Social Media Analytics

Jacek Filipczuk

Gennaio 2015

Indice

Capitolo 1

Introduzione

Social Media è un termine usato per descrivere servizi basati sul web e tecnologie mobile, altamente interattivi, che permettono agli individui di creare profili pubblici, o semi-pubblici, presso un dominio in modo che possano comunicare e connettersi con altri individui, in generale nascono vere e proprie comunità che possono condividere, creare, discutere e modificare contenuti generati dagli utenti stessi. [15].
I Social Media favoriscono enormemente le interazioni online, di condivisione dei contenuti, della diffusione di affermazioni, approcci, valutazioni, influenze, osservazioni, sentimenti e opinioni degli individui, tutte espresse sotto forma di testo in recensioni, blogs, discussioni o notizie.
Questa grandissima fonte di informazioni viene sfruttata dalle grandi organizzazioni, dalle celebrità, perfino dal governo per acquisire conoscenza sui gusti, sulle preferenze, sulle idee più diffuse tra la popolazione.
Le informazioni fornite dai Social Media sono in quantità enormi e questo provoca non poche difficoltà nella procedura della loro raccolta e analisi, difficoltà che si possono superare quasi del tutto tramite un utilizzo accurato delle tecniche di Data Mining.
Questo lavoro si pone come obiettivo quello di individuare quali sono, al momento, i problemi e le sfide di maggior interesse nell'area dei Social Media Analytics e quali sono i modelli più usati per la rappresentazione scientifica dei Social Media. Si vuole, anche, descrivere le tecniche più usate degli ultimi anni nella risoluzione delle sfide individuate ed effettuare un'analisi delle soluzioni esistenti per la trattazione dei Big Data derivanti da Social Media.

Capitolo 2

Social Media

I Social Media sono degli strumenti basati sul web che permettono alle persone di creare, condividere o scambiare informazioni, idee, foto o video all'interno di comunità virtuali e Social Network. La definizione formale dei Social Media è la seguente: *Sono un gruppo di applicazioni basate su internet, costruite sulle fondamenta ideologiche e tecniche del Web 2.0, e che permettono la creazione e lo scambio di contenuti generati dagli utenti.* Estendendo ulteriormente tale definizione si può affermare che i Social Media dipendono da tecnologie mobile e web-based per creare piattaforme altamente interattive grazie alle quali le persone e le comunità possono condividere, co-creare, discutere e modificare contenuti generati dagli utenti. Essi hanno introdotto un cambiamento sostanziale nel modo di comunicare tra le organizzazioni, tra le comunità e tra gli individui, ed è proprio questo cambiamento che ha dato inizio allo studio di metodi e tecniche per l'estrapolazione di informazioni dai Social Media.
I Social Media si differiscono dai media tradizionali in tanti modi, e in particolare sotto il punto di vista della qualità, della raggiungibilità, della usabilità, e della persistenza di questo tipo di media. Secondo uno studio condotto da Nielsen, gli utenti di internet spendono la maggior parte del loro tempo sui siti di Social Media che su altri tipi di siti. La maggior parte della popolazione, quindi, usa i Social Media in una qualche loro forma. Per poter capire la quantità di informazioni generate in questa maniera basti considerare le informazioni viste dagli utenti in un giorno tipico di utilizzo. Nell'ottobre del 2012, gli utenti attivi di Facebook furono circa un bilione, e spendevano circa 20.000 anni online ogni giorno. Nello stesso periodo, YouTube ha riportato più di un bilione di visualizzazioni dei video, mentre i 140 milioni di utenti attivi di Twitter hanno mandato più di 340 milioni di tweet. Tutti questi non sono dei semplici utenti passivi. Dalle analisi di YouTube risulta che ogni settimana più di 100 milioni di utenti effettuano una qualche azione sociale, come scrivere un commento o esprimere una preferenza per un contenuto. Queste azioni si sono duplicate dal 2012 al 2013.

Facebook integra le azioni sociali degli utenti nei suoi messaggi pubblicitari, per esempio permette agli utenti di vedere quali prodotti pubblicizzati sono stati di gradimento dei propri amici. Lo stesso metodo viene usato anche da Twitter che, tramite gli hashtag, offre agli utenti un modo rapido e semplice per esprimere le proprie preferenze, opinioni e altro su un determinato oggetto o fatto.
Le informazioni scambiate sui Social Media sono di notevole importanza per le organizzazioni e le compagnie multinazionali, di seguito viene mostrato un esempio di come i Social Media possano influenzare una compagnia.
Nel 2008, la United Airlines ha rotto la chitarra di Dave Carroll. Questo non era il primo caso in cui uno strumento musicale è stato rotto durante un volo, ma è stata la prima volta in cui il proprietario dello strumento ha registrato un video musicale dell'accaduto e l'ha pubblicato su YouTube. Il video diventò in breve tempo virale ed è stato visto più di nove milioni di volte. Inoltre venne citato sul sito del Times, e in televisione nel programma CNN Situation Room. Tutto questo portò a una crisi nelle relazioni della compagnia United con i suoi clienti, mentre l'accaduto è stato incoraggiato dal pubblico mondiale che capiva troppo bene la frustrazione di avere a che fare con compagnie aventi servizi scadenti. La United non rispose a questo accaduto, mostrando come un'agenzia può risultare impreparata a gestire conversazioni o interazioni sui Social Media.

Tra i Social Media più popolari, nel 2014, vanno menzionati YouTube, Twitter, Instagram e LinkedIn.

2.1 Social Media Analytics

I Social Media sono diventati davvero importanti per le grandi compagnie, per questo motivo vengono sempre più studiati e analizzati nuovi metodi per sfruttare le informazioni contenute in essi. I Social Media Analytics sono appunto l'insieme dei metodi e delle tecniche usate per sfruttare nel migliore dei modi le informazioni contenute nei Social Media.
I Social Media Analytics possono essere considerati come un processo composto da tre fasi: raccolta, comprensione, presentazione [6], ovvero il CUP Framework (Capture, Understand,Present). La raccolta consiste nell'ottenere dati rilevanti dai Social Media monitorando, o ascoltando, svariate sorgenti, archiviando dati di rilievo e estraendo informazioni pertinenti. Tutto questo processo può essere eseguito internamente da una compagnia o da terze parti. Purtroppo non tutti i dati ottenuti sono utili. La fase di comprensione consiste nel selezionare dati importanti e eliminare contemporaneamente dati rumorosi e di scarsa qualità, sfruttando vari metodi avanzati di analisi per avere, anche, una comprensione maggiore dei dati raccolti. Infine la presen-

tazione consiste nel mostrare i dati in maniera significativa. Bisogna notare, anche, che vi è una sovrapposizione delle fasi fra loro. La fase di comprensione crea dei modelli che possono aiutare nella fase di raccolta dati, così come la fase di presentazione aumenta la conoscenza degli utilizzatori sui dati, una conoscenza che può essere applicata per migliorare ulteriormente sia la fase di comprensione che quella di presentazione. Le fasi di questo processo avvengono in maniera iterativa, se ad esempio i modelli della fase di comprensione falliscono nell'individuare modelli utili, allora possono essere migliorati raccogliendo una quantità maggiore di dati per aumentare le loro capacità predittive. In maniera analoga, se i dati visualizzati non risultano interessanti, potrebbe essere necessario tornare nella fase di comprensione o anche di raccolta per aggiustare i parametri usati nell'analisi. Un sistema per l'analisi dei Social Media data può attraversare numerose iterazioni di questo processo prima di diventare realmente utile.

2.1.1 Raccolta Dati

Per un azienda che sfrutta i Social Media Analytics, la fase di raccolta aiuta a identificare le conversazioni sulle piattaforme dei Social Media che sono inerenti le attività e gli interessi aziendali. Questa operazione viene eseguita raccogliendo enormi quantità di dati provenienti da migliaia sorgenti differenti tramite svariati metodi, come ad esempio il crawling. Per preparare il data set così raccolto per la fase di comprensione, possono essere necessari vari passaggi di preprocessing, che includono il data modeling, l'estrazione di feature e altre operazioni di sintassi e semantica che aiutano nell'analisi. Vengono anche estratte le informazioni sull'azienda, sugli utenti, sugli eventi, insieme ai commenti e alle altre informazioni per una successiva analisi e modellazione. La fase di raccolta deve riuscire a raccogliere informazioni da tutte le sorgenti possibili ma porre particolare attenzione alle sorgenti di maggior rilievo e autorevolezza, per una comprensione più accurata.

2.1.2 Comprensione Dati

Nel momento in cui un'azienda riesce a raccogliere le conversazioni degli utenti legate ai proprio prodotti e al proprio operato, deve riuscire a stabilire il loro significato e a generare una metrica utile nel processo di decisione, questa è la fase di comprensione dei dati. Siccome la fase di raccolta consiste nel raccogliere dati da sorgenti differenti, una grande porzione di questi dati potrebbe essere affetta da rumore, cioè essere inutile, e quindi deve essere eliminata prima di passare ad un'analisi significativa. Per effettuare questa procedura di pulizia dei dati rumorosi si può usare un semplice classificatore testuale o classificatori più complessi addestrati su dati etichettati. Dedurre un significato dai dati ripuliti dal rumore può implicare l'uso di metodi statistici provenienti dal data mining, dal natural language processing, dal

machine translation e dal network analysis. La fase di comprensione riesce a fornire informazioni sui sentimenti dei clienti, ad esempio quali sentimenti essi hanno nei confronti dell'azienda e dei suoi prodotti, e sul loro comportamento, ovvero quanto sia probabile che i clienti acquistino un determinato prodotto dopo una campagna pubblicitaria. In questa fase possono essere create numerose metriche utili come gli interessi dei clienti, i loro dubbi e la rete delle loro relazioni. Bisogna notare che la fase di comprensione è il cuore dell'intero processo di analisi, il suo risultato avrà un grosso impatto sulla fase di presentazione e di, conseguenza, sulle future decisioni che l'azienda potrebbe prendere.

2.1.3 Presentazione Dati

In questa ultima fase, i risultati provenienti dalle diverse analisi effettuate, vengono riassunti, valutati e mostrati agli utenti in un formato di facile comprensione. Le tecniche di visualizzazione possono essere usate per presentare informazioni utili, mentre tecniche più sofisticate permettono una presentazione dei dati personalizzata per ogni cliente, riescono a fornire un significato a grandi quantità di dati e mostrano modelli e relazioni tra dati più significative per gli esseri umani che per le macchine.

Capitolo 3

Social Media Challenges

Il fenomeno dei Social Media ha provocato un notevole interesse da parte dei ricercatori di numerose discipline, il motivo di questo interesse viene ampiamente spiegato in [12] e brevemente descritto di seguito. Per la maggior parte di loro l'aspetto cruciale dei dati provenienti dai Social Media è che riguardando *il sociale*. Con questo termine si vuole indicare la natura delle relazioni e dei comportamenti che sussistono tra gli individui. Da questo punto di vista i Social Media offrono una quantità enorme di dati che hanno delle peculiarità. Nel lavoro [25] viene mostrata l'importanza dei Social Media Analytics nello sfruttare questi dati. I dati raccolti dai Social Media riguardano tutta la popolazione e non dei sotto-insiemi, le informazioni sono dinamiche, catturate in tempo reale o in un certo periodo, inoltre riguardano ciò che gli individui dicono e fanno nella vita di tutti i giorni e non quello che dicono in risposta a un questionario o a un'intervista eseguita da un ricercatore. Infine i dati hanno una natura digitale, offrendosi in maniera diretta a tutte quelle tecniche di data mining e linking usate nei casi di studio. D'altra parte esistono numerose e considerevoli sfide da affrontare prima di poter usufruire dei dati offerti dai Social Media. In particolare la più grande sfida è quella di sviluppare sorgenti di dati e metodologie che permettono di interrogare e interpretare i Social Media Data in maniera da poter effettuare richieste complesse di natura sociale. Questa sfida viene affrontata di solito da due punti di vista, quello delle Scienze Sociali e quello delle Scienze Computazionali. Il primo possiede esperienza nel formulare domande di ricerca sociale, ma ha lacune nel trattare i dati dei Social Media. Il secondo invece possiede numerose tecniche e metodi efficaci nella trattazione dei dati forniti dai Social Media, ma ha lacune nel campo teorico riguardante la formulazione di quesiti sociali.
Approfondendo le problematiche riguardanti i Social Media ci si vuole concentrare su quali sono le sfide maggiormente affrontate nel campo dei Social Media Analytics [23]. In particolare le sfide scelte per la trattazione sono le seguenti:

- Topic Detection
- Trend Tracking
- Opinion Mining

3.1 Topic Detection

Nel primo studio sul Topic Detection and Tracking [1], condotto durante il 1996 e il 1997, la nozione di un topic era limitata a essere un evento, e con questo si voleva intendere un qualcosa accaduto durante uno specifico lasso di tempo e in uno specifico luogo. Ad esempio l'eruzione del Monte Pinatubo avvenuta il 15 giugno 1991 è considerato essere un evento, mentre una generica eruzione vulcanica non lo è. Gli eventi possono essere inaspettati, come un incidente aereo, o attesi, come le elezioni politiche. Nello studio successivo sul Topic Detection and Tracking la definizione di un topic è stata ampliata per includere, oltre all'evento scatenante, anche altri eventi e attività che sono direttamente collegate a esso. La definizione del topic è diventata la seguente: *Un topic è definito come un evento o attività scatenante, insieme a tutti gli eventi e le attività direttamente correlate a esso.* Ovviamente bisogna verificare quali eventi o attività sono davvero collegate direttamente all'evento scatenante e limitare l'inclusione solo a essi. Un esempio di Topic Detection viene descritto in dettaglio nel lavoro [10], in cui viene proposta l'estrazione di topic dalle news di *Twitter*.

3.2 Trend Tracking

Per prima cosa viene data una veloce definizione di che cosa sia un trend. Un trend è un concetto che può avere un impatto reale sulla vita di tutti i giorni delle persone e più in generale sulla società. Un trend spesso entra a far parte della corrente principale seguita dalla società sulla quale può avere un impatto su più livelli. A volte un trend nasce come una pratica o un costume locale per poi espandersi velocemente e all'improvviso a gran parte della società. Il Trend Tracking consiste nel monitorare il flusso di informazioni in arrivo che sono collegati a uno specifico trend e seguirne il percorso [11]. Lo scopo principale è quello di identificare e seguire eventi provenienti da sorgenti multiple. Per colpa di Internet, le informazioni correlate a un topic sono spesso disperse su periodi di tempo e luoghi diversi, il Trend Tracking colleziona queste informazioni e le unisce per facilitarne l'uso e la comprensione. Esistono molte aree dell'industria dove la Trend Tracking può essere sfruttata. Può essere usata per avvertire le compagnie quando un prodotto competitivo compare nelle news, permettendo alle compagnie di reagire tempestivamente. Un altro uso è quello in cui una compagnia vuole tenere traccia del proprio andamento e di quello dei propri prodotti. Il Trend

Tracking può anche essere usato nel campo della medicina per individuare nuovi trattamenti che i pazienti applicano. Per ulteriori approfondimenti si rimanda ai lavori [2], [20] e [16] dove sono state proposte alcune applicazioni e usi di Trand Tracking.

3.3 Opinion Mining

La diffusione delle informazioni digitali ha dato il via alla nascita di sistemi automatizzati. Questi sistemi hanno come scopo quello di rendere automatiche alcune operazioni come l'analisi o la classificazione dei dati. Il Text Mining è un metodo interdisciplinare usato in campi differenti come il Machine Learning, Information Retrieval, Statistica e Computational Linguistic. Web Mining è una sub-disciplina del Text Mining usata per lavorare con dati del web semi-strutturati. Opinion Mining, chiamata anche Sentiment Analysis, è il processo di trovare opinioni di utenti su un particolare topic o un particolare prodotto o un problema [7]. Lo scopo dell'Opinion Mining è rendere capace il computer di riconoscere ed esprimere emozioni. Un esempio dell'applicazione dell'*Opinion Mining* si ha nel lavoro [17], dove si cerca di analizzare dati provenienti da Twitter, considerando anche l'influenza delle opinioni presenti in questi dati.

Capitolo 4

Social Media Tools

In questo capitolo si vuole presentare gli strumenti e le tecniche maggiormente usate nell'affrontare le sfide descritte nel capitolo precedente.

4.1 Topic Detection Tools

Lo scopo del Topic Detection, come già detto in precedenza, è quello di trovare eventi di rilievo tra la marea di informazioni che vengono generate ogni istante. In questo tipo di sfida,di solito, non vi è presente una lista dei topic a cui fare riferimento, di conseguenza il problema di identificare e caratterizzare un topic è parte integrante del processo e non è possibile fare affidamento su un training set o altre forme esterne di conoscenza ma solo sfruttare le informazioni contenute nella collezione dei dati da analizzare.
Topic Detection è un attività che viene eseguita su un flusso di documenti in continua evoluzione, per questo motivo necessita di soluzioni basate su algoritmi incrementali. Tra le soluzioni maggiormente utilizzate in questo ambito vi è quella del clustering, e in particolare tecniche che usano reti neurali o il k-means.

Reti Neurali. Per poter usare questa tecnica i documenti da analizzare devono essere espressi secondo un formato standard, ogni documento deve essere rappresentato come un insieme di parole chiave usate come features. Per ridurre la complessità dell'analisi vengono usati tool per individuare la radice delle parole e usare quella per le operazioni. Le reti neurali usate successivamente su questo tipo di input sono spesso delle reti self-organizing. In particolare una classe di queste reti sono le ART Networks che sta per Adaptive Resonance Theory, introdotte da Grossberg nel 1976. Esistono diversi tipi di reti ART, le più interessanti per la Topic Detection sono le Fuzzy ART. Questo particolare tipo di reti unisce la teoria fuzzy insieme alle reti ART, le Fuzzy ART riescono così a distinguere in maniera stabile le categorie sia quando l'input è di tipo analogico sia quando è binario[21]. L'algoritmo che sfrutta queste reti neurali risulta molto efficiente nel gestire

grossi insiemi di dati. Risulta, inoltre, molto robusto anche quando i dati sono affetti da rumore. D'altra parte presenta dei svantaggi. Ad esempio il numero di cluster deve essere specificato a priori. La determinazione del numero di cluster, quindi, risulta essere un'operazione non banale, dato anche il fatto che le caratteristiche dei dati non sono note a priori. Inoltre per poter usare questa tecnica è necessario che tutti gli esempi presenti nel dataset abbiano tutte le proprie caratteristiche avvalorate, un prerequisito che spesso non può essere soddisfatto.

k-Means. Il processo di clustering consiste nel raccogliere documenti, termini o altri oggetti in gruppi sulla base di un qualche criterio di similitudine. Il criterio di similitudine usato più spesso è una funzione di distanza calcolata sulla distribuzione di frequenza dei termini chiave in esame. La distribuzione di frequenza viene calcolata contando le occorrenze delle parole chiave all'interno dei documenti in analisi. Prendendo in esame questo approccio, un topic non è altro che un cluster di parole chiave scelte opportunamente. Questa tecnica sfrutta in particolare il Bisecting k-means algorithm e, informalmente può essere descritta nel seguente modo. All'inizio vengono scelti due elementi aventi la distanza massima, e sono usati come centri di due cluster. Successivamente tutti gli altri elementi vengono assegnati a uno dei due cluster che risulta più vicino. Una volta che tutti gli elementi vengono assegnati a un cluster, i centri di questi ultimi vengono ricalcolati. I centri trovati vengono usati come due nuovi cluster e il processo viene ripetuto fino a ottenere una certa precisione. Se le dimensioni di un cluster sono eccessive, la procedura viene applicata in maniera ricorsiva al cluster troppo grande. Alla fine l'algoritmo fornisce un albero binario di cluster[24]. L'algoritmo che sfrutta il *k-Means* risulta vantaggioso per la sua semplicità, la sua velocità di esecuzione su dati aventi dimensionalità bassa e sulla sua capacità di trovare sub-cluster puri nel caso in cui il parametro k viene specificato molto alto. Tra gli svantaggi di questa tecnica vi sono, innanzitutto, l'incapacità di trattare dati non globulari di dimensioni e densità differenti, di identificare gli outliers e la necessità di specificare il parametro k a priori. Inoltre partizioni di partenza differenti possono portare a cluster finali differenti, per questo motivo gli algoritmi che sfruttano questa tecnica vengono eseguiti più volte.

Per quanto riguarda gli strumenti maggiormente usati nella Topic Detection, di particolare rilievo sono i seguenti:

- *MALLET*:che sta per MAchine Learning for LanguagE Toolkit, è uno strumento per la modellazione dei topic. Questo strumento include un'implementazione estremamente veloce e scalabile del Gibbs Sampling, metodi efficienti per l'ottimizzazione di parametri inerenti i topic di un documento, e altri strumenti per inferire i topic di nuovi documenti dato un modello già addestrato.

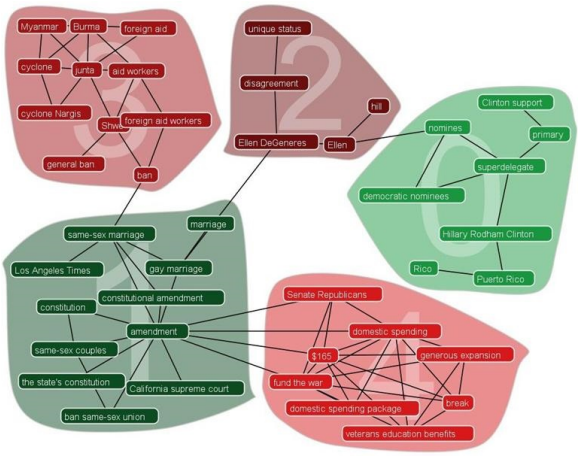

Figura 4.1: Esempio di uso di Keygraph

- *AlchemyAPI*: è uno strumento capace di estrapolare parole chiave dei topic da pagine HTML, testo o contenuti web-based. Esso si basa su un sofisticato algoritmo statistico combinato a tecniche di processing del linguaggio naturale per analizzare i dati, estrapolare parole chiave che possono essere usate per indicizzare il contenuto e altre funzioni aggiuntive. L'estrazione di parole chiave è supportata in una dozzina di linguaggi differenti, permettendo di identificare e categorizzare anche contenuti non in inglese.

- *Apache Mahout*: è uno progetto che ha come obiettivo quello di creare una libreria di machine learning scalabile. In questo momento Mahout supporta tre operazioni principali tra cui il Clustering usato per ragruppare documenti in insiemi correlati da topic simili. Questo strumento viene, quindi, usato anche per la Topic Detection.

Oltre a questi, vi sono altri due strumenti che risultano molto interessanti per eseguire il compito del Topic Detection, ovvero: *KeyGraph* e *TextAlytics*. Questi verranno descritti di seguito.

KeyGraph

La Topic Detection in collezioni di dati di grandi dimensioni e affette da rumore necessita strumenti accurati ma allo stesso tempo scalabili. KeyGraph, mostrato in figura 4.1, è uno strumento efficiente che sfrutta nel suo compito le co-occorrenze delle parole chiave. Esso riesce a filtrare documenti irrilevanti e identificare eventi in grandi e rumorose collezioni di social media, inoltre risulta migliore nei termini di prestazioni rispetto alle soluzioni più comunemente usate[22].

TextAlytics Topic Extraction

TextAlytics è un progetto con diversi strumenti, tra questi vi è Topic Extraction. Esso ha come obiettivo quello di estrarre elementi differenti presenti nelle fonti di informazioni. Questo processo di identificazione viene eseguito combinando un certo numero di complesse tecniche per il processing del linguaggio naturale. Queste tecniche forniscono un'analisi morfologica sintattica e semantica di un testo che viene sfruttata per identificare tipi differenti di elementi significanti. L'alta configurabilità fornita dalle API permette di aggiustare il comportamento di questo strumento per poter affrontare diversi scenari, come ad esempio poter utilizzare diversi tipi di formati, linguaggi e perfino registri di linguaggio.

4.2 Trend Tracking Tools

Numerose sono le tecniche proposte per il Trend Tracking, di seguito si vuole dare una breve descrizione di quelle più diffuse [18]. Le tecniche sono:

- *Modello Probabilistico*: esistono due modelli fondamentali per comparare una storia a un gruppo di topic correlati. Nel primo ciò che viene calcolato è la probabilità p(T|S) dove S è la storia e T rappresenta il gruppo di storie correlate a un topic. La p(T|S) indica qual'è la probabilità, dato il topic T, che S vi sia correlata. In questo modello per ogni parola analizzata o essa appartiene al topic considerato o viene assegnata a un gruppo chiamato modello generale. Nel secondo modello ciò che viene calcolato è p(SisR|T), che è la probabilità che la storia S sia rilevante per il dato modello di topic [11]. Il vantaggio di usare un *Modello Probabilistico* è la semplicità e la velocità di esecuzione, una volta calcolate le probabilità marginali si ottiene un risultato. Lo svantaggio, invece, consiste proprio nel calcolare quelle probabilità marginali, che non sempre risultano chiare o facili da ottenere.

- *Catene Lessicali*: è un metodo di ragruppare termini lessicalmente correlati in catene lessicali, sfruttando tecniche di Natural Language Processing. Una catena è indipendente dalla struttura grammaticale del testo e risulta essere una lista di parole che rappresentano una porzione compatta della struttura del testo. Prima che una catena sia creata, vengono eliminate tutte le stopwords dal testo e un database di nomi viene usato per il controllo dei sinonimi e dei termini correlati [3],[19]. Il vantaggio principale di questo metodo è che prende in considerazione la correlazione tra le parole analizzate e, di conseguenza, riesce a rilevare in maniera più accurata le relazioni tra il testo analizzato e il trend in analisi. Lo svantaggio è che richiede la presenza di un accurato vocabolario lessicale e di un attenta fase di addestramento.

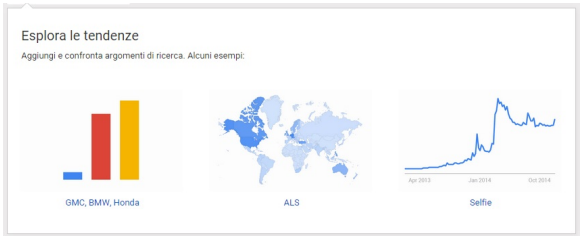

Figura 4.2: Esempio di uso di Google Trends

- *Latent Semantic Analysis SVM*: è un metodo che esegue una profonda analisi delle co-occorrenze delle parole e fornisce un modo per affrontare in maniera automatica il problema dei sinonimi senza la necessità di costruire un thesauro apposito. Questo metodo si basa sull'assunzione che esiste una struttura latente nel modello di utilizzo delle parole nei documenti. I risultati sperimentali hanno dimostrato che LSA-SVM è nettamente superiore ai metodi convenzionali e riduce le percentuali di errore e fallimento nel Topic Tracking. D'altra parte è un metodo che risulta dispendioso sotto il punto di vista computazionale, infatti nella fase di addestramento esso computa matrici di distanza che richiedono grosse quantità di memoria e di potenza computazionale, inoltre, se il problema analizzato è di grandi dimensioni, la fase di addestramento può richiedere tempi lunghi.

Tra gli strumenti maggiormente usati invece due sono quelli che spiccano tra la massa: Google Trends e Topsy. Questi strumenti saranno brevemente descritti di seguito.

Google Trends

Considerando la portata globale di Google, Google Trends, mostrato in figura 4.2, risulta essere uno dei più accurati e potenti strumenti di Trend Tracking. Ha un'interfaccia diretta e semplice che lo rende uno strumento molto intuitivo e conveniente. Google Trends mostra quali sono le grandi cose che accadono in questo momento nel web, basandosi sulle ricerche effettuate dalle persone su Google. Lo strumento offre anche l'opzione di visualizzare su una mappa quali sono gli interessi di una regione o trovare termini correlati a un attività o a un interesse.

Topsy

Topsy è uno strumento largamente usato per scoprire quali sono i Trend attualmente diffusi nel web. Questo strumento offre anche svariate opzioni di analisi, come la ricerca di termini o interessi. La sua caratteristica di spicco, però, è legata a Twitter. Topsy, mostrato in figura 4.3, è forse il

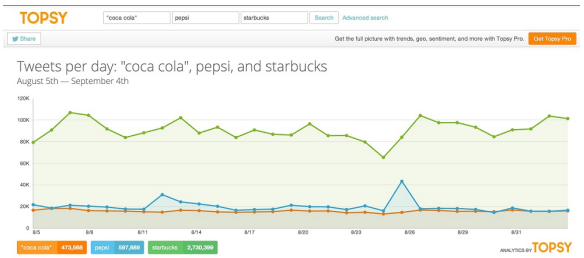

Figura 4.3: Esempio di uso di Topsy Analytics

miglior modo per sapere cosa sta succedendo su Twitter in un particolare momento. Esso è un partner certificato di Twitter e, come tale, ha accesso a tutti i tweet generati fin dalla fondazione di Twitter nel 2006. Questo strumento è la cosa più vicina ad un archivio storico di Twitter.

4.3 Opinion Mining Tools

Le tecniche tutt'oggi maggiormente usate nell'ambito del Data Mining per estrapolare informazioni e conoscenze sono: classificazione, clustering, reti neurali, logica fuzzy, reti Bayesiane, algoritmi genetici e alberi di decisione. In particolare quelle più usate nell'Opinion Mining sono: Supervised Machine Learning, Unsupervised Machine Learning e Case Based Reasoning [7].

Supervised Machine Learning. Questa tecnica di classificazione si basa, in linea di massima, su due fasi, una di addestramento e una di classificazione di nuovi input. Nella prima fase viene usato un data-set contenente, per ogni elemento, l'etichetta della classe di appartenenza. Nella seconda fase, grazie alle conoscenze acquisite durante l'addestramento, nuovi input vengono classificati. *Supervised Machine Learning* è una tecnica che comporta numerose sfide per poter essere usata nel migliore dei modi [14],tra queste si ha: raccogliere il dataset contenente solo le informazioni necessarie, preprocessare e preparare i dati, gestire informazioni mancanti, selezione delle istanze significative, selezione di sottoinsiemi di feature, estrazione delle feature, ecc .

Unsupervised Machine Learning. Al contrario della tecnica precedente, nell'Unsupervised Learning non ci sono target di output specificati per l'input. Le etichette di classe per qualsiasi istanza di input non esistono e per questo motivo devono essere dedotte grazie all'osservazione dei dati. Tra i metodi di Unsupervised Learning il più usato è il Clustering. Questo metodo consiste nel raggruppare oggetti con caratteristiche simili in gruppi chiamati cluster. Di conseguenza oggetti in cluster differenti sono, anch'essi diversi fra loro.

Case Based Reasoning. Questa è un'emergente tecnica di Supervised Learning. Case Based Reasoning [13] risulta essere uno strumento che ragiona e risolve problemi in maniera molto simile a come accadrebbe in uno scenario realistico. Questa tecnica si concentra sulla risoluzione di problemi, non si fonda su regole classiche e la sua base di conoscenza viene rappresentata come una raccolta di eventi passati. Infine le soluzioni vengono depositate in un repository chiamato Knowledge Base o Case Base. Questa tecnica presenta alcuni vantaggi come:

- fornisce soluzioni per ambienti imprevedibili;
- fornisce soluzioni per ambienti mal definiti o aperti;
- notifica gli errori di una soluzione precedente;
- risulta efficiente dal punto di vista temporale.

D'altra parte gli svantaggi di questa tecnica sono:

- crea bias da soluzioni precedenti;
- può non trovare gli insiemi di casi più appropriati;
- segue in maniera cieca i casi precedenti.

Tra gli strumenti più usati nell'analisi dei contenuti generati dagli utenti si trovano:

- **Rewiev Seer Tool**: questo è uno strumento usato per automatizzare il lavoro svolto dai siti di aggregazione. L'approccio utilizzato è quello che sfrutta il classificatore Naive di Bayes per raccogliere opinioni positive e negative. Queste opinioni vengono usate nell'assegnazione di un punteggio a termini estratti come caratteristiche del prodotto. I risultati vengono poi mostrati come delle semplici opinioni sotto forma di sentenze.
- **Red Opal**: permette agli utenti di determinare l'orientamento dell'opinione sui prodotti, basandosi sulle loro caratteristiche. Inoltre questo strumento assegna un punteggio a ogni prodotto in base alle caratteristiche estratte dalle recensioni degli utenti. I risultati vengono visualizzati tramite un interfaccia web.
- **Opinion Observer**: è un sistema di Opinion Mining per l'analisi e il confronto di opinioni su contenuti web generati da utenti. Questo sistema mostra i risultati sotto forma di un grafo e usa il metodo del WordNet Exploring per assegnare una polarità a priori ai contenuti analizzati.

Oltre a questi, vi sono altri due strumenti che risultano i più popolari e usati nell'Opinion Mining: *SentiStrenght* e *Opinion Crawl*. Questi verranno descritti con maggior dettaglio di seguito.

SentiStrength

SentiStrength è uno strumento di Sentiment Analysis che viene sfruttato anche per l'Opinion Mining.
Classificazione dei testi - Per fare in modo che SentiStrength classifichi uno o più testi, è necessario inserire i testi in un file di testo vuoto, in modo che vi sia un testo per linea. Successivamente bisogna selezionare *Analyse All Texts* in *File* dal menu del *Sentiment Strength Analysis* e selezionare il file con i testi. Il risultato sarà una copia del file con, alla fine di ogni riga, una classificazione positiva o negativa del testo. Lo strumento offre anche la possibilità di classificare un singolo testo tramite l'opzione *Analyse One Text* del menu.
Infine questo strumento permette di ottimizzare il peso dei termini usati per il calcolo della polarità di un testo e l'accuratezza di tale calcolo. SentiStrength è stato realizzato come parte del progetto CyberEmotions supportato da EU FP7.

Opinion Crawl

Opinion Crawl, mostrato in figura 4.4, permette agli utenti di stabilire la polarità di un argomento, un evento, una compagnia o un prodotto. Questo strumento offre la possibilità di analizzare un topic e di restituire una polarità calcolata ad-hoc. Per ogni topic, viene creato un diagramma a torta che mostra il sentimento corrente in real-time, una lista delle ultime novità, alcune foto correlate, e un insieme di parole chiave di concetti semantici che il pubblico associa al soggetto in esame. I concetti permettono di capire quali problemi o eventi guidano il sentimento verso uno stato positivo o negativo. Se si vuole sapere qual'è l'andamento del sentimento o si vuole avere un'analisi più approfondita, si può visitare il blog correlato di questo strumento. Tutti i post del blog sono generati automaticamente e aggiornati giornalmente. I Web Crawlers del blog cercano contenuti recentemente pubblicati su molti argomenti popolari e sui problemi della popolazione, per calcolarne il sentimento in maniera continua. I post presenti sul blog mostrano l'andamento nel tempo dei sentimenti e il loro rapporto Positivo su Negativo. Monitorando la nube di concetti nel tempo, è possibile notare il cambiamento nelle problematiche e nei topic chiave che la popolazione associa all'argomento in esame, e come questo influenza il sentimento. I concetti vengono raccolti dai siti Web che sfruttano uno strumento semantico, il SenseBot, associato all'Opinion Crawl.

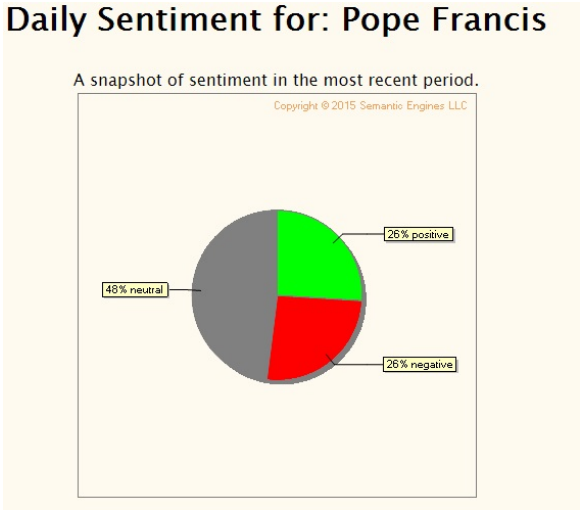

Figura 4.4: Esempio di uso di Opinion Crawl

Capitolo 5

Conclusioni

In questo lavoro è stato affrontato il campo del Social Media Analytics. In particolare sono state presentate le maggiori sfide presenti in questo momento, che vengono affrontate nei Social Media Analytics. Dopo aver descritto tali sfide è stata effettuata un'ampia ricerca dei metodi e delle tecniche più usate nella loro risoluzione. Inoltre è stata, anche, effettuata una ricerca degli strumenti più popolari per la trattazione delle sfide proposte. Questi strumenti sono stati descritti brevemente spiegando, in alcuni casi, l'approccio da loro utilizzato. Il risultato di questo lavoro vuole essere una fonte di informazioni per tutti quelli che intendono affrontare sfide analoghe nei Social Media Analytics o hanno la necessità di strumenti funzionali e performanti in tale ambito.

Capitolo 6

Bibliografia

[1] ". *The 2004 Topic Detection and Tracking (TDT2004) Task Definition and Evaluation Plan.* October 2004.

[2] Tim Althoff, Damian Borth, Jörn Hees, and Andreas Dengel. Analysis and forecasting of trending topics in online media streams. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, pages 907–916, New York, NY, USA, 2013. ACM.

[3] Regina Barzilay and Michael Elhadad. Using lexical chains for text summarization. In *In Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, pages 10–17, 1997.

[4] Shih-Hui Hsiao Clyde Holsapple, Ram Pakath. business social media analytics: Definition,benefits, and challenges. *Twentieth Americas Conference on Information Systems*, 2014.

[5] Courtney D. Corley, AR Mikler, Karan P. Singh, and Diane J. Cook. Monitoring influenza trends through mining social media. In *International Conference on Bioinformatics and Computational Biology (BIOCOMP09)*, Las Vegas, NV, July 2009.

[6] Weiguo Fan and Michael D. Gordon. The power of social media analytics. *Commun. ACM*, 57(6):74–81, June 2014.

[7] Dr.R.ManickaChezian G.Angulakshmi. An analysis on opinion mining: Techniques and tools. *International Journal of Advanced Research in Computer and Communication Engineering*, 3(7), 2014.

[8] W. Scott Spangler Hyung-il Ahn. sales prediction with social media analysis. 2014.

[9] Florian Michahelles Irena Pletikosa Cvijikj. Monitoring trends on facebook. In *IEEE Ninth International Conference on Dependable, Au-*

tonomic and Secure Computing, DASC 2011, 12-14 December 2011, Sydney, Australia, pages 895–902, 2011.

[10] Huijin Jeong Pankoo Kim Jeongin Kim, Byeongkyu Ko. a method for extracting topics in news twitter. *International Journal of Software Engineering and Its Applications*, 7(2), March 2013.

[11] Vishal Gupta Kamaldeep Kaur. a survey of topic tracking techniques. *International Journal of Advanced Research in Computer Science and Software Engineering*, 2(5), May 2012.

[12] Jan H. Kietzmann, Kristopher Hermkens, Ian P. McCarthy, and Bruno S. Silvestre. Social media? get serious! understanding the functional building blocks of social media. *Business Horizons*, 54(3):241–251, May 2011.

[13] J. L. Kolodner. An introduction to Case-Based reasoning. *Artif. Intell. Rev.*, 6(1):3–34, 1992.

[14] S. B. Kotsiantis. Supervised machine learning: A review of classification techniques. In *Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*, pages 3–24, Amsterdam, The Netherlands, The Netherlands, 2007. IOS Press.

[15] Rick Lawrence. social media analytics. April 2011.

[16] Jure Leskovec. Social media analytics: Tracking, modeling and predicting the flow of information through networks. In *Proceedings of the 20th International Conference Companion on World Wide Web*, WWW '11, pages 277–278, New York, NY, USA, 2011. ACM.

[17] Ram Mohana Reddy Guddeti Malhar Anjaria. Influence factor based opinion mining of twitter data using supervised learning. In *Sixth International Conference on Communication Systems and Networks, COMSNETS 2014, Bangalore, India, January 6-10, 2014*, pages 1–8, 2014.

[18] Michael Mathioudakis and Nick Koudas. Twittermonitor: Trend detection over the twitter stream. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, SIGMOD '10, pages 1155–1158, New York, NY, USA, 2010. ACM.

[19] Olena Medelyan. Computing lexical chains with graph clustering. In *Proceedings of the 45th Annual Meeting of the ACL: Student Research Workshop*, ACL '07, pages 85–90, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

[20] Remo Pareschi, Marco Rossetti, and Fabio Stella. Tracking hot topics for the monitoring of open-world processes. In *Proceedings of the 4th International Symposium on Data-driven Process Discovery and Analysis (SIMPDA 2014), Milan, Italy, November 19-21, 2014.*, pages 138–149, 2014.

[21] Kanagasabi Rajaraman and Ah-Hwee Tan. Topic detection, tracking, and trend analysis using self-organizing neural networks. In *Proceedings of the 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, PAKDD '01, pages 102–107, London, UK, UK, 2001. Springer-Verlag.

[22] Hassan Sayyadi and Louiqa Raschid. A graph analytical approach for topic detection. 13(2):4:1–4:??, December 2013.

[23] Ramine Tinati, Olivier Phillipe, Catherine Pope, Les Carr, and Susan Halford. Challenging social media analytics: Web science perspectives. In *Proceedings of the 2014 ACM Conference on Web Science*, WebSci '14, pages 177–181, New York, NY, USA, 2014. ACM.

[24] Christian Wartena and Rogier Brussee. Topic detection by clustering keywords. In *Proceedings of the 2008 19th International Conference on Database and Expert Systems Application*, DEXA '08, pages 54–58, Washington, DC, USA, 2008. IEEE Computer Society.

[25] Michael D. Gordon Weiguo Fan. unveiling the power of social media analytics. 2013.