonte**	**N° documenti**
Standard	ilgiornale.it	36000
Standard	repubblica.it	25800
Standard	ilfattoquotidiano.it	4100
Standard	ilsole24ore.com	6240
Standard	ansa.it	2890
Standard	Corriere.it	2400
Standard	rai.it	2050
Standard	sky.it	380
Social	facebook.com	105000
Social	twitter.com	92000
Social	beppegrillo.it	92
Social	dipietro.it	511
Social	giornalettismo.it	506
Social	polisblog.it	458
Social	politicainrete.it	1820
Social	chatta.it	517

Tabella 4.3 Risultati prima ricerca

Una seconda query effettuata si è basata, invece, su una serie di temi ricorrenti (non propriamente di stretta attualità). I risultati mostrati nella tabella 4.4 rilevano, anche in questo caso, un maggior numero di documenti per le testate web e i social network, rispetto alle altre fonti della stessa tipologia.

"riforma delle pensioni" OR disoccupazione OR imu OR intercettazioni OR "[riforma | legge] elettorale"		
Tipologia	**Fonte**	**N° documenti**
Standard	ilgiornale.it	72200
Standard	repubblica.it	58400
Standard	ilfattoquotidiano.it	14800
Standard	ilsole24ore.com	79100

Standard	ansa.it	127000
Standard	Corriere.it	154000
Standard	rai.it	33700
Standard	sky.it	19700
Social	facebook.com	473000
Social	twitter.com	312000
Social	beppegrillo.it	2900
Social	dipietro.it	1750
Social	giornalettismo.it	25900
Social	polisblog.it	3840
Social	politicainrete.it	37700
Social	chatta.it	118

Tabella 4.4 Risultati della seconda ricerca

Sommando il numero di documenti risultanti dalle ricerche web, possiamo valutare approssimativamente le fonti con la maggior quantità di documenti che trattano i topics di interesse. Sebbene non si tratti di un metodo scientifico, ci fornisce delle indicazioni utili per determinare le fonti più rilevanti per l'ambito della politica italiana.

Tipologia	**Fonte**	**N° documenti**
Standard	corriere.it	157060
Standard	ansa.it	130453
Standard	ilgiornale.it	109065
Standard	ilsole24ore.com	86780
Standard	repubblica.it	85520
Standard	rai.it	36260
Standard	sky.it	20161
Standard	ilfattoquotidiano.it	18904

Tabella 4.5 Classifica fonti standard

Per la tipologia delle fonti standard, quelle con il maggior numero di risultati, sono le testate giornalistiche on-line (probabilmente le più note), come è possibile osservare dalla tabella 4.5. Per la parte social, come ci aspettavamo, nei primi posti della classifica compaiono i social network come Facebook e Twitter (figura 4.6).

Tipologia	Fonte	N° documenti
Social	facebook.com	615300
Social	twitter.com	433100
Social	politicainrete.it	40046
Social	polisblog.it	32498
Social	giornalettismo.com	26514
Social	chatta.it	10175
Social	beppegrillo.it	3455
Social	antoniodipietro.it	2455

Tabella 4.6 Classifica fonti social

Osservando i risultati ottenuti, come selezione finale delle fonti abbiamo scelto le prime cinque del gruppo standard (che presentato un numero di risultati abbastanza omogeneo) e le prime due del gruppo delle fonti social:

- **Standard**
  - corriere.it
  - ansa.it
  - ilgiornale.it
  - ilsole24ore.com
  - repubblica.it
- **Social**
  - facebook.com
  - twitter.com

Capitolo 5

Metodologia di verticalizzazione

Nei capitoli precedenti abbiamo discusso delle problematiche relative al reperimento dei documenti (crawling) e del caso di studio (politica italiana) con l'analisi dell'ambito di ascolto e la creazione della tassonomia di dominio. Inoltre abbiamo descritto l'architettura del sistema, di Social Business Intelligence, realizzato e la sua implementazione. In questo capitolo affronteremo tutte le fasi del processo di verticalizzazione del motore di Text Analytics, fornendo una descrizione della metodologia impiegata e di tutte le problematiche incontrate.

5.1 Introduzione al tema della verticalizzazione

Prima di entrare nel merito della metodologia di verticalizzazione, approfondiamo alcuni aspetti importanti, come la funzione del motore di Text Anlytics (si veda anche capitolo dedicato all'architettura del sistema e delle tecnologie impiegate) e cosa si intende per verticalizzazione sul dominio d'ascolto e l'importanza che ricopre. Infine si fornirà una breve descrizione della base di conoscenza da cui ha origine il processo di arricchimento della conoscenza di dominio.

Come già osservato nel capitolo dedicato alle tecnologie di Text Mining, il linguaggio è molto variegato ed è altrettanto ricco di ambiguità. Il significato, che

attribuiamo ad un termine, è molto spesso legato al contesto in cui risiede, infatti non è sempre possibile conferire ad un concetto un significato univoco. Dal nostro punto di vista (umano) appare un fatto logico, ma la macchina non è in grado (ancora) di distinguere i diversi significati se non gli viene fornita una conoscenza in più: il contesto. Ma per fornire questa conoscenza al sistema è necessario effettuare un processo, denominato di "verticalizzazione" del sistema, definendo per ogni termine rilevante nel dominio, il significato o il ruolo univoco che possiede in quel preciso ambito di ascolto. Con un processo iterativo di verticalizzazione vedremo migliorare progressivamente le capacità del sistema nell'estrazione dei concetti di interesse (dell'ambito politico) e nell'identificazione delle relazioni tra di essi. Operativamente la verticalizzazione si effettua, mediante l'addestramento iterativo del sistema, osservandone la prestazioni con il testing periodico su una serie di dataset. Si rimanda ai paragrafi 5.2 per la costruzione dei dataset e al 5.3 per l'approfondimento sulla metodologia di verticalizzazione adottata.

Verticalizzare significa come già affermato, fornire al sistema, la conoscenza su un determinato dominio di ascolto. Per fare ciò, è necessario che il sistema sia dotato di una knowledge-base iniziale, che anche se non specializzata, sia in grado di riconoscere a grandi linee (senza pretendere grandi risultati) il linguaggio del testo, nel nostro caso della lingua italiana. Come già accennato nel capitolo due, la base di conoscenza italiana, che corrisponde al cuore del motore di Text Analytics, si è dimostrata deficitaria sotto tanti aspetti. In definitiva, si è preferito utilizzare come base di conoscenza iniziale, quella relativa ad un "vertical" di un altro ambito di ascolto (quello alimentare), realizzato in una precedente collaborazione del centro di ricerca, con una azienda del settore. Questa scelta ha aperto ulteriori opportunità, tra le quali, il comprendere in prima persona, come la verticalizzazione su di un particolare dominio, ne pregiudichi il funzionamento su di un altro. Per questo motivo, le operazioni di verticalizzazione, hanno previsto due processi paralleli e

simultanei: quello di "deverticalizzazione" del dominio alimentare e verticalizzazione di quello politico.

5.2 Costruzione dei dataset

Per il training e testing del motore, occorrono una serie di testi (dataset) che siano rappresentativi della forma e nel contenuto del dominio applicativo. A tale scopo abbiamo preselezionato una serie di testi provenienti dalle fonti determinate nella fase di reperimento dei documenti, creando tre dataset differenti, ciascuno composto da 483 frasi (sentences, più o meno lunghe), di cui:

- due (ds1 e ds2), contenenti testi provenienti da fonti di tipo standard;
- uno (ds Social), contente testi provenienti da fonti di tipo social (twitter e facebook).

Le frasi dei diversi dataset sono state etichettate manualmente, definendo per ognuna di esse le seguenti caratteristiche:

- Sentiment percepito: inteso come la polarizzazione dell'opinione espressa nel testo, relativamente al dominio di ascolto, che si divide in tre classi:
    - *negativa* (-1): il testo contiene una maggioranza di opinioni con accezione negativa.
    - *positiva* (+1): il testo contiene una maggioranza di opinioni con accezione positiva.
    - *neutra* (0): non contiene nessun tipo di opinione.

- Difficoltà del testo: intesa come valutazione della difficoltà di rilevazione del sentiment nel testo da un sistema automatico. Si divide in due classi:
    - *Difficile*: se il sentiment viene espresso, ad esempio, in maniera non evidente, velata o addirittura in chiave ironica (difficilmente percepibile dalla macchina)

- *Facile*: se il sentiment viene espresso, in maniera diretta, chiara e inequivocabile.

Come può subito apparire, stiamo considerando valutazioni soggettive, che possono quindi variare da individuo ad individuo. Per stabilizzare questa soggettività, i dataset sono stati etichettati da tre persone diverse e per ognuna delle caratteristiche di ogni testo, è stato definito il cosiddetto "majority vote" (voto di maggioranza). In sintesi, il *sentiment* finale di ogni sentence viene assegnato alla classe che ha avuto più "voti", la stessa cosa accade per la *difficoltà del testo*. Per il *sentiment* è possibile (in qualche sporadico caso) che accada una situazione di non accordo tra le tre persone (cosiddetto "no-majority" con tre rilevazioni differenti -1,0 e +1), mentre invece per la *difficoltà* trattandosi di una classificazione binaria, esiste sempre un voto di maggioranza. Per una panoramica completa, si può osservare la Tabella 5.1 che sintetizza la composizione dei dataset utilizzati. Il primo dei due (composto dai dataset ds1 e ds2) è costituito da testi recuperati dalle cosiddette fonti "qualificate" o standard, ovvero tutti quei siti autorevoli (selezionati nel capitolo relativo all'ambito di ascolto) che pubblicano notizie di politica italiana, come quotidiani e riviste on-line.

		Positivi	**Negativi**	**Neutri**	**Totale**	**Concordanza Media**
Dataset Standard	**Facile**	174	255	222	656	92,10%
	Difficile	82	122	99	310	85,10%
	Totale	256	377	321	966	89,20%
Dataset Social	**Facile**	83	119	37	244	93,88%
	Difficile	43	148	47	239	83,30%
	Totale	126	267	84	483	88,54%

Tabella 5.1 Composizione dei dataset per il traning e testing

Il poter attingere informazioni da queste fonti garantisce l'estrazione di testi corretti dal punto di vista sintattico e grammaticale, aspetto che deve essere tenuto in considerazione quando si valutano i risultati, ottenuti dall'elaborazione di un

sistema di analisi del testo. In figura 5.1, possiamo osservare la composizione del dataset relativo alle fonti standard, in particolare la suddivisione del campione estratto nelle classi di difficoltà e polarizzazione.

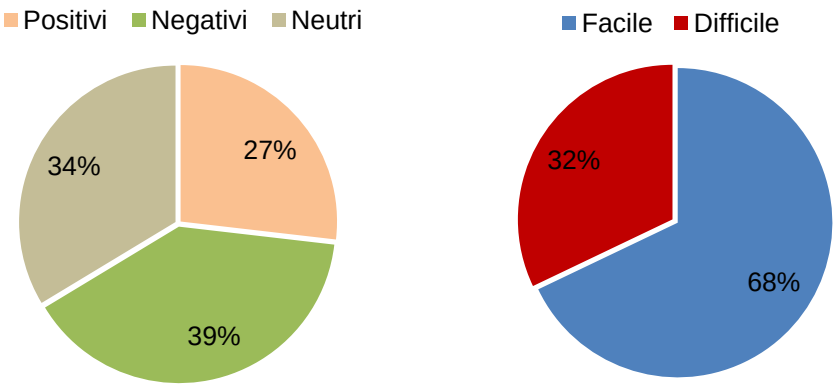

Figura 5.1 Composizione del dataset standard

Il secondo dataset utilizzato (ds Social) è stato invece costruito sulla base dei dati estratti dalle fonti social, nella fattispecie i social network Facebook e Twitter. I testi provenienti da queste fonti, sono caratterizzati da alcune peculiarità, tra le quali quella di essere scritti spesso con linguaggi dialettali e generalmente con forme sintattiche poco aderenti alla grammatica italiana. Questa caratteristica non può essere trascurata in fase di valutazione dei risultati di analisi del testo, in quanto potrebbe influire negativamente sui risultati. Infatti se vediamo la figura 5.2 possiamo osservare la composizione del dataset rappresentativo delle fonti social, che rispetto a quello relativo alle fonti standard, presenta un numero maggiore di testi catalogati come difficili. Inoltre la composizione della polarizzazione delle opinioni, risulta nettamente sbilanciata verso quelle negative, fatto tutto sommato comprensibile considerato l'ambito di ascolto.

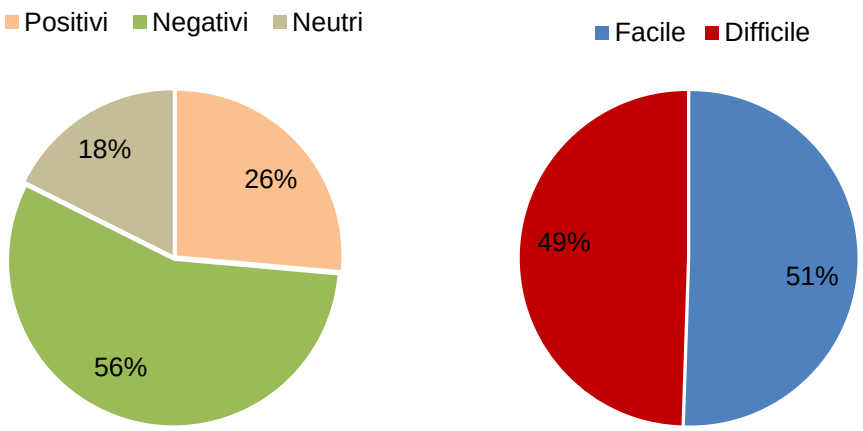

Figura 5.2 Composizione del dataset social

Si evidenzia che, al fine di separare i fattori di complessità crawler e motore semantico, i dataset sono composti da testi selezionati manualmente tra quelli restituiti dal crawler. Pertanto assumiamo che il crawler sia in grado di eliminare eventuali tag presenti nel testo e di segmentare correttamente un documento in più frasi. Oltre alla composizione in termini quantitativi la Tabella 5.1 riporta anche la *concordanza media* dei testi nel loro complesso o suddivisa per classe di polarizzazione. La concordanza media non è altro che il valore medio di concordanza del giudizio ottenuto dagli esperti, rispetto al voto di maggioranza, che viene considerato come un oracolo teorico. Questo valore può fornire un'indicazione sulla complessità del dataset, in quanto, dato un tasso medio di accordo nella valutazione del sentiment tra individui, non possiamo sperare che il sistema restituisca un risultato migliore. La risposta dell'oracolo è quindi definita come l'opinione di maggioranza, *Majority Group*, tra le valutazioni degli esperti. Il valore medio di concordanza calcolato, è vicino al valore medio di *inter-tagger agreement* definito tra individui, che viene stimato intorno dell' 80% in alcuni studi accademici (Gliozzo & Strapparava, 2009). Questo risultato ci permette di stabilire un punto di riferimento sul grado di efficacia di un sistema automatico di opinion

mining, proprio perché risulta inverosimile che la macchina oltrepassi le capacità di un utente umano.

A conclusione del percorso di ottimizzazione del sistema è stato creato un quarto dataset con la funzione esclusivamente di testset, ricavato dai documenti estratti dal crawler, in quel momento più recenti e composto da un totale di 300 testi. La tabella 5.2 mostra come la concordanza media sia sensibilmente inferiore a quella dei dataset iniziali, di conseguenza risulta un upper bound prestazionale inferiore rispetto ai precedenti dataset.

		Positivi	**Negativi**	**Neutri**	**Totale**	**Concordanza Media**
Dataset Finale	**Facile**	31	44	60	164	84,69%
	Difficile	18	64	79	136	67,70%
	Totale	49	108	139	300	74,44%

Tabella 5.2 Composizione del dataset finale

Ciò è dovuto alla composizione del dataset, che come possiamo osservare dalla figura 5.3 è caratterizzato da una prevalenza di testi con difficoltà *difficile*, mentre la composizione delle polarizzazioni risulta sbilanciata vero i testi neutri e negativi.

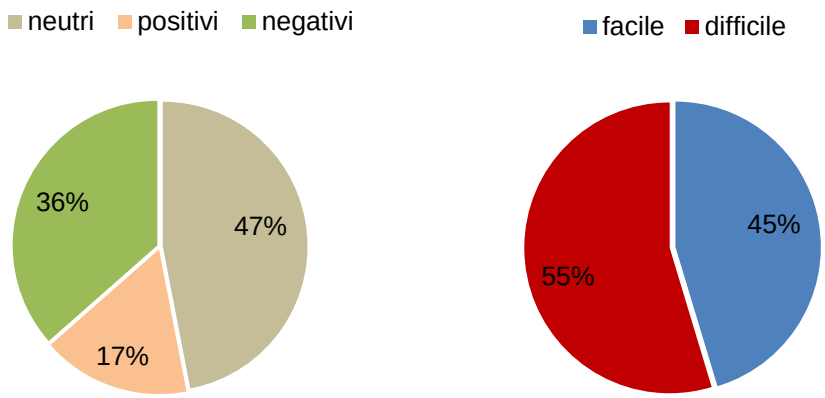

Figura 5.3 Composizione dataset per il test finale

Per questo dataset abbiamo aggiunto un'ulteriore valutazione, dettata dal fatto che la suddivisione in frasi viene effettuata automaticamente dal crawler (e quindi con possibilità di errore). Abbiamo valutato anche se il tipo di segmentazione potesse inficiare o meno la rilevazione del sentiment corretto, da parte del motore di analisi.

5.3 Arricchimento della base di conoscenza

Come abbiamo potuto constatare l'arricchimento della base di conoscenza di un sistema di Text Analytics, è un procedimento abbastanza laborioso, costituito da una serie di fasi come il testing, l'analisi, valutazioni, e solo al termine di questo processo è prevista l'effettiva fase di raffinamento e di modifica della base di conoscenza (creazione di una nuova release della knowledge-base). La metodologia adottata (rappresentata dalla figura 5.4), è un procedimento iterativo, in cui ad ogni iterazione si effettua:

- Testing della release
- Analisi dei risultati
- Modifica della base di conoscenza

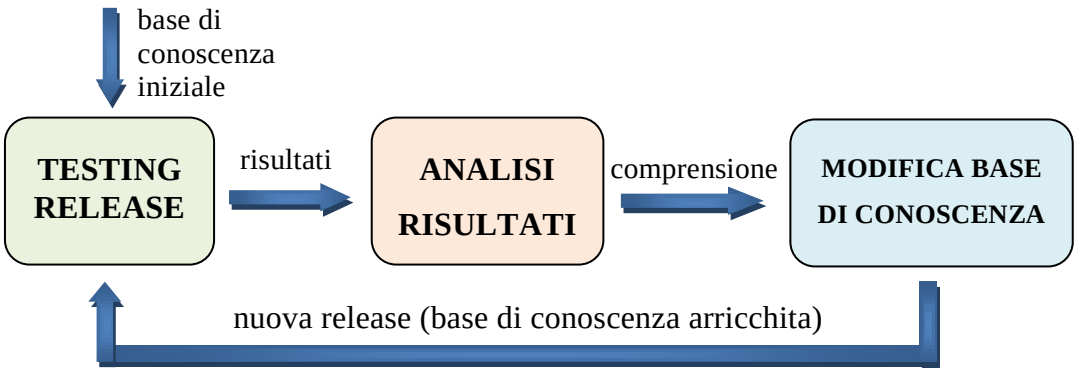

Fig. 5.4 Ciclo della metodologia di verticalizzazione del motore di Text Analytics

Per effettuare il testing della release (versione corrente della knowledge-base), si sottopongono al motore di analisi i testi dei vari dataset e si esegue l'elaborazione del testo. Dell'output prodotto, si considerano i risultati derivati dalla funzione di sentiment analysis del motore (che è costituito dalla release corrente della base di conoscenza). Successivamente si analizzano i risultati e si valutano le prestazioni in base alla concordanza del sentiment del sistema rispetto al sentiment di majority (d'ora in poi *mj*), su diversi criteri di valutazione:

- Totale: numero di testi totale, in cui c'è concordanza tra il sentiment del sistema con il *mj*
- Neutri: numero di testi con *mj sentiment* neutro correttamente rilevati come neutri
- Positivi: numero di testi con *mj sentiment* positivo correttamente rilevati come positivi
- Negativi: numero di testi con *mj sentiment* negativo correttamente rilevati come negativi
- Difficile: numero di testi con difficoltà *difficile* in cui c'è concordanza tra il sentiment del sistema e il *mj*
- Facile: numero di testi con difficoltà *facile* in cui c'è concordanza tra il sentiment del sistema e il *mj*.

In più per il dataset ds-Social si aggiungono due ulteriori parametri di valutazione:

- Twitter: numero di testi recuperati da twitter, in cui in cui c'è concordanza tra il sentiment del sistema e il *mj*.
- Facebook: numero di testi recuperati da facebook, in cui c'è concordanza tra il sentiment del sistema e il *mj*.

Nella fase di valutazione delle prestazioni, si confrontano le performance di ogni criterio sopra elencato, rispetto alla release precedente, calcolando gli scostamenti e valutando eventuali miglioramenti o peggioramenti.

La figura 5.5 rappresenta una serie di curve teoriche di miglioramento delle performance. Dall'andamento delle curve possiamo intravedere due fasi: nelle prime iterazioni della verticalizzazione si può assistere ad una forte crescita della qualità dei risultati, mentre dopo aver superato un certo grado di precisione si osserva una fase di stabilità asintotica, in cui raggiunto un livello prossimo al bound teorico, non si è più in grado di migliorare significativamente le performance. Le motivazioni che portano all'interruzione della crescita delle performance, possono essere molteplici e derivanti da alcuni fattori preponderanti come il limite imposto dalle tecnologie impiegate e/o l'intrinseca indeterminatezza dei testi.

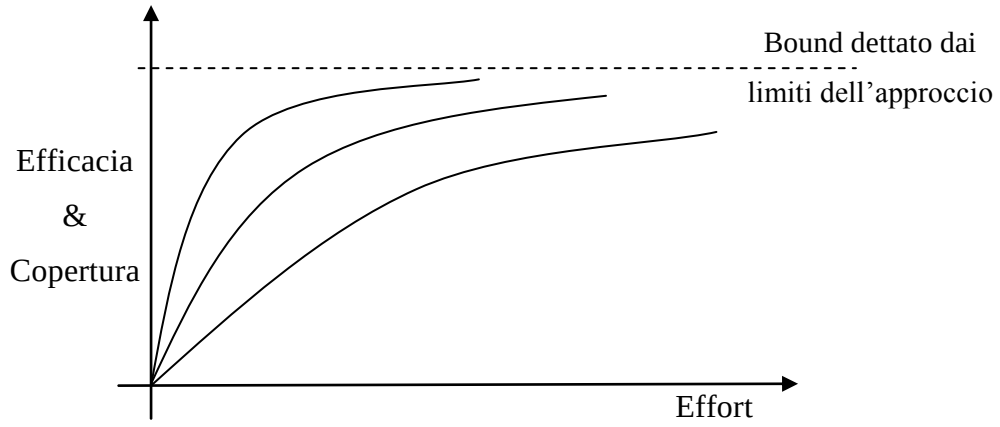

Figura 5.5 Curve teoriche di miglioramento delle performance di un sistema di social BI.

La fase di verifica della correttezza e completezza dei risultati è oltremodo importante perché fornisce l'input per la successiva iterazione del processo. Sottovalutare questa attività può comportare grossi rischi tra cui:

- L'introduzione di una quantità di errore superiore ai benefici apportati con una modifica. Un cambiamento qualsiasi della base di conoscenza (ad esempio l'inserimento o rimozione di uno o più concetti nelle librerie, oppure la variazione della polarizzazione di uno o più termini e così via)

potrebbe correggere il sentiment di una clip ma contemporaneamente potrebbe introdurre effetti collaterali, non controllati, su un altre clip.

- Prolungare oltremodo l'*effort* di arricchimento, anche quando sia ha già abbondantemente superato la parte di forte crescita della curva, provocando uno spreco di risorse a fronte di un limitato (se non addirittura nullo) ritorno, in termini di miglioramento delle prestazioni.

Tuttavia, sono tanti i fattori che possono influenzare il reale comportamento della curva di miglioramento (ad esempio le caratteristiche del progetto), per questo motivo deve essere provato empiricamente.

La verifica della correttezza ha richiesto, per forza di cose, l'etichettatura manuale dei dataset, descritta nella sezione precedente. Come abbiamo visto, la suddetta attività può riguardare uno o più aspetti (es. correttezza del sentiment, correttezza dell'identificazione di un termine/concetto, correttezza nella categorizzazione del documento) e risulta fondamentale per poter confrontare i risultati di sentiment estratti automaticamente dal sistema. Per determinare la curva di Figura 5.4 sarà quindi necessario calcolare al termine di ogni iterazione di arricchimento una opportuna misura quale ad esempio:

$$correttezza = \frac{\#\ clip\ correttamente\ etichettate}{\#\ clip\ considerate}$$

Questa misura può essere anche calcolata per aspetti specifici (per sorgente, per livello difficoltà, ecc.) in modo tale da comprendere precisamente per ognuno degli aspetti considerati, il mutamento delle prestazioni.

Per quanto riguarda la verifica della copertura ci troviamo di fronte ad un problema più difficoltoso considerato che nella quasi totalità dei casi è praticamente impossibile conoscere esattamente l'ampiezza del parlato per un certo dominio di ascolto. In riferimento alla Figura 5.6 l'obiettivo della fase di arricchimento della

base di conoscenza (figurativamente) è quello di avvicinare fino a far sovrapporre l'insieme del parlato (azzurro) con quello dell'ascoltato (arancione).

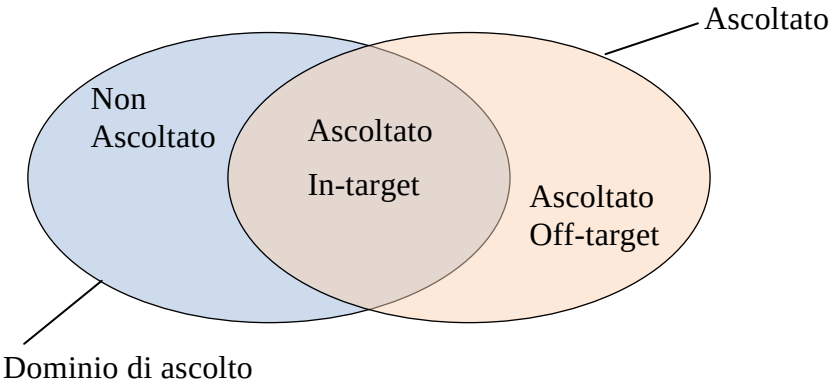

Figura 5.6 Rappresentazione grafica del rapporto tra ascoltato e reale dominio di ascolto

Capire la quantità di ascoltato Off-target, ovvero il numero di testi erroneamente recuperati dal crawler non appartenenti al dominio di ascolto, è relativamente semplice, può essere determinato con la misura seguente:

$$precision = \frac{\#\ clip\ InTarget}{\#\ clip\ catturate}$$

Risulta, invece molto più difficile, comprendere l'estensione del *Non Ascoltato* in quanto richiederebbe un sistema precisa che permetta di definire totalmente ed esclusivamente il dominio di ascolto. Per aggirare questo limite altrimenti difficilmente superabile, possiamo determinare un'approssimazione che ci permetta di caratterizzare il parlato di un certo dominio, mediante l'eseguzione di alcune interrogazioni sui più comuni motori di ricerca (es. Google, Bing e Yahoo). Se si suppone di conoscere la quantità di clip presenti all'interno del dominio di ascolto,

si può ricavare la percentuale delle clip In-target avvalendosi della seguente misura:

$$recall = \frac{\#\ clip\ InTarget}{\#\ clip\ nel\ dominio\ di\ ascolto}$$

Osservate le performance della release corrente, si passa all'analisi effettiva dei testi, con assoluta precedenza a quelle in cui il sentiment rilevato non trova corrispondenza con il majoriry vote, che quindi consideriamo come un errore di rilevazione. In questa analisi è importante cercare di comprendere se l'errore riscontrato è dovuto ad una carenza della base di conoscenza oppure se è totalmente (o almeno in parte) ascrivibile alle caratteristiche intrinseche del testo (ambiguità, modi di dire, ironia ecc). Si cerca quindi di effettuare una classificazione dell'errore, in modo da migliorare la comprensione sul problema. Ottenuta una panoramica dei casi di errore, passiamo alla fase di modifica della base di conoscenza. Le modifiche apportabili sono numerose e incidono su diverse attività dell'analisi del testo, più precisamente possiamo:

- Aggiungere un concetto nella libreria(termine singolo o multi word), nel caso possieda un'accezione polarizzata per l'ambito di ascolto, in particolare:
    - Se positiva aggiungendo il termine nei qualificatori di tipo positivo.
    - Se negativa aggiungendo il termine nei qualificatori di tipo negativo.
- Aggiungere un concetto nella libreria(termine singolo o multi word), a causa di una mancata rilevazione di un topic di interesse.
- Rimuovere un concetto presente nella libreria, per causa di un precedente errore di valutazione o perché si richiede un'operazione di "deverticalizzazione" di un altro dominio di ascolto.

- Creare una nuova regola, se si ritiene che le regole attuali non siano sufficienti per catturare un determinato pattern.

Durante il periodo dedicato alla verticalizzazione, abbiamo constatato come il processo di modifica sia delicato e incida fortemente sull'efficacia del sistema di analisi. Le criticità di questa fase, come abbiamo anticipato in precedenza, derivano soprattutto dal rischio di attuare correzioni non ben ponderate alla base di conoscenza, apportando benefici sul singolo testo, ma causando contemporaneamente effetti collaterali, in alcuni casi nemmeno visibili dal testing dei dataset di training (situazione di massima criticità). In questo senso, come approccio di base, si consiglia di approntare ogni modifica riflettendo sul dominio nel suo complesso, evitando di soffermarsi sul singolo errore e di perdere quindi di generalità. Perdere di generalità sul problema significa concentrarsi eccessivamente sui casi specifici dei dataset di training introducendo il cosiddetto problema dell'overfitting, situazione nella quale ci troveremmo se effettuassimo una modellazione della base di conoscenza, troppo connessa alla casistica particolare dei dataset di training. Si ricorda che in presenza di overfitting, si otterrebbero ovviamente migliori prestazioni (cioè la capacità di adattarsi e prevedere) sui dati di allenamento, ma avremmo prestazioni inferiori sui dati non visionati.

Capitolo 6

Analisi dell’efficacia

6.1 Introduzione all’analisi

Conclusa la realizzazione del prototipo del sistema e introdotta la metodologia di arricchimento della base di conoscenza, è arrivato il momento di valutare quantitativamente l’efficacia del sistema di Social Business Intelligence. Per la valutazione del prototipo realizzato, analizzeremo i risultati ottenuti da l’insieme di test svolti sul dominio di ascolto della politica italiana, utilizzato come caso di studio. L’attività di testing si è concentrata sulla valutazione della correttezza dei risultati di opinion mining (sentiment analysis) che risulta la funzionalità più importante del sistema di Social BI implementato e la più complessa da automatizzare, necessitando di una costante attività di perfezionamento. I test che presenteremo hanno un duplice scopo: fornire una valutazione effettiva dell’efficacia del sistema di analisi ed offrire indicazioni sperimentali sulla variazione delle performance riscontrate nelle attività di verticalizzazione e arricchimento.

La verticalizzazione della base di conoscenza sul dominio di ascolto, ha seguito un processo iterativo di arricchimento, descritto in tutte le sue fasi nella metodologia di verticalizzazione, nel capitolo precedente. Al termine di ogni iterazione, come esito delle modifiche, corrisponde una nuova versione della base di conoscenza. In

questa sezione riportiamo, per ogni release realizzata, i risultati dell'efficacia dell'analisi del sentiment, una loro breve interpretazione e i propositi di miglioramento per l'iterazione successiva. Osserveremo con particolare interesse, l'evoluzione delle prestazioni ottenute con i progressivi rilasci (quindi con il continuo perfezionamento) della base di conoscenza.

Per dare una visione completa sui perfezionamenti approntati, per ogni release realizzata, forniamo le seguenti informazioni:

- *Dataset di training*: si indica il dataset utilizzato per la fase di training, ossia quali testi sono stati analizzati, per fornire indicazioni al processo di modifica alla base di conoscenza. Utile per valutare la capacità di generalizzazione delle correzioni, sugli altri dataset.
- *Giorni richiesti*: si indica il numero di giorni uomo impiegati per l'analisi, per la comprensione e conseguente fase di modifica (ovvero il tempo dell'intera iterazione, che ha portato al conseguimento della release).
- *Modifiche apportate*: si indicano brevemente le tipologie di modifica apportate al sistema di analisi, differenziando i cambiamenti in correzioni alle librerie, alle TLA e di altri aspetti del sistema.

In più per ogni release, si riportano sottoforma di tabella, i risultati di efficacia suddivisi per :

- le diverse polarizzazioni (neutri, positivi, negativi);
- le tipologie di difficoltà (facile, difficile)
- le fonti Twitter e Facebook (solo Dataset Social)

Per meglio valutare i test che riporteremo nella prossima sezione, ricordiamo al lettore, che per i risultati di efficacia sul sentiment, il range di prestazioni ammissibile va dal 33% (performance minima ottenibile lanciando un ipotetico dado a 3 facce, una per ogni polarizzazione del sentiment) al 75-80% che possiamo definire come "upper bound" prestazionale, considerata la concordanza media tra utenti umani.

6.2 Test delle release

- **Release 1 (Base di conoscenza iniziale)**

La prima release rappresenta la base di conoscenza iniziale. Ricordiamo che le librerie dei termini sono state implementate per un altro dominio di ascolto, mentre le regole per l'individuazione dei pattern presentano carenze evidenti.

Dataset	Polariz.	Majority	Concordanza (Motore, MJ)	% Efficacia
ds1	**Neutri**	176	145	82,39*
	Positivi	109	0	0
	Negativi	190	87	45,79
	Totale	475	232	**48,84**
ds2	**Neutri**	145	122	84,14*
	Positivi	147	0	0
	Negativi	187	80	42,78
	Totale	479	202	**42,17**
ds Social	**Neutri**	84	69	82,14
	Positivi	126	0	0
	Negativi	267	94	35,21
	Totale	477	163	**34,17**

Tabella 6.1 Risultati della Release 1, differenziati per polarizzazione.

I risultati osservabili nelle tabelle 6.1 e 6.2, sono causati da gravi mancanze della base di conoscenza. Innanzitutto la totale assenza di regole per l'individuazione di pattern contenenti opinioni positive, ha ovviamente precluso la rilevazione di quest'ultime, mentre i termini presenti nelle librerie non sono risultati globalmente centrati sul dominio di ascolto politico. L'apparente ottimo risultato riscontrato sui testi con polarizzazione neutra, non deve trarre in inganno il lettore, in quanto palesemente condizionato dalla mancanza di rilevazione della polarizzazione positiva. Infatti la falsa rilevazione delle polarizzazione neutre è circa del 60%.

Dataset	Tipologia	Majority	Concordanza (Motore, MJ)	% Efficacia
ds1	**Difficile**	178	69	38,76
	Facile	297	163	54,88
ds2	**Difficile**	125	51	40,80
	Facile	354	151	42,66
ds Social	**Difficile**	238	74	31,09
	Facile	239	89	37,24
	Twitter	249	90	36,14
	Facebook	228	73	32,02

Tabella 6.2 Risultati della Release 1, differenziati per tipologie di difficoltà e fonte.

Si noti il divario prestazionale tra i diversi dataset e si osservi come le percentuali di efficacia sui testi provenienti dalle fonti social, risultino molto penalizzate rispetto ai dataset standard. Inoltre le prestazioni confermano le nostre valutazioni fatte con la classificazione della difficoltà del testo: in tutte le situazioni i testi etichettati "facili" hanno prestazioni significativamente più elevate di quelli "difficili".

- **Release 2**

L'iterazione che ha portato alla realizzazione di questa release, ha richiesto 5 giorni di attività. Per l'analisi dei testi, abbiamo utilizzato come training-set il dataset ds1 (fonti standard) e le modifiche apportate alla base di conoscenza, sono state:

- Librerie
  - Inserimento di 43 termini classificati negativi
  - Inserimento di 19 termini classificati positivi
  - Rimozione di 2 termini classificati negativi
  - Rimozione di 87 termini classificati positivi
- TLA
  - Implementazione 1 regola (per l'individuazione delle opinioni positive)

Dataset	Polariz.	Majority	Concordanza (Motore, MJ)	% Efficacia	%Risp. realese prec.
ds1	**Neutri**	176	126	71,59	-9,55
	Positivi	109	60	55,05	+53,57
	Negativi	190	85	44,74	-1,05
	Totale	475	271	**57,05**	**+8,23**
ds2	**Neutri**	145	96	66,21	-17,93
	Positivi	147	80	54,42	+54,42
	Negativi	187	56	29,95	-12,83
	Totale	479	232	**48,43**	**+6,26**
ds Social	**Neutri**	84	64	76,19	-5,95
	Positivi	126	82	65,08	+65,08
	Negativi	267	72	26,97	-8,24
	Totale	477	218	**45,70%**	**+11,53**

Tabella 6.3 Risultati della Release 2, differenziati per polarizzazione.

I risultati osservabili nelle tabelle 6.3 e 6.4, dimostrano un netto miglioramento complessivo su tutti i dataset testati. L'implementazione della regola per

l'individuazione delle opinioni positive ha generato il risultato sperato (quasi +58% in media). Inoltre il training effettuato sui testi del dataset ds1, ha portato benefici anche sugli altri dataset (+6,26 sul ds2 e +11,53 sul ds-Social). Questo rappresenta un risultato importante perché dimostra che le modifiche apportate possiedono ottime capacità di generalizzazione.

Dataset	Tipologia	Majority	Concordanza (Motore, MJ)	% Efficacia	Risp. realese prec.
ds1	**Difficile**	178	81	45,51	+7,87
	Facile	297	190	63,97	+9,09
ds2	**Difficile**	125	54	43,20	+2,40
	Facile	354	178	50,28	+7,63
ds Social	**Difficile**	238	78	32,77	+1.68
	Facile	239	140	58,58	+21,34
	Twitter	249	117	46,99	+10.85
	Facebook	228	101	44,30	+12.28

Tabella 6.4 Risultati della Release 2, differenziati per tipologie di difficoltà e fonte.

Si noti come l'efficacia, soprattutto per i dataset non di training, abbia un miglioramento molto più significativo per i testi di tipologia facile rispetto a quelli difficili.

- **Release 3**

L'iterazione che ha portato alla realizzazione di questa release, ha richiesto 4 giorni di attività. Per l'analisi dei testi, abbiamo utilizzato come training-set il dataset ds1 (fonti standard) e le modifiche apportate alla base di conoscenza, sono state:

- Librerie
  - Inserimento di 167 termini classificati negativi
  - Inserimento di 58 termini classificati positivi
  - Rimozione di 12 termini classificati negativi
  - Rimozione di 30 termini classificati positivi

Dataset	Polariz.	Majority	Concordanza (Motore, MJ)	% Efficacia	Risp. realese prec.
ds1	**Neutri**	173	114	64,77	-8,06
	Positivi	112	71	65,14	+11,57
	Negativi	190	130	68,42	+23,68
	Totale	475	315	**66,74**	**+9,69**
ds2	**Neutri**	145	97	66,90	+0,69
	Positivi	147	74	50,34	-4,06
	Negativi	187	74	39,57	+9,62
	Totale	479	245	**51,15**	**+2,30**
ds Social	**Neutri**	84	64	76,19	0
	Positivi	126	85	67,46	+2,38
	Negativi	267	82	30,71	+3,74
	Totale	477	231	**48,43**	**+2,73**

Tabella 6.5 Risultati della Release 3, differenziati per polarizzazione.

I risultati osservabili nella tabella 6.5, dimostrano un buon miglioramento generale in tutti i dataset testati. L'arricchimento della base di conoscenza con l'inserimento di concetti negativi (per l'ambito di ascolto) ha migliorato sensibilmente le prestazioni per l'individuazione della polarizzazione negativa (quasi +13% in media). Inoltre il training effettuato sui testi del dataset ds1, ha portato benefici

sugli altri dataset, anche se in maniera meno marcata. L'efficacia sulla polarizzazione negativa risulta ancora non sufficiente, soprattutto per i dataset ds2 e per quello Social. Sarà necessario nelle prossime release concentrarsi anche su questi testi per ampliare l'insieme di conoscenza del motore di analisi.

Dataset	Tipologia	Majority	Concordanza (Motore, MJ)	% Efficacia	Risp. realese prec.
ds1	**Difficile**	178	104	58,43	+11,80
	Facile	297	211	71,04	+7,75
ds2	**Difficile**	125	52	41,60	-1,60
	Facile	354	193	54,52	+4,24
ds Social	**Difficile**	238	81	34,03	+1,26
	Facile	239	150	62,76	+4,18
	Twitter	249	124	49,80	+2,81
	Facebook	228	107	46,93	+2,63

Tabella 6.6 Risultati della Release 3, differenziati per tipologie di difficoltà e fonte.

Si noti dalla tabella 6.6 come, anche in questo caso, l'efficacia per i dataset non impiegati nel training (nella fattispecie ds2 e ds-Social), abbia un miglioramento molto più significativo per i testi di tipologia facile rispetto a quelli difficili. Questo ci porta a pensare, che le capacità di generalizzazione incidano maggiormente su questa tipologia di frasi.

- **Release 4**

L'iterazione che ha portato alla realizzazione di questa release, ha richiesto 4 giorni di attività. Per l'analisi dei testi, abbiamo utilizzato come training-set il dataset ds2 (fonti standard) e le modifiche apportate alla base di conoscenza, sono state:

- Adozione di una procedura esterna di lemmatizzazione
- Librerie
  - Inserimento di 151 termini classificati negativi
  - Inserimento di 46 termini classificati positivi
  - Rimozione di 82 termini classificati negativi
  - Rimozione di 129 termini classificati positivi
- TLA
  - Implementazione 4 regole

Dataset	Polariz.	Majority	Concordanza (Motore, MJ)	% Efficacia	Risp. realese prec.
ds1	**Neutri**	176	101	57,39	-7,38
	Positivi	109	74	67,89	2,75
	Negativi	190	135	71,05	+2,63
	Totale	475	310	**65,26**	**-1,48**
ds2	**Neutri**	145	98	67,59	+0,69
	Positivi	147	95	64,63	+14,29
	Negativi	187	135	72,19	+32,62
	Totale	479	328	**68,48**	**+17,75**
ds Social	**Neutri**	84	51	60,71	-15,48
	Positivi	126	73	57,94	-9,52
	Negativi	267	105	39,33	8,62
	Totale	477	229	**48,01**	**-0,42**

Tabella 6.7 Risultati della Release 4, differenziati per tipologie di difficoltà e fonte.

I risultati osservabili nelle tabelle 6.7 e 6.8, dimostrano un miglioramento evidente per i testi contenuti nel dataset di training (ds2). Considerate le buone prestazioni

già raggiunte sui dataset standard, questa condizione potrebbe voler significare che siamo vicini al limite prestazionale raggiungibile con questo tipo di approccio. I risultati piuttosto deludenti sul dataset ds-Social, richiedono invece, un'analisi approfondita dei testi provenienti da queste fonti, in quanto le diverse caratteristiche morfologiche riscontrate (già evidenziate) potrebbero essere la causa delle scarse prestazioni.

Dataset	Tipologia	Majority	Concordanza (Motore, MJ)	% Efficacia	Risp. realese prec.
ds1	**Difficile**	178	95	53,37	-5,06
	Facile	297	215	72,39	+0,67
ds2	**Difficile**	125	64	51,20	+9,60
	Facile	354	264	74,58	+20,06
ds Social	**Difficile**	238	76	31,93	-2,10
	Facile	239	153	64,02	+1,26
	Twitter	249	128	51,41	+1,61
	Facebook	228	101	44,30	-2,63

Tabella 6.8 Risultati della Release 4, differenziati per tipologie di difficoltà e fonte

Si noti come, il risultato di l'efficacia ottenga un buon miglioramento per i testi di tipologia facile, soprattutto per i dataset non impiegati nel training (ds1 e ds-Social). Si consideri anche il divario prestazionale formatosi tra le fonte Twitter (51,41%) e Facebook (44,30%). Ancora non possiamo dare un giudizio definitivo in merito, ma la tendenza attuale, vede i testi provenienti dalla fonte Twitter maggiormente comprensibili rispetto a quelli di Facebook.

- **Release 5**

L'iterazione che ha portato alla realizzazione di questa release, ha richiesto 3 giorni di attività. Per l'analisi dei testi, abbiamo utilizzato come training-set il dataset ds-Social (fonti social) e le modifiche apportate alla base di conoscenza, sono state:

- Librerie
  - Inserimento di 90 termini classificati negativi
  - Inserimento di 21 termini classificati positivi
  - Rimozione di 0 termini classificati negativi
  - Rimozione di 27 termini classificati positivi
- TLA
  - Implementazione 4 regole

Dataset	Polariz.	Majority	Concordanza (Motore, MJ)	% Efficacia	Risp. realese prec.
ds1	**Neutri**	176	105	59,66	+2,27
	Positivi	109	73	66,97	-0,92
	Negativi	190	135	71,05	0
	Totale	475	313	**65,89**	**+0,63**
ds2	**Neutri**	145	99	68,28	+0,69
	Positivi	147	97	65,99	+1,36
	Negativi	187	132	70,89	-1,60
	Totale	479	328	**68,48**	**0**
ds Social	**Neutri**	84	54	64,29	+3,58
	Positivi	126	90	71,43	+13,49
	Negativi	267	146	54,68	+15,35
	Totale	477	290	**60,80**	**+12,79**

Tabella 6.9 Risultati della Release 5, differenziati per tipologie di difficoltà e fonte.

La release 5 è la prima basata sull'osservazione dei testi della tipologia social e quindi dell'analisi delle sue peculiarità morfologiche. I risultati osservabili nella tabella 6.9, dimostrano un miglioramento complessivo evidente per il dataset

ds-Social, mentre per i dataset standard i risultati sono pressoché invariati. Le modifiche apportate con l'inserimento di termini negativi come imprecazioni ed epiteti (prima non previsti nella base di conoscenza) tipici del mondo social, hanno giovato soprattutto a quest'ultimi testi. Il risultato dell'efficacia delle opinioni negative, sul ds-Social, ha raggiunto un risultato confortante (54,68%). Le difficoltà su questa tipologia di testi, sono maggiormente accentuate dal forte contenuto ironico, spesso presente nei testi provenienti dalle fonti social, (soprattutto per questo ambito di ascolto), difficilmente riconoscibile da un motore di analisi.

Dataset	Tipologia	Majority	Concordanza (Motore, MJ)	% Efficacia	Risp. realese prec.
ds1	**Difficile**	178	96	53,93	+0,56
	Facile	297	217	73,06	+0,67
ds2	**Difficile**	125	69	55,20	+4,00
	Facile	354	259	73,16	-1,42
ds Social	**Difficile**	238	100	42,02	+10,09
	Facile	239	190	79,50	+15,09
	Twitter	249	152	61,04	+9,63
	Facebook	228	138	60,53	+16,23

Tabella 6.10 Risultati della Release 5, differenziati per tipologie di difficoltà e fonte

Anche i risultati riportati in tabella 6.10, indicano una stabilizzazione delle prestazioni per le fonti standard ed inoltre, si noti come il divario prestazionale tra *Twitter* e *Facebook* mostrato dalla precedente release, sia stato quasi interamene colmato dalle ultime modifiche.

- **Release 6**

L'iterazione che ha portato alla realizzazione di questa release, ha richiesto un solo giorno di attività. Per l'analisi dei testi, abbiamo utilizzato come training-set il dataset ds-Social (fonti social) e le modifiche apportate alla base di conoscenza, sono state:

- Librerie
  - Inserimento di 34 termini classificati negativi
  - Inserimento di 3 termini classificati positivi
  - Rimozione di 0 termini classificati negativi
  - Rimozione di 6 termini classificati positivi

Dataset	Polariz.	Majority	Concordanza (Motore, MJ)	% Efficacia	Risp. realese prec.
ds1	**Neutri**	176	104	59,09	-0,57
	Positivi	109	70	64,22	-2,75
	Negativi	190	135	71,05	0
	Totale	475	309	**65,05**	**-0,84**
ds2	**Neutri**	145	100	68,97	+0,69
	Positivi	147	95	64,63	-1,36
	Negativi	187	133	71,12	+0,53
	Totale	479	328	**68,48**	**0**
ds Social	**Neutri**	84	55	65,48	+1,19
	Positivi	126	82	65,08	-6,35
	Negativi	267	167	62,55	+7,87
	Totale	477	304	**63,73**	**+2,93**

Tabella 6.11 Risultati della Release 6, differenziati per tipologie di difficoltà e fonte.

L'iterazione 6, che ha prodotto questa release, può essere considerate come un prolungamento dell'iterazione precedente, infatti è risultata la più corta dell'intera verticalizzazione ed ha avuto come principale obiettivo quello di identificare con certezza, il limite prestazionale raggiungibile sulle fonti Social. I risultati

osservabili nella tabella 6.11, dimostrano un piccolo miglioramento per il dataset ds-Social, mentre per i dataset standard i risultati sono pressoché invariati. Dall'analisi dei risultati della corretta individuazione delle diverse polarizzazioni, si evidenzia come il divario prestazionale tra l'efficacia di rilevazione delle opinioni positive e negative sia divenuto molto ridotto. Infatti il miglioramento evidente conseguito sulle opinioni negative (+7,87%) è stato controbilanciato da un calo importante su quelle positive (-6,35%). Quando, per migliorare una componente del nostro sistema, siamo costretti ad introdurre un effetto collaterale evidente su un altro aspetto, tale condizione è sintomatica del fatto che siamo vicini al limite raggiungibile dall'approccio impiegato.

Dataset	Tipologia	Majority	Concordanza (Motore, MJ)	% Efficacia	Risp. realese prec.
ds1	**Difficile**	178	96	53,93	0
	Facile	297	213	71,72	-1,34
ds2	**Difficile**	125	69	55,20	0
	Facile	354	259	73,16	0
ds Social	**Difficile**	238	115	48,32	+6,30
	Facile	239	189	79,08	-0,42
	Twitter	249	165	66,27	+5,23
	Facebook	228	139	60,96	+0,43

Tabella 6.12 Risultati della Release 6, differenziati per tipologie di difficoltà e fonte

Si noti, dalla tabella 6.12, come il divario tra *Facebook* e *Twitter* (a favore di quest'ultimo) sia tornato abbastanza marcato con le ultime modifiche.

6.3 Sintesi dei risultati e considerazioni

Data l'ingente mole di test effettuati sulle diverse versioni della base di conoscenza, occorre sintetizzare i risultati ottenuti ed individuare gli aspetti più significativi. Se osserviamo l'evoluzione delle prestazioni riscontrata nelle diverse release, per i testi provenienti dalle fonti qualificate (raggruppate in un unico dataset), possiamo notare un rapido e costante miglioramento delle prestazioni (figura 6.1) fino alla release 4, in cui abbiamo raggiunto il limite prestazionale dell'approccio impiegato. Come ci aspettavamo, sono state le prime iterazioni quelle più proficue, proprio perché, sono le prime fasi quelle che introducono le modifiche più fruttuose, in grado di generare i miglioramenti più evidenti. Un altro aspetto importante, già evidenziato nell'analisi dei risultati delle singole release, è il divario prestazionale sempre più marcato tra i testi di tipologia difficile e quelli di tipologia facile (da una iniziale differenza di 8,63 si arriva fino ad uno scarto massimo di 21,1 punti). Risulta evidente come l'arricchimento della base di conoscenza apporti i maggiori benefici ai testi valutati come facilmente analizzabili, mentre si ha la netta sensazione che su quelli considerati difficili non si possa incidere più di tanto. Non è cosa del tutto inaspettata, se ricordiamo le caratteristiche di questi testi e se consideriamo le forti difficoltà nell'infondere al motore la comprensione dell'ironia, dei modi di dire e di forme gergali poco comuni. Per le fonti social, che valutiamo distintamente da quelle qualificate considerate le innumerevoli differenze del linguaggio utilizzato, riscontiamo un comportamento apparentemente diverso.

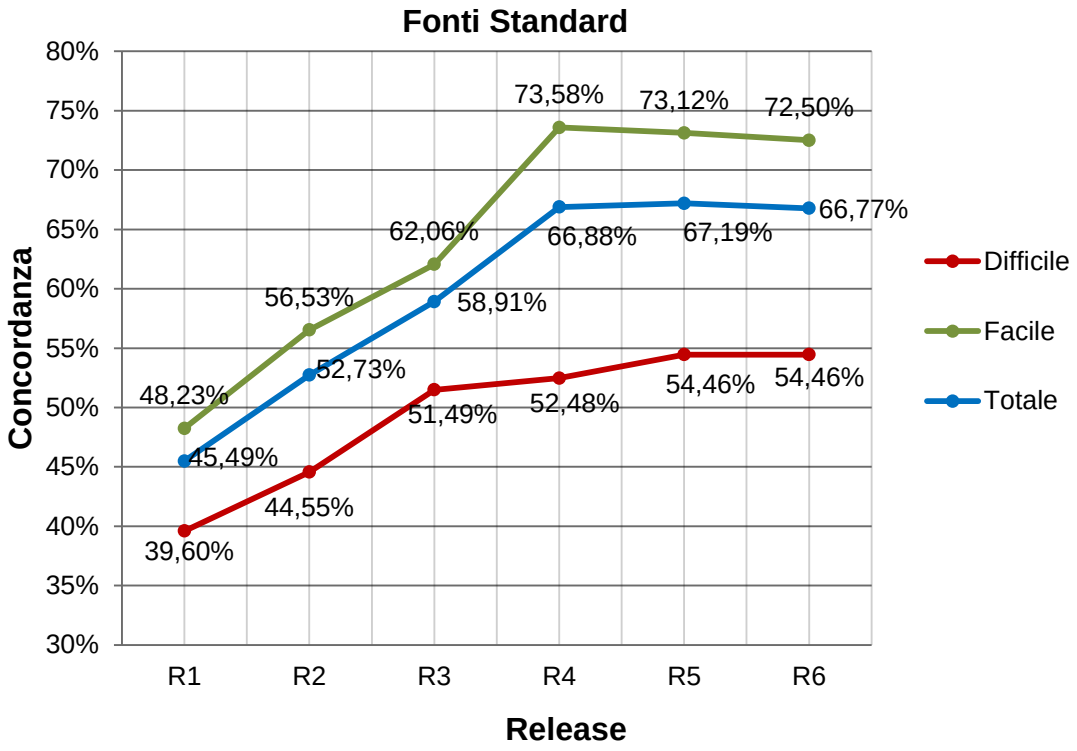

Figura 6.1 Evoluzione dell'efficacia per i testi provenienti dalle fonti standard

L'evoluzione dei risultati riportati in figura 6.2 evidenzia un netto miglioramento iniziale, dovuto in parte anche da un risultato di partenza (34,17%), inferiore rispetto a quello relativo alle fonti standard (45,49%). Successivamente (nelle release 2,3 e 4 prodotte sulla base dei testi dei dataset standard) si osserva subito un assestamento delle prestazioni. Da questo comportamento comprendiamo come le specificità del linguaggio impiegato nell'ambito social non siano ancora comprese dal motore di analisi. Con le release 5 e 6, introduciamo nella base di conoscenza quegli elementi caratteristici del linguaggio dei social network, e otteniamo di conseguenza un considerevole salto di qualità dei risultati di efficienza (+12,79% complessivo).

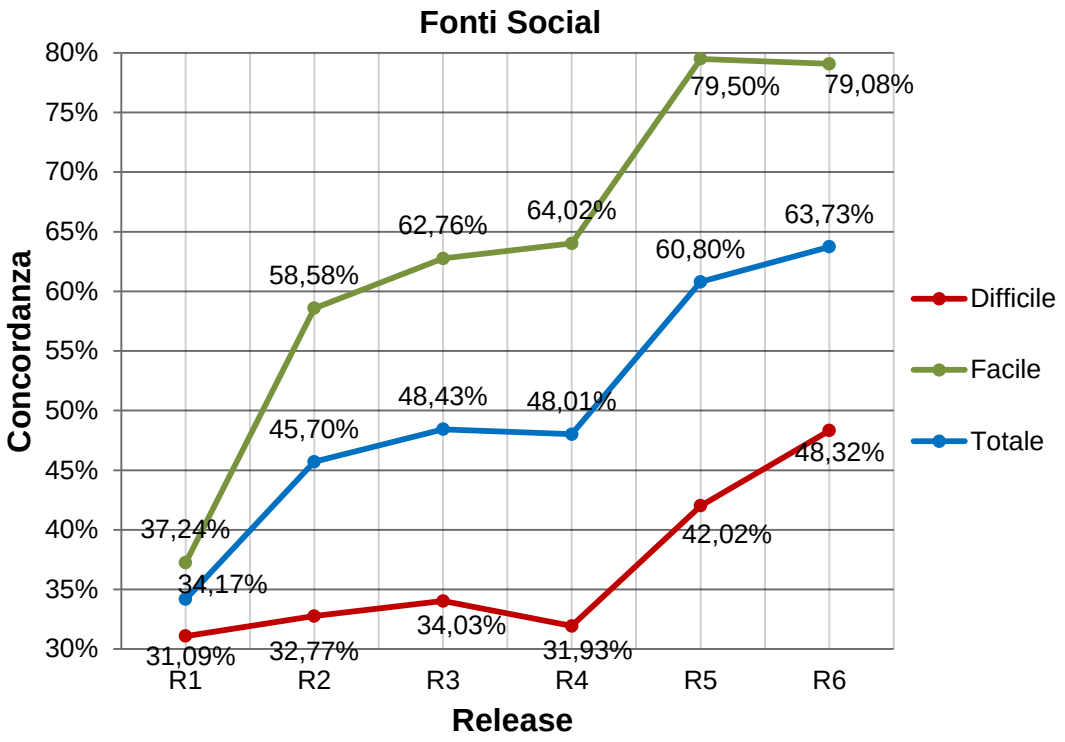

Figura 6.2 Evoluzione dell'efficacia per i testi provenienti dalle fonti social

La release 5 può essere considerata come la prima vera release per le fonti social, dove introduciamo la maggiore quantità di conoscenza sul dominio politico dal punto di vista dei social network. Interessante è anche notare l'evoluzione delle prestazioni per tipologia di polarizzazione, osservabile nella figura 6.3 per le fonti standard e 6.4 per le fonti social. Per le fonti qualificate, l'andamento delle tre polarizzazioni è molto diseguale, a causa delle lacune iniziali della prime release della base di conoscenza, nelle quali, se ricordiamo, l'efficacia sui sentiment positivi e negativi era molto scadente.

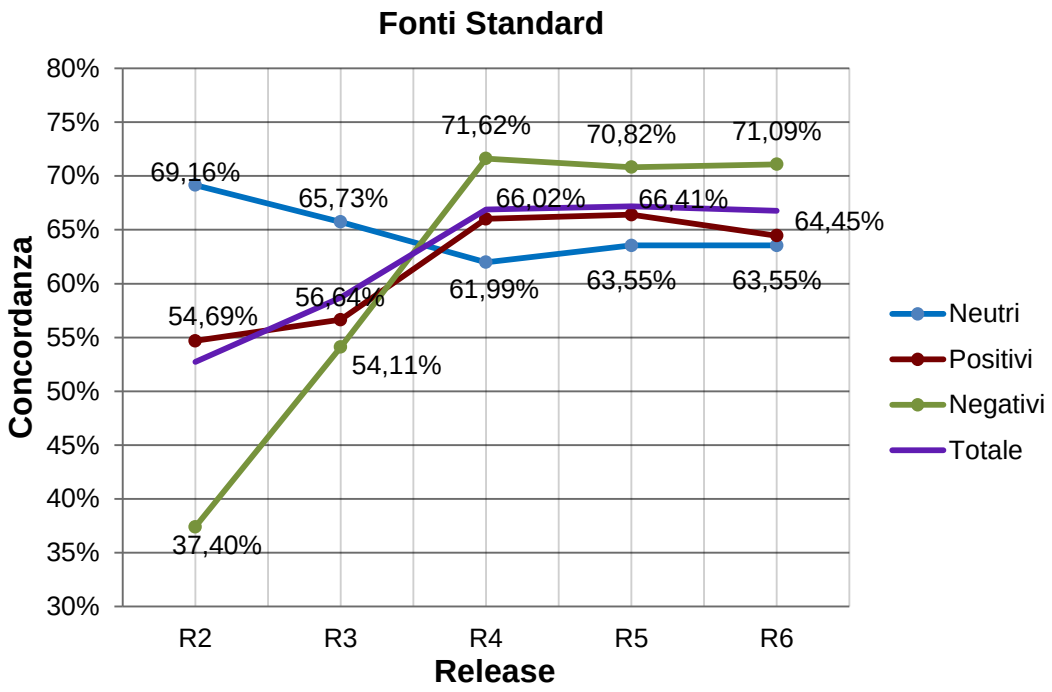

Figura 6.3 Evoluzione dell'efficacia per tipologia di polarizzazione, per le fonti qualificate.

L'arricchimento della base di conoscenza, ha però portato ad una progressiva crescita dell'efficacia dell'individuazione delle due polarizzazioni (positiva e negativa), che ha generato, tra le tre diverse tipologie, uno scarto finale molto ridotto (tutte comprese in un range tra il 64% e il 71%). Un comportamento analogo lo evidenziamo anche per le fonti social, in cui notiamo una progressiva convergenza delle tre diverse valorizzazioni del sentiment, verso un preciso range di efficacia (dal 62,5% al 65,5%).

I testi che hanno maggiormente beneficiato delle attività di verticalizzazione sono senza dubbio quelli con polarizzazione negativa. Grazie all'inserimento nella base di conoscenza di circa 500 concetti con accezione negativa per l'ambito di ascolto (contro solo 170 concetti con accezione positiva), abbiamo ottenuto un miglioramento dell'efficacia molto marcato, fino a raggiungere il 71% circa nelle fonti standard e 62,5% per quelle social.

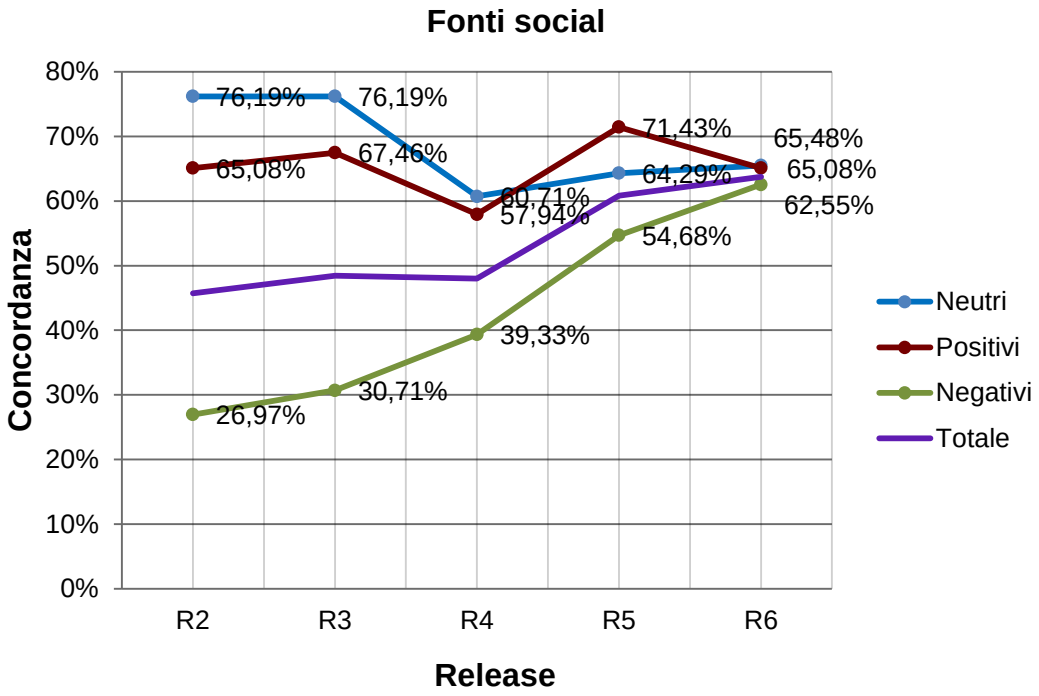

Figura 6.4 Evoluzione dell'efficacia, per tipologia di polarizzazione, per le fonti social.

Dedichiamo lo spazio finale di questa analisi dei risultati, alla prova finale effettuata sul Testset. Questo dataset, lo ricordiamo, è composto da 300 frasi di testi di tipologia standard recuperati dal crawler nel periodo di fine gennaio 2013, sui cui non è stato effettuata nessuna attività di traning. La release, della base di conoscenza, impiegata è l'ultima realizzata (release 6) che fa tutt'ora parte del motore di analisi del testo. Dalla tabella 6.13, possiamo notare un netto calo dell'efficacia rispetto al dataset di training. Però se osserviamo meglio le percentuali di efficacia per tipologia di difficoltà possiamo constatare come il maggior decremento, lo si avverte per i testi etichettati come difficili (-8,49%), mentre per quelli facili, la diminuzione è del tutto fisiologica (-2,87%).

Dataset	Difficoltà	Majority	Concordanza (Motore, MJ)	% Efficacia	% Risp. Standard
Testset Finale	**Difficile**	161	74	**45,96**	**-8,49**
	Facile	135	94	**69,63**	**-2,87**
	Totale	296	168	56,76	-10,01

Tabella 6.13 Risultati del test finale sull'efficacia dell'analisi del sentiment.

Infatti, se ricordiamo, le composizioni dei due dataset, che riportiamo per comodità in figura 6.5, sono fortemente diverse. Per quello di training abbiamo una netta prevalenza di testi di tipo facile (circa il 68%), che in quello di testing risultano essere invece la minoranza (circa il 45%). Questo ci impone, per poter comparare equamente i risultati dei due dataset, di bilanciare le proporzioni del dataset di testing.

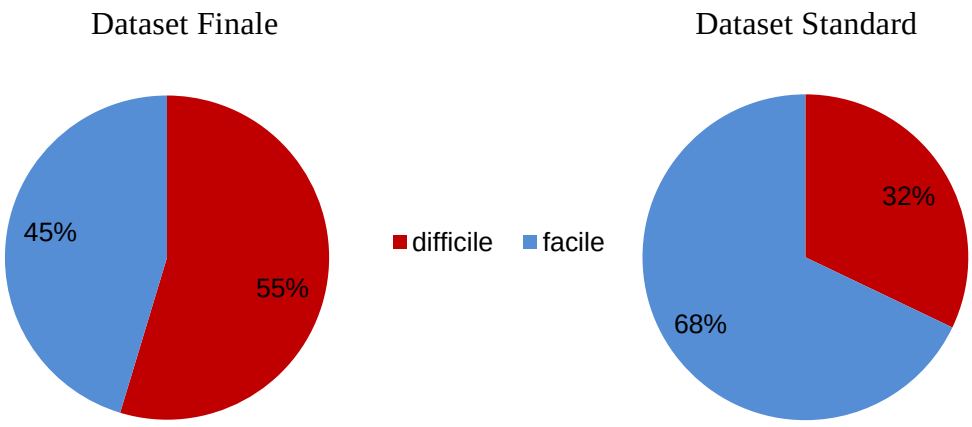

Figura 6.5 Differenza nella composizione del dataset di training e quello di testing

Se supponiamo, infatti, di poter disporre per il dataset di testing, su di una percentuale di testi facili del 68% e di quelli difficili del 32% (le stesse proporzioni del dataset di training), l'efficacia complessiva risulterebbe del 62,03%, solamente

di circa 4 punti percentuali inferiore del dato ottenuto nel dataset di training, quota che consideriamo del tutto accettabile.

Nel complesso le attività di arricchimento della base di conoscenza per l'ambito politico, ha generato un sensibile miglioramento generale delle performance del sistema, nel rilevare correttamente il sentiment dell'opinione espressa nel testo. Per poter accrescere maggiormente le capacità su dataset reali, occorre analizzare e studiare ulteriori testi ed effettuare ulteriori iterazioni di traning, necessarie per comprendere tutte quelle casistiche non individuate nei training-set fin qui analizzati. Infatti l'approccio statistico, impiegato dallo strumento che sta alla base di questo sistema, per apprendere un determinato costrutto o locuzione, richiede che sia compreso nella base di conoscenza. Rispetto ad altri approcci più linguistici, come da previsione, SPSS Text Analytics, ha richiesto uno sforzo maggiore nella verticalizzazione del base di conoscenza. Tuttavia ha confermato ottime doti di versatilità sulle diverse tipologie di testi, infatti le prestazioni raggiunte sono risultate piuttosto soddisfacenti in tutti gli scenari sperimentati, anche in quello social, notoriamente più ostico e complesso.

Conclusioni

Abbiamo constatato come la realizzazione di un sistema di Social Business Intelligence, seppure ad un livello prototipale, richieda un percorso di lavoro lungo e laborioso, che coinvolge tante discipline diverse al suo interno. I temi affrontati sono stati tanti, partendo da quelli più tecnologici, come il reperimento dei dati dal web, la progettazione e implementazione dell'architettura e la parametrizzazione degli strumenti di analisi del testo. Senza dimenticare le esperienze maturate, per così dire, "meno tecniche", come lo studio approfondito delle caratteristiche dell'ambito di ascolto (la politica italiana) e delle peculiarità linguistiche del linguaggio scritto per le diverse fonti web previste, come i Social Network o le fonti qualificate standard (quotidiani web o aggregatori di news on-line). Allo stato dell'arte, le competenze che si incrociano e devono cooperare per la buona riuscita di un progetto ambizioso come l'inserimento di un sistema di Social BI in azienda, sono molteplici e vanno, come già anticipato, dai canonici aspetti architetturali a quelli linguistici, in aggiunta alle conoscenze richieste sul dominio di ascolto. Probabilmente la trasversalità del problema è una delle maggiori macro-criticità della realizzazione di un sistema di Social BI che estenda le piattaforme di BI già presenti in azienda. In primo luogo il dato va recuperato al di fuori dei consueti confini aziendali (nell'intero Web) e questo introduce un primo livello di

complessità, ulteriore rispetto ai tradizionali sistemi di BI. Si consideri anche la tipologia destrutturata dei dati che si intendono analizzare, che rende necessara quindi l'adozione di strumenti e tecnologie apposite, non ancora diffuse nei sistemi tradizionali aziendali.

Tutti gli aspetti sopra considerati, sono stati oggetto di questo studio di tesi grazie alle esperienze maturate sulla metodologia sperimentata e sugli strumenti impiegati, lungo tutto il percorso che ha portato alla realizzazione del prototipo finale del sistema. Durante la realizzazione del sistema, sono state riscontrate alcune problematiche relative, ad esempio, al reperimento dei dati dal web o alle scadenti capacità iniziali della base di conoscenza italiana, inclusa nello strumento di analisi del testo. Anche grazie alle esperienze maturate su questi aspetti del sistema, sono state evidenziate alcune delle criticità più importanti del processo di realizzazione di un sistema di Social BI. Innanzitutto per il conseguimento di risultati soddisfacenti è indispensabile che la qualità dei dati recuperati sia elevata, il che si traduce in: documenti in-target (ridurre al minimo i documenti non inerenti al dominio di ascolto) e puliti (rimuovere possibili fonti di errore alla radice). Altre criticità appurate sono relative al sottosistema di analisi del testo, che deve essere di accurata implementazione e soprattutto orientato alla lingua e al tipo di linguaggio utilizzato nei documenti. Infatti una delle problematicità riscontrate nel caso in esame, sono state le lacune iniziali della base di conoscenza per la lingua italiana. Probabilmente a causa della natura anglosassone dello strumento impiegato, il motore di analisi iniziale è risultato carente sotto tanti aspetti. Uno degli aspetti principali è stata la mancanza di una corretta lemmatizzazione del testo, di fondamentale importanza per tutte le successive fasi del processo di analisi. Inoltre, anche il tipo di linguaggio ha influito pesantemente sui risultati conseguibili dal sistema di analisi; è bene quindi valutare ed esaminare con attenzione le fonti da cui si intendono recuperare i documenti, soprattutto quelle dei Social Network, che destano probabilmente il maggior interesse in ambito aziendale.

Non ultime, le criticità affrontate durante il processo di verticalizzazione del sistema di analisi, sull'ambito di ascolto previsto nel caso di studio in esame. Abbiamo constatato come l'attuazione delle modifiche da apportare per infondere al motore conoscenza sul dominio, fosse un procedimento molto influente sulla qualità dei risultati conseguibili e di conseguenza non privo di rischi. Si è resa necessaria l'adozione di una metodologia che comprendesse una prima fase di testing della versione corrente del motore di analisi e successivamente una seconda fase, di analisi e studio approfondito delle categorie di errori riscontrati. E' stato ideato questo processo, al fine di conseguire una maggiore comprensione sul dominio. Quella stessa comprensione che, con la fase finale di effettiva modifica della base di conoscenza, si tenta di infondere al motore di analisi. I rischi che si potevano correre senza l'impiego di una metodologia precisa non erano pochi, primo fra tutti il rischio di introdurre, con modifiche non ben ponderate, effetti collaterali che potessero limitare, se non annullare totalmente, i benefici attesi dalla nuova versione del motore di analisi.

Per concludere questa visione d'insieme delle esperienze maturate in questo studio di tesi, esaminiamo nel complesso i test effettuati sulla funzionalità di sentiment analysis del sistema. I test ci hanno mostrato che, mediante l'impiego di un approccio statistico, nello specifico del software SPSS Text Analytics, si raggiungono complessivamente prestazioni soddisfacenti in tutti gli scenari sperimentati. E' giusto riportare che, considerate le già citate caratteristiche dei diversi linguaggi, l'ambito dei Social Network si è dimostrato più ostico rispetto altri scenari, come per'altro ci attendavamo. L'evoluzione della curva di miglioramento e il test finale sull'efficacia, hanno evidenziato, invece, come l'approccio impiegato, richieda molte iterazioni di arricchimento della base di conoscenza per comprendere al meglio le caratteristiche del dominio di ascolto. Si sono rese necessarie più iterazioni per infondere, al motore di analisi, una completa comprensione e la capacità di generalizzazione sul dominio.

Infine, valutato il prototipo di sistema realizzato, quello che si preannuncia come sviluppo futuro, è la concretizzazione di una struttura completa di Social BI, che costituisca una piattaforma integrata, con i sistemi di BI già presenti. I passi successivi per realizzazione della piattaforma prevedono ad esempio la modellazione multidimensionale attraverso l'implementazione di un data warehouse alimentato dal database operazionale già presente. Il traguardo finale, è quello di ottenere una piattaforma completa che permetta di estrarre tutto il valore informativo contenuto nei dati estratti, mettendo a disposizione dell'utente di business, apposite funzioni avanzate di front-end. Queste funzioni, devono essere pensate specificamente per social media e devono essere appositamente progettate sulla base delle macro-funzionalità fornite dal sistema. Ad esempio per le funzionalità di individuazione e osservazione dei topic di interesse (New Topic, Trend Topic e Top Topic), possiamo prevedere l'implementazione di apposite dashboard che diano evidenza del numero di occorrenze catturate nel web, analizzandone l'andamento nel tempo. Mentre per la funzioni relative alla co-occorrenza possiamo pensare alla definizione di una dashboard che, mediante un grafo relazionale, rappresenti per un topic di interesse, i concetti ad esso collegati. Infatti non dobbiamo sottovalutare l'importanza della progettazione e della realizzazione di appositi tools di front-end, se vogliamo rendere vantaggioso al business aziendale, l'enorme valore informativo contenuto in questi dati.

Bibliografia

Jacques Bughin, Angela Hung Byers, and Michael Chui (2011). *How social technologies are extending the organization*, http://www.mckinseyquarterly.com/High_Tech/Strategy_Analysis/

H Cunningham, K Humphreys, R Gaizauskas, Yorick Wilks (1997). *Software infrastructure for natural language processing.* http://dl.acm.org/citation.cfm?id=974592

A.Gliozzo, C.Strapparava (2009). *Semantic Domain in Computational Linguistics*

Seth Grimes, Beye Network (2010), *Social Media and the Enterprise Business Intelligence/Analytics Connection.* http://altaplana.com/SMI_v4.pdf

Dion Hinchcliffe (2013), *Realizing social business: Enterprise 2.0 success stories.* http://www.zdnet.com/blog/hinchcliffe/realizing-social-business-enterprise-2-0-success-stories/1908

IBM. IBM SPSS (2011). *Manuale dell'utente di IBM SPSS Modeler 14.2*. http://www.ibm.com/spss

IBM. IBM SPSS (2011). *Text Analytics 14.2 User's Guide*. http://www.ibm.com/spss

Tom Malloy (2012). *Revolutionizing Digital Marketing with Big Data*. http://www.cikm2012.org/doc/cikm2012_Malloy.pdf

Talend Inc (2006) *Talend Open Studio* http://www.talend.com/products/data-integration

Ian H. Witten (2012). *Text Mining. http://www.cs.waikato.ac.nz/~ihw/papers/04-IHW-Textmining.pdf*

Eros Zanchetta, Marco Baroni (2005). *Morph-it! A free corpus-based morphological resource for the Italian language, proceedings of Corpus Linguistics* University of Birmingham, Birmingham, UK.