POLITECNICO DI MILANO

FACOLTÁ DI INGEGNERIA DELL'INFORMAZIONE

Corso di Laurea in Ingegneria Informatica

STUDIO E ANALISI DI CORRELAZIONE TRA VENDITE E SENTIMENT WEB APPLICATO AI BOTTEGHINI CINEMATOGRAFICI

Relatore: Prof. CHIARA FRANCALANCI
Correlatore: Ing. ALESSANDRO POLI

Tesi di laurea di:
Dama Alessandro
Matr. 771175

Anno Accademico 2012 - 2013

Alla mia famiglia

Ringraziamenti

Desidero ringraziare la Prof.ssa Chiara Francalanci e l'Ing. Alessandro Poli che mi hanno aiutato, consigliato e seguito nella realizzazione di questo progetto

Desidero ringraziare soprattutto i miei famigliari, a partire dai miei genitori che mi hanno sempre seguito e appoggiato con affetto lungo tutto il mio percorso di studi e senza di loro non sarei arrivato cosí lontano, e mia sorella e mio fratello che hanno sempre creduto incondizionatamente in me e nelle mie capacitÃ .

Ringrazio anche la mia fidanzata che mi é stata accanto e mi ha sostenuto e accompagnato in quest'ultimo anno e mezzo.

Inoltre, per ultimi ma non meno importanti, ringrazio tutti miei amici e compagni di studio che oltre ad aiutarmi nella preparazione sono stati ottimi compagni di vita e hanno reso tutto questo percorso piú leggero e divertente.

Prefazione

Il presente lavoro di tesi ha come obiettivo quello di mostrare come la grande quantitá di dati presente sui Social Media sia una fonte di informazioni preziose utili in diversi ambiti. Noi ci prefiggiamo di andare ad analizzare il settore delle vendite cinematografiche assieme all'informazione contenuta su uno dei Social Media piú diffusi come la piattaforma di microbloggin di Twitter.

In questo contesto molte aziende stanno lavorando e operando pre sfruttare al meglio questi dati e risciure ad essere piú vicini ai clienti che sono i benefattori dei servizi che producono. Tutto questo é nato con l'avvetto del cosiddetto Web 2.0 che ha avvicitato gli utenti al web e che li ha resi partecipi del suo stesso contenuto. Questi aspetti sono piú che mai attuali e sono in continua evoluzione ed esplorazione.

Inoltre andremo ad esplorare e ad applicare alcune tecniche di Sentiment Analisys che consistono nell'analizzare un testo estraendo da esso delle caratteristiche come ad esempio la polaritá, cioé se esprimono un giudizio positivo o negativo.

Noi vedremo ed osserveremo questo contesto e cercheremo di dimostrare, all'interno del nostro ambito, che il contenuto presente su Twitter rispecchia abbastanza fedelmente quali sono i reali sviluppi delle vendite.

Indice

Capitolo 1

INTRODUZIONE

La nascita del Web 2.0 e la diffusione dei Social Media hanno portato alla creazione di un bene molto prezioso che corrisponde alla grande mole di dati che i semplici utenti del Web lasciano su portali, blog e social network. Queste informazioni sono costituite da dati non strutturati, complessi e disordinati che se filtrati, manipolati ed estratti con certe modalitá permettono di entrare in contatto con una banca dati dalle enormi potenzialitá. Queste potenzialitá sono state comprese dalle imprese che in questi ultimi anni stanno investendo in ricerca e sviluppo per lo sfruttamento di dati che puo' permettere di comprendere l'andamento di mercati, la reputazione di certi brand o aziende, scoprire nuove esigenze e nuovi servizi. Tutto questo é possibile grazie al coinvolgimento totale della gente comune che il Web 2.0 comporta e quindi la possibilitá di arrivare a informazioni che in passato erano irraggiungibili e altamente costosi. Grazie a tutto questo sono nate tecniche e attivitá d'impresa nuove come il Social Media Marketing o l'analisi della Brand Reputation. Inoltre, con la nascita delle tecniche di Sentiment Analisys si é arrivati alla possibilitá di captare i gusti e i pareri delle persone utili alle imprese per fornire sempre piú beni e servizi che rispecchiano i bisogni concreti e reali delle persone.

In questo progetto di tesi vogliamo cercare di dimostrare che tutta questa informazione é realmente utile, che ha il potenziale concreto per poter essere usata per capire gli andamenti dei mercati in vari settori e poter basare i sistemi previsionali anche su questo tipo di informazioni. Piú nel dettaglio, in questo lavoro andremo ad analizzare il mercato cinematografico e sfrutteremo le informazioni estraibili da uno dei principali e diffusi social media, Twitter, per capire se questi dati consentono di prevedere le vendite di un film e se i pareri, le considerazioni e le recensioni presenti su questo media di nuova generazione, possono influenzare o meno i consumatori. Questo elaborato di tesi é strutturato come segue:

Nel Capitolo 2 presenteremo le caratteristiche, le novitá e le potenzialitá che il Web 2.0 ha portato, dalla diffusione e utilizzo quotidiano dei Social Media, alle caratteristiche fondamentali dell'informazione presente su questi canali, alla

nascita e allo sviluppo del Social Media Marketing e delle tecniche di Sentiment Analisys.

Nel Capitolo 3 mostreremo i motivi per cui é stato scelto il contesto cinematografico, come si presentano i dati di vendita e le varie informazioni utili per la nostra analisi, ed infine si presenterá l'idea di come estrarre solo l'informazione necessaria da Twitter.

Nel Capitolo 4 mostreremo concretamente gli strumenti software realizzati per l'estrazione delle informazioni dal Web ed infine come si é estratto il sentiment da questi dati

Nel Capitolo 5 vedremo come sono stati trattati e manipolati i dati precedentemente estratti verificando o meno la presenza di correlazione tra le vendite ai botteghini dei film ed i vari aspetti ricavati dal Social Media.

Nel Capitolo 6 trarremo le conclusioni di questo progetto vedendo se le nostre supposizioni per quanto riguarda il significato che si cela dietro alle informazioni presenti sui Social Media corrispondono a qualcosa di concreto oppure no.

Capitolo 2

STATO DELL'ARTE

2.1 WEB 2.0 E SOCIAL MEDIA

Con il termine Web 2.0 [1] si vuole indicare l'evoluzione che ha avuto il Web fino a portarlo a come lo conosciamo ai giorni nostri. Questo termine nasce alla fine del 2004 durante una conferenza tra la O'Reilly Media e MediaLive International dove, per l'occasione, Dale Dougherty, pioniere del web e Vice-Presidente di O'Reilly, sottolineó l'importanza di nuove applicazioni web attraverso questo nuovo concetto. Essendo complesso e articolato il nuovo Web, esso fu presentato sottolineando le differenze sostanziali che si sono create rispetto al Web iniziale chiamato di conseguenza in modo retroattivo Web 1.0. Queste differenze si focalizzano sostanzialmente sulla visione dinamica del web e soprattutto al coinvolgimento in prima persona dell'utente grazie ad una notevole interazione tra esso e i siti come blog, forum, chat, wiki, piattaforme di condivisione di media come Flickr, YouTube, Vimeo, social network come Facebook, Myspace, Twitter, Google+, Linkedin, Foursquare, ecc. Per esemplificare al meglio tutto ció si puo' fare riferimento al decalogo definito dal futurologo Vito di Bari proprio sul Web 2.0 [2]:

- il web é una piattaforma, si passa dai software installati sul pc degli utenti ai servizi online, ormai tutto é online non piú scorporato.
- ll web é funzionalitá, i siti web diventano fonti di contenuto e di servizi e non piú centri di informazione statici.
- Il web é semplice, sia intermini di accesso che di utilizzo, é user-friendly.
- Il web é leggero, la leggerezza é connotata dalla condivisione di contenuti e servizi e abilitata dall'implementazione di elementi modulari intuitivi e di facile utilizzo.
- Il web é sociale, le persone fanno il web.

- Il web é flusso, gli utenti divengono co-sviluppatori, cosí si cambia e migliora perpetuamente.
- Il web é flessibile, il software si colloca a un livello superiore rispetto al singolo dispositivo (device) per far leva sul potere della long tail attraverso il customer self-service per raggiungere l'intero web.
- Il web é mixabile, gli utenti stessi riescono ad ottenere applicazioni nuove dall'incrocio di applicazioni giá esistenti, si creano i cosiddetti mashup
- Il web é partecipativo, gli utenti aggiungono valore all'applicazione mentre la usano (o la modificano).
- Il web é nelle nostre mani, sono gli stessi utenti che categorizzano e organizzano i contenuti: il social tagging é un potere nelle mani degli utenti.

La nascita di questo nuovo tipo di Web é stata dovuta all'espansione dell'utilizzo del web e al maggior facilitá di accesso/acquisto di una connessione Internet efficace e veloce. Dall'ideologia e dalla tecnologia alla base del Web 2.0 nascono i Social Media [3] che consentono la creazione e lo scambio di User Generated Contet (UGC), termine che sta a rappresentare le varie forme dei contenuti multimediali. Questi contenuti devono essere accessibili a tutti gli utenti di Internet o ad una parte di utenti facente parte di un sito di social networking, essere creativi e non avere scopi commerciali. I social media rappresentano fondamentalmente un cambiamento nel modo in cui la gente apprende, legge e condivide informazioni e contenuti. In essi si verifica una fusione tra sociologia e tecnologia che trasforma il monologo (da uno a molti) in dialogo (da molti a molti) e ha luogo una democratizzazione dell'informazione che trasforma le persone da fruitori di contenuti ad editori. Sono diventati molto popolari perché permettono alle persone di utilizzare il web per stabilire relazioni di tipo personale o lavorativo. Una prima classificazione dei Social Media che puo' essere fatta, considerando tutte le piattaforme social, é la seguente:

- Blogging: giornale o diario online, generalmente aggiornato giornalmente o settimanalmente.
- Microblogging: Blog di ridotte dimensioni, massimo 140 caratteri
- Social Network: rete che permette interazioni social con i propri amici e colleghi.
- Social news: gli utenti caricano notizie che vengono votate dalla comunitá online.
- Podcasting: social media in cui é possibile distribuire i contenuti in un formato maggiormente interattivo

- Social shopping: gli utenti condividono e recensiscono prodotti venduti online

In base ai vari scopi dei Social Media, possiamo classificarli anche nel seguente modo:

- Social Media di comunicazione, es. Facebook, Twitter, Linkedin
- Social Media di collaborazione, es. Wikipedia, Slideshare
- Social Media multimedia, es. Youtube
- Social Media di comunitá virtuali
- Social Media per aziende

Si possono anche individuare le funzionalitá che permettono di distinguere e catalogare i Social Media, queste sono presentate nella figura seguente [4]:

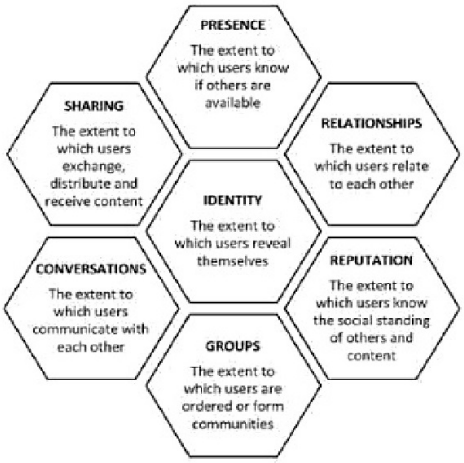
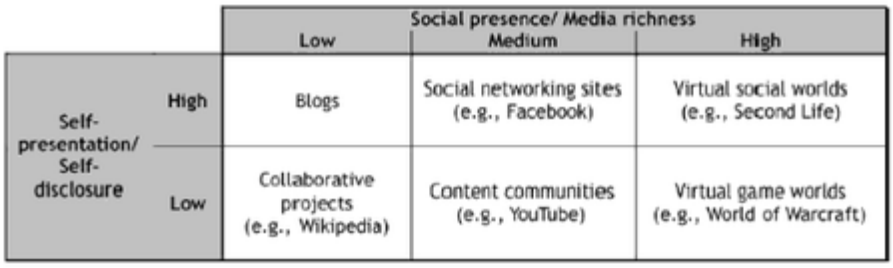


Figura 2.1: Figura che rappresenta le funzionalitá dei social Media

Inoltre é possibile definire delle dimensioni di analisi sul quale stilare una classificazione visibile nella seguente tabella: Con Social presence s'intende il livello di contatto che un social media permette di raggiungere con un altro utente, Self-presentation consiste nel cercare di definire le impressioni che gli altri utenti si fanno su di loro, Media richness si riferisce alla velocitá con cui le informazioni circolano mentre la self-disclosure rappresenta la quantitá di dati personali dell'utente richieste da un dato mezzo.

2.2 SOCIAL MEDIA DEMOGRAPHICS

I Social Media hanno preso sempre piú piede e si sono diffusi in modo esponenziale negli ultimi 10 anni e questa crescita non sembra minimamente arrestarsi.



		Social presence/ Media richness		
		Low	Medium	High
Self-presentation/ Self-disclosure	High	Blogs	Social networking sites (e.g., Facebook)	Virtual social worlds (e.g., Second Life)
	Low	Collaborative projects (e.g., Wikipedia)	Content communities (e.g., YouTube)	Virtual game worlds (e.g., World of Warcraft)

Table 1: Classification of Social media by social presence/media richness and self-presentation/self disclosure (Kaplan & Haenlein, 2009)

Figura 2.2: Figura che rappresenta la classificazione dei social Media

Ma chi sono gli utenti dei Social Media? Secondo una ricerca effettuata da Online MBA [5] piú del 66% degli utenti adulti sono connessi ad uno o piú piattaforme social. Per rendere ancor piú le dimensioni dei Social Media basti pensare che Facebook conta 845.000.000 di utenti attivi, Twitter 127.000.000, Linkedin 150.000.000, Google+ 90.000.000 e questi sono solo i principali social media. In dettaglio sono state analizzate le ripartizioni secondo livello di istruzione, reddito, etá e sesso degli utenti, insieme ad altri fattori. Per quanto riguarda il livello di istruzione é risultato che gli studenti universitari, o coloro che hanno completato un college, rappresentano la maggioranza sui siti di social media come Facebook, Twitter,Pinterest, Digg e Reddit. Tra gli utenti di Facebook, il 57% hanno completato l'universitá, e il 24% ha conseguito un bachelor o master, dati simili per Twitter con il 59% di laureati e il 24% di chi ha conseguito un bachelor o master. Su Linkedin crescono le percentuali per chi possiede un master, 37%, mentre i laureati sono sempre una percentuale corposa che si assesta sul 50%. Nel seguente grafico sono presenti in dettaglio le divisioni secondo l'istruzione, degli utenti dei principali social network: Per quanto riguarda il sesso degli utenti vi é una maggioranza di utenza femminile, infatti il 57% su Facebook e il 59% su Twitter sono donne. In piú si ha una preponderanza femminile su Pinterest che registra il piú pesante squilibrio tra i sessi, 82% degli utenti sono donne, che pinnano, idee regalo, hobbistica, interior design e moda. Mentre su Google+ c'é una maggioranza, abbastanza evidente, di utenza maschile con il 71%, contro il 29% femminile: Inoltre per quanto riguarda le fasce d'etá si ha un risultato curioso su Google+ dove ben il 50% degli utenti ha un etá inferiore hai 25 anni e questi sono spesso giovani Ingegneri e sviluppatori. Di seguito é mostrata la distribuzione sui principali social media della loro popolazione in base alle fasce d'etá: Dopo aver illustrato come é composto il popolo dei social media andiamo ad illustrare i motivi per il quale milioni di persone si registrano e partecipano attivamente in questi siti.

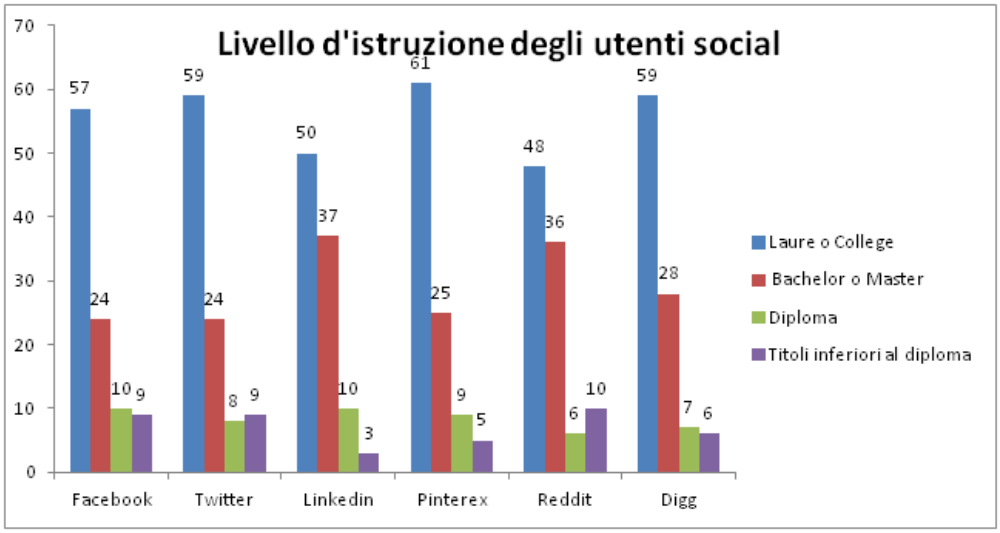


Figura 2.3: Grafico che rappresenta il livello d'istruzione degli utenti social

Secondo uno studio sempre offertoci da Online MBA, gli americani utilizzano i social network principalmente per 3 motivi connessi tra loro:

- Stare in contatto con gli amici (67%)
- Stare in contatto con i propri familiari (64%)
- Riallacciare i rapporti con amici di vecchia data (50%)

Da questi dati si evince il fatto che lo scopo principale dei Social Media é proprio quello di permettere facilmente di instaurare contatti con le persone, in questo caso, soprattutto persone che si conoscono giá nel mondo reale e con il quale si vuole proseguire un rapporto di conoscenza attraverso il web nel caso in cui sia difficoltoso proseguirlo in un contesto reale. Con percentuali notevolmente piú basse si utilizzano i social media per:

- Avere contatti con persone aventi interessi comuni (14%)
- Farsi nuovi amici (9%)
- Leggere commenti di celebritá, atleti o politici (5%)
- Trovare potenziali partner amorosi (3%)

Questi risultati sono in linea anche con gli utenti italiani, infatti secondo i dati di un indagine condotta dall'azienda Swg di Trieste, su un campione analizzato di 1.326 tra iscritti a Facebook, My- Space, Linkedin, Badoo e Giovani.it, ben l'88% utilizza questi canali per riuscire a restare in contatto con amici e conoscenti, ossia quindi con persone con cui si aveva giá una relazione prima che esistesse tale tipologia di siti, ma soprattutto se ne fa anche un uso creativo, commentando

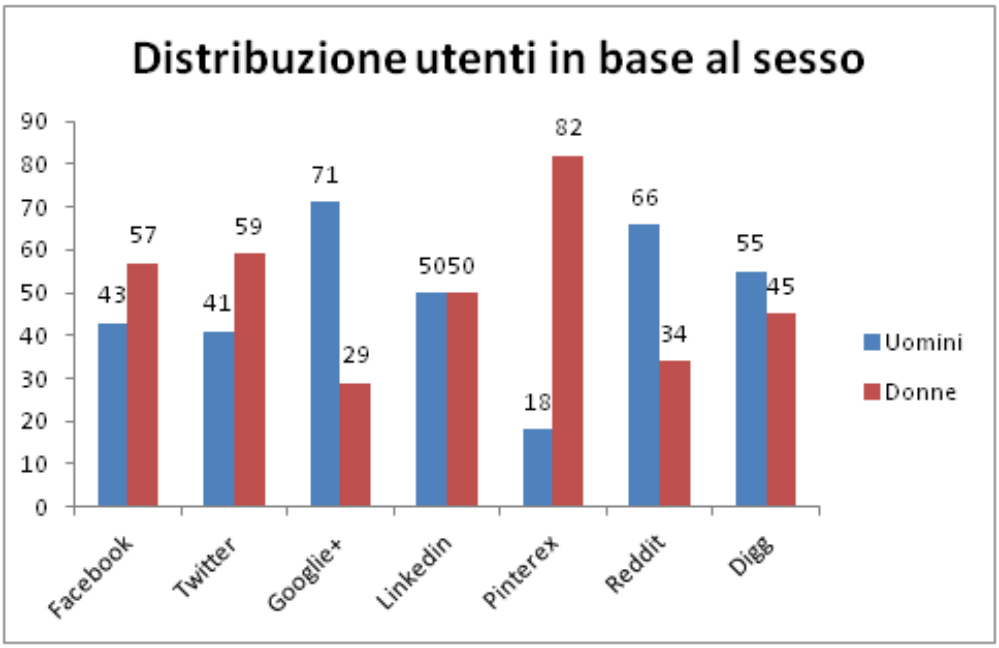


Figura 2.4: Grafico che rappresenta la Distribuzione degli Utenti in base al Sesso

in gruppo eventi o programmi di interesse comune. Invece conoscere nuovi amici riguarda infatti solo il 33% degli intervistati, mentre trovare nuovi partner addirittura é considerata una motivazione poco o per niente considerata da ben l'86% del campione. La maggior parte degli intervistati comunque pare avere le idee chiare anche su questo: il 76% degli iscritti sostiene infatti di usare molto o abbastanza questi strumenti per condividere informazioni. Spesso infatti gli iscritti non postano solo i propri pensieri personali, ma anche e soprattutto articoli di giornali e di blog che hanno trovato in giro per la Rete. Il risultato é che chi partecipa a queste reti sociali si trova pronta ogni giorno una sorta di rassegna stampa, con in piú la possibilitá, per niente secondaria, di commentarla e discuterla con gli altri. In Italia il Social Network piú utilizzato é Facebook con il 96%, poi segue MySpace con il 21% e infine Twitter col 17%. Inoltre nel nostro paese siamo molto socievoli: rispetto alla media europea, l'Italia, risulta il paese con numero di amici per ogni utente piú elevato [8]: 88 amici, davanti a Regno Unito (77), Olanda (77), Spagna (69), Germania (57) e Francia (53). In piú il rapporto con le marche o brand sui social media é molto buono: il 29% degli utenti social network si dichiara fan o follower di un brand. Piú coinvolti di noi, in Europa, solo gli utenti di UK con il 32%. Infine ben il 76% degli utenti italiani dichiara di accedere ai social media una o piú volte al giorno e il 18% vi accede almeno una volta a settimana. Ancor piú sorprendente é vedere quanto a livello mondiali, gli italiani siano i maggiori utilizzatori di social media: la quota dei nostri connazionali che li frequenta (86% di tutti gli internauti) ha superato il dato statunitense (79%). E' uno dei dati che

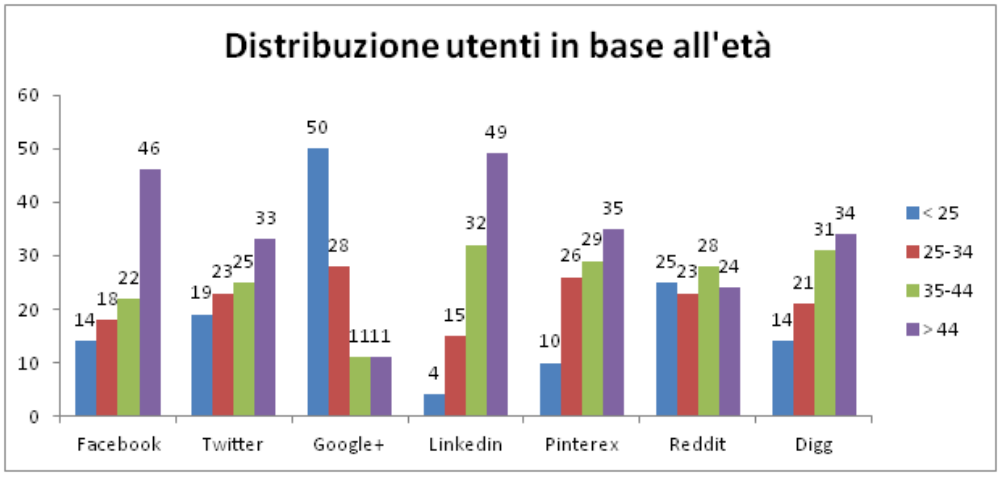


Figura 2.5: Grafico che rappresenta la Distribuzione degli Utenti in base all'etá

emergono dal report State of the Media: The Social Media Report di Nielsen, [8] che esplora i cambiamenti in atto nel panorama dei social media, le modalitá di interazione dei consumatori con questi strumenti e le piattaforme digitali che ne trainano l'utilizzo. Di seguito vediamo un grafico che stila la classifica dei paesi mondiali con la maggior percentuale di utenti sui social media rispetto agli utilizzatori totali di internet: Gli italiani non sono solo i principali utilizzatori,

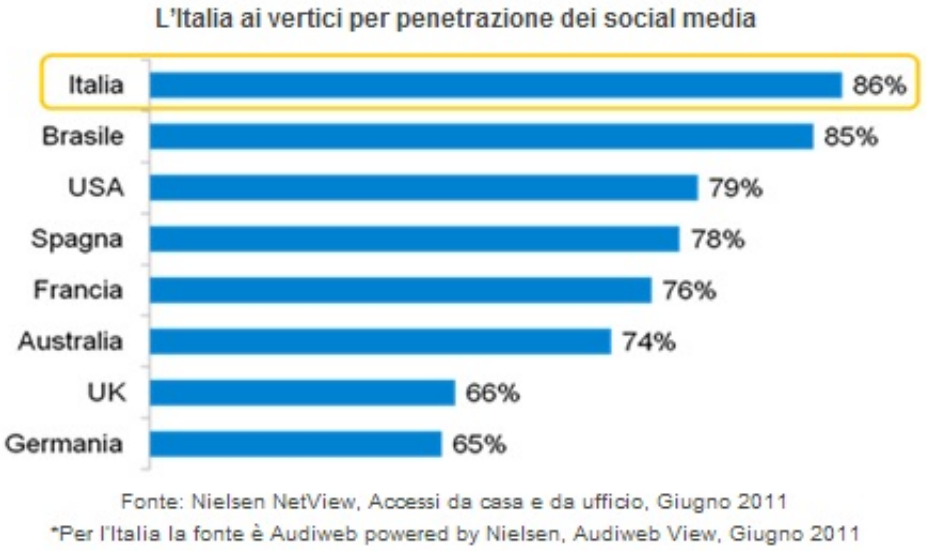


Figura 2.6: Grafico che rappresenta la Distribuzione degli Utenti per Paese

ma anche tra quelli che passano piú tempo su blog e social network: all'incirca un terzo di tutto il tempo trascorso online, contro un quarto degli americani. In piú si sta notevolmente diffondendo anche l'utilizzo dei social con i dispositivi mobile che nel biennio 2010 2011 é piú che raddoppiato ed é sempre piú in

aumento fino ad oggi, infatti come si puo' notare nel seguente grafico, che le applicazioni di social networking sono la terza categoria piú scaricata, dopo giochi e mappe in Italia, dopo giochi e meteo negli Stati Uniti.

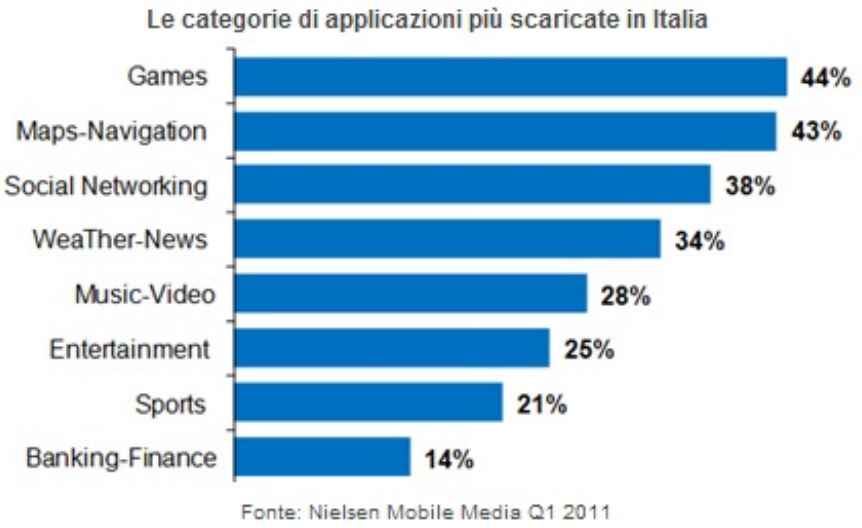


Figura 2.7: Grafico che rappresenta le applicazioni piú scaricate in Italia

2.3 DIFFERENZE TRA SOCIAL MEDIA E MEDIA TRADIZIONALI

I social media sono diversi dai media industriali come giornali, televisione e cinema. Mentre i social media sono strumenti relativamente a basso costo che permettono a chiunque (anche soggetti privati) di pubblicare ed avere accesso alle informazioni, i media tradizionali richiedono cospicui investimenti finanziari per pubblicare informazioni. Tra le voci di spesa dei media industriali possiamo annoverare ad esempio la pressa tipografica o le autorizzazioni statali. Una caratteristica che accomuna social media e media industriali é la capacitá di ottenere un'audience sia vasta che ridotta; sia il post di un blog che una trasmissione televisiva possono raggiungere milioni di persone oppure nessuno. I parametri che aiutano a descrivere le differenze tra i due tipi di media variano a seconda del tipo di analisi. Alcuni di questi parametri sono [7]:

- bacino d'utenza: sia i social media che i media industriali offrono a ciascuno l'opportunitá di ottenere un'audience globale.

- accessibilitá: i mezzi di produzione dei media industriali sono generalmente di proprietá privata o statale; gli strumenti dei social media sono disponibili da ciascuno a un costo contenuto o gratuitamente.

- fruibilitá: la produzione di mezzi industriali richiede in genere formazione e competenze specialistiche; i social media invece no, o in qualche caso reinventano le competenze, cosicché ciascuno puó gestire i mezzi di produzione.

- velocitá: il tempo che intercorre tra le informazioni prodotte dai media industriali puó essere lungo (giorni, settimane o anche mesi) in confronto al tempo impiegato dai social media (che hanno la possibilitá tecnica di reagire istantaneamente, solo la mancanza di reattivitá dei partecipanti puó comportare ritardi). Poiché ormai anche i media industriali si avvalgono degli strumenti dei social media, questo potrebbe non essere piú un tratto distintivo.

- permanenza: una volta creati, i mezzi industriali non possono essere piú modificati (una volta stampato e distribuito, l'articolo di una rivista non puó piú ricevere modifiche), mentre i social media possono essere cambiati quasi istantaneamente mediante commenti e modifiche.

Un'ulteriore distinzione riguarda la responsabilitá. I media industriali sono tenuti a rendere conto alla societá della qualitá dei contenuti e dei risultati delle loro attivitá in termini di interesse pubblico, responsabilitá sociale ed indipendenza editoriale. I social media non hanno altrettante responsabilitá in merito alle loro attivitá editoriali. I media tradizionali si caratterizzano per la scarsa interattivitá e l'asimmetria comunicativa. Le informazioni vengono trasmesse da uno a molti; il messaggio é selezionato e filtrato da un redattore che segue una precisa linea editoriale; il lettore/spettatore non ha possibilitá di scegliere o di intervenire sui contenuti. I social media, invece, sono caratterizzati da un approccio orizzontale da molti a molti, in cui gli utenti scelgono le informazioni da creare o diffondere e possono interagire con altri utenti per approfondire, migliorare, criticare quelle informazioni. Di seguito presentiamo una tabella riepilogativa delle principali differenze tra media tradizionali e social media: Dalla

	Media Tradizionali	**Social Media**
Attendibilità	Brand	Individuale
Contribuzione	Linea editoriale controllataAutoreferenzialità	Libera
Pubblicazione	Pianificata	In tempo reale
Stile di scrittura	Distaccato e lineare	Personale
Relazione con il pubblico	Indiretta (sondaggi e vendite)	Interattiva e partecipativa
Limiti fisici	Pagina o durata della trasmissione	Non hanno limiti di pagine o di ore

Figura 2.8: Tabella media tradizionali vs social media

tabella emerge che i social media hanno abbattuto le barriere di accesso alla comunicazione e tutti possono diffondere messaggi o contenuti selezionandoli sulla base delle preferenze personali. Gli utenti determinano l'importanza delle news sostituendosi al processo editoriale. Con i social media é cambiato i modo di consumare le informazioni ed é cambiato il pubblico di riferimento: Le caratteristiche innovative dei social media riguardano soprattutto:

- il ruolo del cittadino (da fruitore/spettatore a attore)
- rapiditá nella diffusione delle notizie
- esposizione di contenuti che si basano sull'esperienza e sulle preferenze personali
- interattivitá e modificabilitá dei contenuti
- multimedialitá
- trasparenza (non é possibile mentire)
- persistenza (l'informazione resta sempre presente dopo la pubblicazione)
- viralitá (il messaggio tende a diffondersi senza interruzioni nel tempo)

Queste caratteristiche, di portata rivoluzionaria, hanno messo in ombra i media tradizionali che hanno perso di credibilitá. La crescente mole di immagini omologate che le grandi agenzie mondiali dell'informazione producono ogni giorno inibiscono la capacitá di approfondire gli argomenti trattati in modo originale. Il mercato editoriale presenta alti indici di concentrazione ed é accusato di autoreferenzialitá poiché appare come una rete chiusa che si autoalimenta. Questi aspetti hanno condotto ad una progressiva sfiducia verso i media tradizionali e ad un sempre maggiore successo dei social media. Una recente indagine della fondazione statunitense Project for Excellence in Journalism (Pej) ha messo ha confronto per una settimana, nel giugno scorso, l'agenda dei media tradizionali e quella di alcuni dei siti web 2.0 piú popolari in America (Digg, Reddit, Del.icio.us). Dall'analisi sono emersi alcuni dati interessanti:

- Molte storie tra le piú lette dagli utenti non sono apparse su alcuna testata tradizionale di informazione
- Gli interessi dei lettori sono diversi da quelli dell'editoria tradizionale
- I media tradizionali tendono a rivisitare per piú giorni la stessa storia da diversi punti di vista, mentre i lettori tendono a privilegiare la varietá degli argomenti
- interattivitá e modificabilitá dei contenuti

- I lettori tendono a riaggregare le notizie nello stile del citizen journalism piuttosto che seguire l'agenda selezionata dai media tradizionali

Dallo studio é emerso che il 70% di quanto segnalato e discusso online proveniva da blog o community come YouTube e MySpace, mentre solo il 25% discendeva dai portali delle testate tradizionali. Un altro aspetto significativo é che sebbene Internet rappresenti la globalitá per eccellenza, la fruizione dei contenuti su web 2.0 non privilegia i temi transnazionali ma fatti locali, regionali e informazioni pratiche legate alla vita quotidiana. I social media, inoltre, non si focalizzano a lungo su un singolo evento, come avviene per i portali nazionali che riportano per diversi giorni articoli su un singolo avvenimento. Nei social media la notizia piú importante non mantiene a lungo la prima posizione ed é facilmente soppiantata da nuove storie.

2.4 SOCIAL MEDIA MARKETING

Il Social Media Marketing [9] é il modo per aumentare la popolaritá e diffondere ad un vasto pubblico (community) il proprio brand, i prodotti e i servizi, attraverso i Social Network. La presenza sui Social Network é fondamentale e proficua. Una qualsiasi notizia che normalmente avrebbe bisogno di mesi se non di anni per la diffusione, su Twitter e su Facebook si espande a macchia d'olio nel giro di poche ore. La vera forza della comunicazione sui Social Network é rappresentata dal passaparola online, che si propaga semplicemente grazie, ad esempio, ai tasti condividi, mi piace e Retweet. Questo funzionerá ovviamente se il messaggio, il prodotto o il servizio sará oggettivamente interessante, coinvolgente, innovativo. A differenza della pubblicitá sui mass media, dove il consumatore riceve passivamente il messaggio, senza alcun contraddittorio, sui social network é la community che puó decretare o meno il successo di una impresa, interagendo e lasciando un commento. Una coordinata azione marketing sui Social Network, se ben implementata, puó avere un fortissimo impatto a livello commerciale e, in breve tempo, far aumentare il business dell'impresa. Una buona strategia di web marketing oltre alla presenza sui social network, deve necessariamente prevedere un sito internet aziendale, ben indicizzato sui motori di ricerca e un blog. I contenuti del sito e quelli del blog (testi, immagini, filmati), devono essere costantemente aggiornati. Cosí come i diversi profili su Facebook, Twitter, Youtube ecc. devono contenere nuovi post ai quali occorre rispondere ai commenti lasciati dagli utenti, per creare audience e, nel caso in cui fosse necessario, difendere la tua reputazione online. Il Social Media Marketing quindi risulta avere molta importanza soprattutto vedendo i risultati di un'indagine realizzata da eGain a livello europeo, analizzando il comportamento d'acquisto dei consumatori e le modalitá d'interazione preferite. Il sondaggio, svolto su un campione di 3 mila consumatori europei, di cui circa 500 italiani ha tratto come risultati il fatto che per gli italiani, come il resto

degli europei, sono influenzati positivamente relativamente ad un acquisto se l'azienda o il venditore associato risponde velocemente alle mail e possiede una buon sezione di FAQ sul proprio sito web. Inoltre dal sondaggio emerge che gli italiani si fanno influenzare meno, rispetto al resto degli europei, dalle recensioni negative degli altri utenti su beni e servizi. Inoltre risulta utilizzato sempre con piú frequenza il canale social per lamentele o problemi di assistenza. Piú precisamente, dal sondaggio, risulta ancora il telefono il mezzo piú utilizzato, soprattutto dagli italiani, seguito dalle mail o dal web self-service e infine il 17% dei consumatori usa le web chat o i social media. Quindi risulta importante che si tratti di piccole imprese o grandi, che realizzare anche solo le basi di quello che viene generalmente proposto per il social media marketing significa quanto meno investire una certa quantitá di tempo delle persone interne o assumerne di nuove, coinvolgere e pagare degli esterni, tipo agenzie, formatori, consulenti, e talvolta anche rivedere o stravolgere l'organizzazione interna. Questo perché banalmente, aprire un blog e seguirlo, gestirlo, promuoverlo, frequentare Twitter, creare una Facebook Fan Page e rispondere ai commenti dei fan, realizzare strategie conversazionali e dunque conversare quotidianamente, offrire promozioni attraverso i social network, contest, monitorare ció che viene detto e talvolta rispondere, intervenire alle discussioni su temi di interesse, tutto questo e molto altro é spesso troppo, di piú di quello che molte aziende possono fare, visto che l'occupazione principale é quella di produrre beni o realizzare servizi. Esistono delle strategie di social media marketing [10], alternative o complementari, che le aziende possono realizzare in base alle proprie risorse e al proprio tempo, e i diversi gradi di copertura possibili, attraverso questa semplificazione delle possibili soluzioni:

- Strategia Presenzialista

L'azienda apre i suoi canali nel Web utilizzandoli in maniera scarsa. Piú che una strategia é la realtá di molte aziende, che non riescono a fare di meglio. Puó avere un senso quando non si possono dedicare risorse a questa attivitá e si vuole comunque lasciare una porta da cui ogni tanto sbirciare, senza per questo aspettarsi di ottenere alcun risultato specifico e stando attenti a non creare confusione, come purtroppo a volte avviene.

- Strategia Editoriale

L'azienda produce giá dei contenuti che inizia a veicolare attraverso il Social Web, o decide di iniziare a produrne con questo intento. La presenza sui social media é legata alla divulgazione di notizie, articoli e informazioni, e un obiettivo cui si tende é l'aumento della quantitá di pubblico. I vantaggi di questa strategia sono legati alla visibilitá, sui motori come sui siti, blog e forum del proprio settore, ma il tempo richiesto, se i contenuti vengono prodotti solo con questo fine, é molto elevato, e bisogna dedicare risorse alla promozione e visibilitá.

- Strategia Conversazionale

L'azienda interviene nei social media con l'intento di partecipare alle conversazioni e di creare valore per i propri clienti. Monitora ció che viene detto su di lei e sui propri prodotti, interviene quando necessario, accoglie domande e risolve le critiche che le vengono mosse attraverso prevalentemente i social network aperti come Twitter. Con questo approccio instaura delle relazioni, fidelizza clienti e ne raggiunge di nuovi. Il problema piú diffuso di chi utilizza questa strategia, lasciando stare i casi in cui si verifichino forti attacchi e critiche, é che all'apertura sul Web non corrisponda un servizio coerente nell'offline.

- Strategia Strutturale

L'azienda decide di utilizzare i social media non (soltanto) per creare la propria presenza, ma per abilitare le persone a condividere con il minor sforzo possibile la propria passione per il brand e a consigliarne l'utilizzo agli amici attraverso il proprio profilo su Facebook, su Twitter o il proprio blog. E' il caso degli share e degli invita presenti nei siti Web, delle applicazioni per Facebook, dei widget embeddabili ed embeddati: soluzioni tecniche che utilizzano l'aspetto piú interessante dei social media, le persone, per ottenere visibilitá e guadagnarne in reputazione.

- Strategia Analitica

L'azienda non utilizza in maniera attiva il Social Web ma ne sfrutta le conversazioni per trarne informazioni e spunti strategici, oltre che per migliorare i propri servizi e la propria comunicazione al di fuori degli ambienti sociali. Sono le aziende che pur non investendo in attivitá dirette, si rivolgono ad agenzie specializzate che offrono loro report dettagliati sulle conversazioni e consigli strategici derivanti da questi studi.

- Strategia Virale

Si tratta di utilizzare i social media per diffondere, in qualsiasi modo, non un insieme di contenuti ma uno o pochi oggetti specifici, che in generale non riguardano l'azienda ma la richiamano in qualche forma.

- Strategia Strisciante

L'azienda non utilizza i social network per creare una presenza ufficiale, ma per distribuire sotto mentite spoglie dei contenuti di rilevanza strategica. Nascono in questo modo i flog e i profili fake, le marchette nei blog e alcuni ambienti Internet un po' ambigui, siti informativi creati con secondi fini e gruppi Facebook attraenti ma costruiti con l'intento unico di veicolare offerte o messaggi aziendali.
Visto la non banalitá del Social Media Marketing nasce una nuova figura professionale, il Social Media Manager, un professionista del web marketing che lavora

al fianco delle aziende e di imprenditori e che si prende cura della comunicazione aziendale sui canali sociali. Oltre alla figura del Social Media Manager nasce anche un nuovo distretto aziendale costituito appositamente per la gestione del brand aziendale sul web, come viene definito sul blog tagliaerbe.com [11], il brand reputation management 2.0. In questo blog vengono definite delle fasi relative a questo distretto che consistono:

1. Monitorare: sapere cosa si dice del proprio brand online é un primo step fondamentale che deve essere fatto da qualsiasi azienda presente sul Web. Era stato giá affermato in precedenza che la capacitá di saper ascoltare ció che si dice della propria azienda é una componente fondamentale di un'azienda capace di interagire con la propria audience utilizzando i social media.

2. Condividere: logica conseguenza di una efficace politica di content marketing. I potenziali consumatori si conquistano soprattutto mettendo a disposizione contenuti di qualitá. Un brillante esempio che permette di comprendere un corretto utilizzo della condivisione di contenuti sul World Wide Web é indicato sul sito internet contentmarketinginstitute.com, ed é citato il caso della compagnia AtTask, una azienda B2B che grazie a regolari aggiornamenti del suo corporate blog e del suo podcast é in grado di attrarre quotidianamente potenziali consumatori sui suoi Owned Media.

3. Partecipare: i mercati sono conversazioni, l'ulteriore passo da compiere per migliorare la propria reputazione é partecipare alla conversazione. E gli strumenti per farlo sono sempre piú numerosi, a cominciare da un eventuale corporate blog, fino ad arrivare all'utilizzo di strumenti piú nuovi come Twitter, Facebook, ecc...

Puo' risultare pericoloso non seguire queste fasi per il brand reputation management, come ad esempio non seguire e partecipare alle conversazioni o non interessarsi al rapporto azienda social media. Inoltre lo scambio di opinioni e quindi le conversazioni sulle varie piattaforme possono andare verso un lato non previsto e quindi cercare di essere in grado di gestire situazioni eccezionali. Per rispondere al meglio alle reazioni del mondo social é necessario essere social e quindi l'azienda deve abbassarsi al livello di tutti gli altri utenti delle varie piattaforme e seguire certi comportamenti definiti da Andreas Kaplan and Michael haenlein:

- Essere attivi sulla piattaforma social: coloro che presenziano sui social media hanno il desiderio di diventare produttori e consumatori di informazione, e l'unico modo per ottenere ció é di stimolarmi con una costante pubblicazione e condivisione di informazioni e contenuti

- Essere interessanti: é un obiettivo importante ma difficile da realizzare ed é necessario prima essere disponibili ad ascoltare per capire le esigenze dei consumatori
- Essere umili: le aziende devono tenere in considerazione che il mondo e la cultura social sono nati prima che le aziende entrassero a farne parte e quindi é loro compito comprendere l'effettivo funzionamento dei social col quale si vuole partecipare e mantenere un livello paritetico con gli altri utenti
- Essere non professionali: non é necessario strafare spendendo migliaia di euro su blog o siti, bisogna mettersi alla prova e cercare di guadagnarsi la fiducia degli utenti anche commettendo e ammettendo qualche errore
- Essere onesti: rispettare le regole é l'ultimo passo fondamentale per partecipare all'attivitá sui social media

Tutte questi aspetti strategici e questi comportamenti non sono banali da attuare specialmente tutti insieme e in modo non superficiale. Proprio per questo molte aziende ignorano questa via ma tante altre invece, cercano di trarre benefici da questo aspetto implementando piccole strategie di Social Media Marketing. Di seguito vediamo qualche caso di applicazione reale di strategie o espedienti associabili al Social Media Marketing [12]:

1. Adobe pubblica una rivista online che si chiama CMO (Chief marketing Officer), dove fornisce numerose informazioni rilevanti sul settore del web marketing, senza menzionare se stessa od i propri prodotti, ma solo proponendo notizie relative ad altre aziende e servizi. Stessa cosa per Intel con Free Press, un magazine online che tratta le notizie relative alla tecnologia ed alle applicazioni dei prodotti Intel, senza peró fare pubblicitá diretta alla propria azienda. Testi e video possono essere ripubblicati liberamente su altri siti. Questo viene fatto per porre le aziende in questione come punti di riferimento per chi vuole tenersi aggiornato su ció che accade nella propria nicchia di mercato fornendo un servizio aggiuntivo a chi segue il brand sul web risultando peró distaccato dall'esplicita pubblicitá, infatti questi magazine dai contenuti *curati* non devono avere alcuna finalitá promozionale legata all'azienda o ai prodotti che vende.
2. La banca Intesa Sanpaolo possiede una fan page dedicata su Facebook, dove un team dell'azienda risponde quotidianamente alle domande degli attuali clienti, cercando di risolvere rapidamente i loro problemi, e fornisce informazioni puntuali ai possibili nuovi clienti. Anche Alitalia ha una fan page su Facebook dove il proprio staff, con nome e cognome e non con l'account della fan page, risponde alle lamentele dei passeggeri dopo un viaggio, fornisce informazioni per lo svolgimento di un volo aereo ed anche alle operazioni preliminari di viaggio (prenotazione volo, possibilitá di

portare animali, dimensioni dei bagagli, ecc.). In alcuni casi, sono gli stessi clienti ad aiutarsi tra loro. Stesso servizio fornito anche da Telecom Italia che possiede un account su Twitter che utilizza per fornire feedback ai propri clienti, ed informazioni utili a coloro che vogliono passare a Telecom da un altro operatore, o hanno problemi con la propria linea telefonica. Dietro, ci sono membri dello staff di Telecom Italia che rispondono con il proprio nome e cognome, non con il profilo dell'azienda.

3. Oral-B Italia propone una serie di video che spiegano in modo semplice e chiaro come funzionano le loro apparecchiature, i modi migliori di utilizzarle, la manutenzione necessaria, e come ogni prodotto puó risolvere un problema specifico legato all'igiene dentale. Inoltre esiste anche un'altra serie di video che rappresentano delle testimonianze di clienti reali che raccontano la loro esperienza di utilizzo dei prodotti dell'azienda. Sono tutte persone riprese nella loro casa o attivitá quotidiana, che spiegano i vantaggi che hanno tratto dall'utilizzare una specifica soluzione per la cura dei loro denti. Abbiamo anche IKEA che con il proprio canale Youtube presenta numerosi video a disposizione dei clienti e potenziali tali dove si possono trovare tante idee per abbellire o sfruttare in modo originale i prodotti dell'azienda, semplici istruzioni di montaggio e di utilizzo dei mobili acquistati, ed anche divertenti cartoni animati basati sui prodotti IKEA. Sempre il marchio svedese ha creato sul web Hemma, la comunitá online dei clienti IKEA. In questa comunitá si trovano ricette di cucina, idee migliori per personalizzare i mobili, mercatini per vendere / acquistare prodotti IKEA di seconda mano, forum dove fare domande e rispondere a dubbi di altri acquirenti.

4. Nike ha messo a disposizione sul web la possibilitá di disegnare la tua scarpa preferita, personalizzandone i colori, le scritte ed il look totalmente a discrezione del cliente. E' possibile modificare ogni aspetto: suola, lacci, punta, ecc, e poi anche salvare il risultato finale sul computer o condividerlo sui social network.

5. Grasshopper, azienda statunitense che fornisce servizi di telefonia online per le piccole aziende, possiede una sezione Happy Customers (clienti felici) sul sito, dove i clienti raccontano la loro esperienza di utilizzo dei prodotti dell'azienda, e come questi li abbiano portati a migliorare il proprio business.

6. Fiat crea spesso dei blog, ad es. Quelli che Bravo, dove interagisce con gli appassionati del modello o comunque del marchio cercando di valorizzare il prodotto. Con questa attivitá l'immagine del marchio é migliorata e in piú riesce a raccogliere feedback preziosi sui propri modelli.

7. Lego ha creato una community (www.lugnet.com) composta da un pubblico appassionato adulto, attraverso il quale é riuscita ad individuare migliaia di ambasciatori del marchio lego e a trovare individui che potessero dare giudizi produttivi da utilizzare nella creazione di nuovi prodotti

8. Dell possiede un forum comunitario dedicato alla risoluzione dei problemi tecnici e grazie a questo gli ha permesso di sfruttare le competenze della comunitá e migliorare l'esperienza d'uso dei clienti facendogli risparmiare tempo e denaro.

2.5 SENTIMENT ANALYSIS

Il termine Sentiment Analysis si riferisce all'applicazione di elaborazione del linguaggio naturale e analisi del testo per identificare ed estrarre informazioni soggettive in contenuti di base. In generale, ll sentiment analysis mira a determinare l'atteggiamento di un oratore o di uno scrittore rispetto a qualche argomento o la polaritá complessiva contestuale di un documento. L'atteggiamento puó essere il suo giudizio o una valutazione,lo stato affettivo, cioé lo stato emotivo dell'autore durante la scrittura, oppure la comunicazione emotiva, vale a dire, l'impatto emotivo che l'autore desidera avere sul lettore. Risulta importante cercare di capire come si puo' presentare un testo e secondo quali aspetti andare ad analizzarlo per poter definire il cosiddetto sentiment. Le caratteristiche che un testo che esprime un'opinione puo' avere sono:

- Tipo di valutazione: sostanzialmente ci sono due modi principali per esprimere sentimenti che sono, un'opinione o confronto diretto. I pareri diretti di solito nel descrivere un oggetto utilizzano alcuni aggettivi che fanno riferimento ad esso, esempio la qualitá dell'immagine di questo televisore é ottimo. Al contrario, i prospetti comparativi, parlano di piú di un oggetto e descrivono qualche tipo di relazione tra loro, ad esempio la qualitá delle immagini del televisore A é molto meglio di quella del televisore B).

- Tipo di contesto: Per estrarre un parere corretto si deve sapere qual é il contesto in cui il testo o la frase vengono inseriti. A seconda della posizione o del portale nel quale le informazioni sono descrittive, esse possono avere significati molto diversi. Su portali di recensioni spesso é relativamente facile estrarre le informazioni relative al sentiment perché si conosce il soggetto della recensione, mentre per esempio su un forum, é molto piú difficile individuare l'oggetto di discussione o oggetto di un singolo post. Infatti tutti i vari software e i vari algoritmi sono creati e ottimizzati per effettuare il riconoscimento automatico del sentiment in contesti specifici. Secondo alcuni studi applicare la sentiment analysis ad un testo in uno specifico contesto ha tipicamente un intervallo di precisione dal 70% al 80% a seconda della quantitá di input e del tipo di testo. I motori e gli

algoritmi generici funzionano molto peggio e sono in grado di analizzare solo particolari tipi di testo.

- Livello d'interesse: Le persone possono esprimere le proprie opinioni con dettagli diversi. Alcuni danno informazioni di carattere generale, mentre altri provvedono ad un analisi piú approfondita. Inoltre una persona puo' giudicare un prodotto con solo una breve descrizione, mentre altri approfondire molto di piú alcune caratteristiche. Questo fattore ha una particolare importanza durante la classificazione complessiva dell'orientamento del testo. Si deve giudicare se separare le frasi si riferiscono allo stesso attributo / oggetto o diversificare. Allo stesso modo a seconda del parere dell'utente, una frase puó esprimere molte opinioni all'interno.

- Formula d'interrogazione: A seconda della persona e il luogo in cui le persone condividono le loro opinioni, le dichiarazioni e le richieste possono essere espressi in modo diverso. Alcuni utenti tendono ad usare parole chiave o frasi brevi, mentre altri forniscono un testo piú completo

- Tipo di vocabolario: le opinioni possono essere espresse in diversi modi a seconda della modalitá di utilizzo del vocabolario. Si possono usare parole che si riferiscono direttamente al soggetto della frase, vale a dire penso che questo articolo é orrendo!, o parole che esprimono piú le emozioni che possono essere molto piú difficili da riconoscere, esempio Mi piace il modo in cui questo telefono funziona! o Sono rimasto sbalordito nel vedere tutti quegli effetti speciali).

- Modo di esprimersi: una frase puo' presentare contenuti o giudizi espliciti oppure impliciti.

Un altro aspetto da definire é il tipo di valutazione dobbiamo trarre dal nostro testo, questa puo' essere:

- Scalare: cioé assumere un valore binario, positivo o negativo, oppure un valore su una scala Likert, cioé un valore da 1 a 5 che mi rappresenta l'intensitá di positivitá(o di negativitá) del testo.

- Multidimensionale: il giudizio viene valutato su piú caratteristiche che si possono individuare sul soggetto della valutazione

L'analisi del sentiment puo' essere effettuata su tre livelli differenti:

1. Livello di Documento: l'obiettivo é quello di analizzare un intero documento o un testo composto da piú frasi e definire una valutazione complessiva di esso attraverso un giudizio discreto, positivo, negativo o neutro

2. Livello di frase: l'obiettivo é quello di analizzare frase per frase il testo e identificare quale di queste esprimono un giudizio, positivo o negativo

che sia, dell'autore e quali invece siano semplicemente descrittive, quindi effettuare una distinzione tra gli aspetti soggettivi e obiettivi

3. Livello di Caratteristica: l'obiettivo é quello di non dare solo un parere positivo o negativo globale ma di estrapolare i singoli giudizi relativi a singoli aspetti dell'articolo che é oggetto del testo o dell'argomento in questione

2.5.1 TEORIE DI RICERCA DEL SENTIMENT

Di seguito sono presentate alcune teorie di soluzione di ricerca del sentiment collaudate da alcuni ricercatori [13]:

- L'analisi secondo Turney si basa essenzialmente sulla misurazione della distanza della polaritá tra gli aggettivi presenti nel testo e quelli presenti in una lista predefinita. Gli aggettivi presenti in questa lista sono aggettivi attinenti al contesto di analisi e che hanno una polaritá ben definita, o estremamente positiva o estremamente negativa. Turney definisce 3 step per l'analisi di un documento senza la supervisione umana:
  - Vengono estrapolati tutti gli aggettivi dal testo fornendo peró all'algoritmo il contesto in cui si colloca il documento. Gli aggettivi vengono identificati secondo pattern predefiniti
  - Misurazione dell'orientamento semantico. Questa viene effettuata misurando la distanza tra l'aggettivo e quelli presenti nella lista predefinita con polaritá nota. Le mutue dipendenze tra le due parole vengono scovate attraverso il calcolo degli hit che si ricavano dalle ricerche sul motore di ricerca AltaVista in documenti che possiedono le due parole con una certa prossimitá l'una dall'altra
  - Un algoritmo calcola la media dell'orientamento semantico di ogni coppia di parole e identifica e classifica il documento come raccomandato o no.

- L'analisi secondo Pang é un lavoro basato sul classico tema delle tecniche di classificazione. L'approccio proposto mira a testare se attraverso un gruppo selezionato di machine learning algorithms possono produrre risultati quando la sentiment analysis é percepita come l'analisi di un documento con due temi, positivo o negativo. I risultati sono stati presentati in 3 varianti in base ai seguenti algoritmi:
  - Naive Bayes
  - Maximum Entropy
  - Support Vector Machine

- Riloff e Wiebe incentrarono il loro studio sull'analisi della frase, piú propriamente andare ad identificare la soggettivitá o l'obiettivitá di una frase. Proposero l'utilizzo di un algoritmo che possedeva un'altra precisione e un basso recall nell'individuazione delle frasi soggettive. Gli step di questo studio erano:
  - Suddividere ed etichettare le frasi del testo in fortemente soggettive o altamente obiettive. Le frasi che non si riuscivano ad etichettare con certezza venivano scartate. I classificatori associavano le etichette grazie ad una lista predefinita di parole che indicavano la soggettivitá. Questi classificatori avevano una precisione intorno al 90%.
  - Utilizzare i dati raccolti per la formazione di un algoritmo di estrazione che genera modelli per frasi soggettive. I modelli sono utilizzati per estrarre piú frasi dallo stesso testo. Il metodo proposto é applicato in modo tale da aumentare il recall. Durante la fase di esecuzione l'algoritmo utilizza una serie predefinita di modelli sintattici che vengono confrontati con le frasi soggettivi (vedi Figura **??**). Dopo il training set l'elaborazione dei modelli estratti sono ordinati in base alla loro frequenza di occorrenza e, secondo alcune condizioni predefinite, solo i modelli migliori sono selezionati per iterazioni successive all'analisi di base del testo (vedi Figura **??**).

SYNTACTIC FORM	EXAMPLE PATTERN
<subj> passive-verb	<subj> was satisfied
<subj> active-verb	<subj> complained
<subj> active-verb dobj	<subj> dealt blow
<subj> verb infinitive	<subj> appear to be
<subj> aux noun	<subj> has position
active-verb <dobj>	endorsed <dobj>
infinitive <dobj>	to condemn <dobj>
verb infinitive <dobj>	get to know <dobj>
noun aux <dobj>	fact is <dobj>
noun prep <np>	opinion on <np>
active-verb prep <np>	agrees with <np>
passive-verb prep <np>	was worried about <np>
infinitive prep <np>	to resort to <np>

Figura 2.9: Syntactic Template

- Yu e Hatzivassiloglou studiarono il rilevamento sia oggettivitá/soggettivitá delle frasi, sia il loro orientamento (positivo, negativo, neutro). Per il rilevamento della soggettivitá usarono tre tipi di algoritmi:
  - Sentence similarity detection

PATTERN	FREQ	%SUBJ
<subj> was asked	11	100%
<subj> asked	128	63%
<subj> is talk	5	100%
talk of <np>	10	90%
<subj> will talk	28	71%
<subj> put an end	10	90%
<subj> put	187	67%
<subj> is going to be	11	82%
<subj> is going	182	67%
was expected from <np>	5	100%
<subj> was expected	45	42%
<subj> is fact	38	100%
fact is <dobj>	12	100%

Figura 2.10: Patterns Frequences

- Naive Bayens classification
- Multiple Naive Bayens classification

  Per la definizione dell'orientamento utilizzarono una tecnica molto simile a quella presentata per Turney, l'unica differenza principale stava nel fatto che utilizzarono piú categorie, quindi piú parole, nella lista predefinita di aggettivi, cioé non solo associati ai due gruppi estremamente positiva o estremamente negativa ma anche a orientamenti intermedi.

Un altro tipo di analisi piú dettagliata é quella della definizione del sentiment a livello de caratteristiche del soggetto di un articolo. Essendo piú utile é anche il piú difficile da eseguire. L'obiettivo é quello di determinare non solo la soggettivitá del testo e la polaritá, ma anche ció che, in particolare, all'autore del testo é piaciuto dell'oggetto. Tipiche di queste analisi sono le seguenti attivitá:

- estrarre le caratteristiche degli oggetti che vengono commentati
- determinare l'orientamento dei pareri (positivo / negativo / neutro)
- individuare gruppi di caratteristiche similari e produrre una sintesi (vedi Figura 2.11)

Questo tipo di analisi vengono fatte spesso solo su specifici tipi di testo soprattutto dal punto di vista del formato, come ad esempio le recensioni che

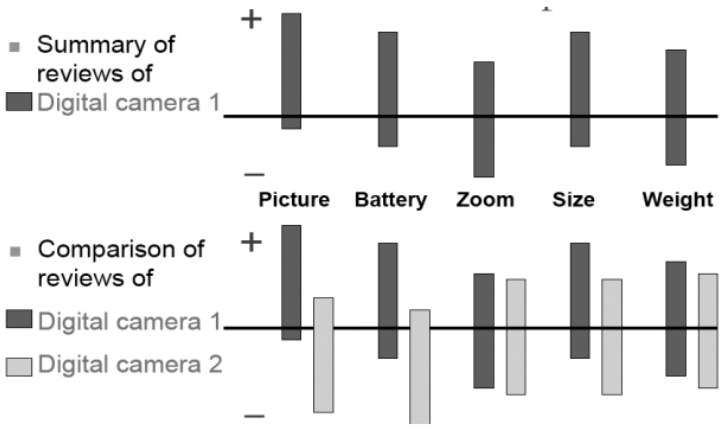


Figura 2.11: Sentiment in base alle caratteristiche

possiedono un'area contenente gli aspetti positivi e un'altra con gli aspetti negativi. In piú nella valutazione finale vengono tenute in considerazione solo le caratteristiche piú frequenti presenti in piú valutazioni.

2.5.2 AMBITI DI UTILITÁ DELLA SENTIMENT ANALYSIS

La sentiment analysis é uno strumento che ha una grossa potenzialitá ed é sempre piú frutto di studi e di utilizzo. Ma a cosa realmente puo' servire? Di seguito sono presentate alcune delle principali utilitá di applicazione della sentiment analysis:

- Benchmarking di prodotti e conoscenza del mercato La chiave per vendere un prodotto sta nel saper risponde alle esigenze dei clienti in un tempo immediato e nel giusto modo. Molte aziende spendono quantitá enormi di soldi l'analisi del mercato e assumono numerosi consulenti esterni alla societá. L'utilizzo della sentiment analysis potrebbe aiutare in questo ambito e permetterebbe di minimizzare i costi e potrebbe creare un vantaggio competitivo per la societá nei confronti dei suoi concorrenti potendo piú rapidamente rispondere alle esigenze dei clienti.

- Collocamento della pubblicitá sul web Al giorno d'oggi é divenuta molto importante anche la pubblicitá sul web, peró essendo il pubblico molto vasto, é importante cercare di far arrivare la pubblicitá a chi possa essere veramente interessato. Questo puo' avvenire grazie all'analisi del sentiment perché ci permette di capire quali sono i contesti del contenuto che si trova su certi blog o forum e posizionare la pubblicitá dove si parla di argomenti attinenti al soggetto della propaganda pubblicitaria. Inoltre é anche opportuno capire se in quel sito, o blog, o forum, ecc si parli in modo

positivo o negativo della categoria di oggetti oppure dell'oggetto in questione perché sarebbe totalmente controproducente inserire la pubblicitá di un oggetto su un forum dove si parla male del medesimo

- Motori di ricerca del parere Spesso prima di effettuare un acquisto si cerca di capire se questo abbia degli aspetti negativi tali da far passare in secondo piano i pregi, per questo spesso si va alla ricerca di giudizi, pareri, recensioni sull'oggetto che si vuole acquistare. Potrebbe risultare molto utile avere un servizio che dato in input il nome di un oggetto, mi restituisca solo i pareri e i giudizi relativi a questo oggetti in modo tale da poter fare delle valutazioni.

- Rilevamento delle opinioni spam Le opinioni spam sono il risultato diretto della popolaritá degli user generated content. I pareri forniti dagli utenti su diversi prodotti e servizi hanno guadagnato un'enorme valore commerciale nel corso degli ultimi anni. Internet é spietatamente utilizzato, proprio come altri mezzi di comunicazione, come un fronte di battaglia tra le imprese e le societá per accaparrarsi clienti. Per questo, non a caso, i sistemi che consentono di inviare opinioni vengono abusati nel loro utilizzo. Infatti spesso capitano commenti falsi o comunque vi é una certa disinformazione volontaria atta a trarre in inganno il potenziale cliente a fare un acquisto o a non effettuarne un altro. Questo puó essere fatto sia automaticamente, ma anche dagli esseri umani. In scala ridotta é facile moderare la possibilitá di commentare certi oggetti, tuttavia nei forum grandi e popolari, bacheche o anche negozi in internet, questo puó essere molto difficile. La creazione di sistemi che rilevano i commenti falsi sui prodotti potrebbero migliorare la credibilitá di qualsiasi portale della comunitá aumentando cosí le vendite potenziali per oggetti che sono realmente validi e affidabili. Tuttavia questo dominio non é ancora molto discusso. Uno dei motivi é che i clienti ingannevoli cercano sottilmente di penalizzare o criticare l'oggetto e quindi é difficile rilevare l'intenzionalitá dolosa del commento, sia da parte di una macchina ma anche da parte di un essere umano.

2.5.3 TOOL PER LA SENTIMENT ANALYSIS

In questa sezione verranno presentati alcuni tool esistenti, creati da vari gruppi di ricerca finanziati da grosse imprese per il rilevamento del sentiment [14]:

- BlogPulse o NM Incite: nata dalla Joint venture tra Nielsen e McKinsey si occupa di trovare soluzioni di marketing attraverso i social media. Fornisce analisi dei volumi di dati, confronti tra vari brand, individuazione dei cosiddetti opinion leader e possiede un modulo per la sentiment analysis. Tutti questi dati sono ricavati da una vasta gamma di sorgenti tra cui milioni di blog forum e altri user-generated content

- Addictomatic: cerca i migliori siti in diretta sul web per le ultime notizie, blog, video e immagini. Le pagine di notizie forniscono le ultime news su temi come spettacolo, politica, shopping, sport e altro ancora. é possibile ottenere tutti i feed dalle migliori fonti di ogni categoria. E come dashboard dei risultati di ricerca, é possibile personalizzare il layout delle caselle principali, cancellare quelli che non ti piacciono e aggiungere ai preferiti la pagina personalizzata. I risultati vengono estratti via rss dai migliori notiziari e siti di ricerca sul web tra cui Google, Yahoo, Technorati, Ask, YouTube, Truveo, Flickr, Blinkx, Ice Rocket, Digg, Topix, Newsvine e Tweetscan.

- Alterian: consente ai clienti di prendere decisioni attraverso e fornendo analisi e previsioni fornendo supporto decisionale, strategico, consulenza e ascolto intorno a soluzioni personalizzate. Nello spazio dei social media, dove il personale con competenze adeguate é spesso difficile da assumere e costoso. Fornisce soluzioni di servizi confezionati intorno alla piattaforma tecnologica SM2 per guidare le azioni di marketing, indipendentemente dal canale. Social Media Insights servizi sociali trasforma i dati multimediali in business intelligence perseguibile. Alterian permette di rispondere a queste domande:

  - Che cosa dicono del tuo brand?
  - Chi sono i fattori di influenza e come si puó raggiungere?
  - Come funziona il vostro marchio in confronto con la concorrenza?
  - Ricerche di mercato: Che cosa si puó imparare su un determinato argomento? Rapporto di Ricerca di mercato
  - Come puo' essere fatta una campagna o il lancio di un prodotto sui social media?

- Opinion Finder: E' un sistema che esegue l'analisi soggettivitá identificando automaticamente quando sono presenti in un testo opinioni, sentimenti, speculazioni ecc. In particolare, si propone di individuare e contrassegnare nelle frasi soggettive i vari aspetti della soggettivitá, tra cui la fonte della soggettivitá e le parole che sono incluse nelle frasi che esprimono sentimenti positivi o negativi. Possiede due modalitá di utilizzo, batch e interactive. La batch prende in input una lista di documenti da processare mentre la interactive fornisce un front-end che permette all'utente di interrogare on-line le fonti di notizie per i documenti da elaborare.

- Trendrr: Presenta soluzioni per il pubblico attraversano una vasta gamma di dispositivi e servizi di social media per aiutare ad identificare la migliore esperienza sociale per raggiungere i gli obiettivi e realizzare le idee su tutte le piattaforme multimediali - on-air, on-line, o mobile.

- Media Analysis Platform (MAP): É il software realizzato dalla Sysomos, compagnia canadese, che estrae e analizza il contenuto dei social media o di qualsiasi user-generated content. Permette di realizzare in tempo reale un quadro della situazione per quanto riguarda la reputazione di prodotti, persone o brand sui Social Media. Inoltre cerca di capire come ha avuto origine e come si é sviluppata la reputazione analizzando le conversazioni su vari argomenti e scovando i cosiddetti Influencer chiave. Inoltre é in grado di presentare i risultati ottenuti segmentandoli secondo aree geografiche, classi demografiche, secondo le lingue, ecc.

- Cogito: Rappresenta la tecnologia semantica progettata dall'azienda italiana Expert System per effettuare vari tipi di analisi tra cui la sentiment analysis. Cogito possiede un ricco database lessicale che gli permette di leggere e comprendere correttamente i testi, identificare informazioni e relazioni nascoste, trasformare i dati non strutturati in dati strutturati, rendendoli cosí immediatamente fruibili. Il core di Cogito é il Sensigrafo,che consiste in una vasta rete semantica multilingue in cui le parole sono rappresentate secondo il concetto che esprimono e sono legate fra loro anche in base al significato che possiedono. Possiede la capacitá di identificare il significato corretto di una parola in base al contesto (es. panda: auto o animale?) e quindi rende possibile la disambiguazione dei testi. Il Sensigrafo contiene piú di 1 milione di concetti e 4 milioni di relazioni. Ogni concetto, poi, é caratterizzato da diversi attributi che concorrono insieme a rendere l'identificazione del contesto il piú corretta possibile, con enormi vantaggi in termini di precisione e accuratezza nell'ambito di tutte le attivitá legate alla gestione delle informazioni.

- Blogmeter: E' un servizio di ascolto dei social media fondato su una piattaforma tecnologica proprietaria, che permette di analizzare ció che viene detto online su un tema, un'azienda, un brand o un personaggio pubblico. Utilizza una metodologia che definisce un processo iterativo di ascolto, comprensione e analisi che consente di misurare, attraverso l'analisi della passaparola online, l'efficacia sia da un punto di vista qualitativo che quantitativo delle strategie di comunicazione dell'azienda. La metodologia di analisi del passaparola online, sulla quale si fonda la piattaforma BlogMeter, puó dunque essere suddivisa in tre fasi principali:

  - Ascolto: la piattaforma Blogmeter, grazie allo sviluppo di appositi sistemi di acquisizione dati, preleva dal web tutti i contenuti user generated potenzialmente interessanti

  - Comprensione: il motore proprietario di analisi semantica viene utilizzato per strutturare e classificare le conversazioni secondo le dimensioni di analisi previste

– Analisi: la piattaforma web based di analisi e reportistica consente all'utente di navigare le conversazioni in maniera strutturata e aggregare le dimensioni in piú modi.

Figura 2.12: Blogmeter

- Liquida: E' un sito internet made in italy che puo' essere definito come un aggregatore di User-Generated Content cioé rileva dai social network le varie notizie, individua quelle attendibili, le ordina dandogli una struttura e le presenta come se fosse un editoriale. Inoltre permette di effettuare ricerche mirate su oggetti, brand, persone che si possono indicare nella casella di ricerca e mi permette di visualizzare il trand del soggetto e il sentiment, positivo o negativo, estratto dalla rete

- TweetFeel: Tool di sentiment analisysis basato su Twitter. Calcola il sentiment di brand, oggetti, personaggi attraverso il confronto delle keyword di ricerca con degli indicatori passando attraverso algoritmi complessi cercando di fornire risultati il piú attendibili possibili. Fornisce una casella di testo dove inserire le proprie keyword di ricerca da includere ma consente anche di specificare le keyword da escludere.

- Sentimetrix: Permette il tracciamento delle opinioni, su qualsiasi argomento, espresse sui mezzi di informazione online, nei blog e nei forum, nelle recensioni dei utenti ed anche nei database forniti dai clienti. Fornisce anche un'analisi del sentiment su scala continua e non solo positivo o negativo. Inoltre fornisce l'accesso ai dati raccolti da una moltitudine di sorgenti da tutto il mondo, inclusi piú di 50.000 news/media outlets ed un milione di top blog. Fonti aggiuntive personalizzate possono essere aggiunte su richiesta. Questa soluzione é composta essenzialmente da due parti:

  - SentiGrade Dashboard - un'interfaccia web-based semplice da usare che permette di tracciare fino a 10 argomenti contemporaneamente, confrontare i cambiamenti di opinione nel tempo in base all'argomento, alla lingua ed alla fonte dell'informazione.
  - SentiGrade Data Service - serie di servizi web per i clienti che giá possiedono giá soluzioni proprie e desiderano aggiungere ai propri dati le informazioni sulle opinioni. SentiGrade fornisce l'accesso ai dati raccolti da una moltitudine di sorgenti da tutto il mondo, inclusi piú di 50.000 news/media outlets ed un milione di top blog. Fonti aggiuntive personalizzate possono essere aggiunte su richiesta.

- Radian6: Strumento che permette di definire la web reputation di una azienda. Effettua analisi sia in tempo reale che con dati storici sfruttando un'ampia gamma di fonti e permette di effettuare comparazioni fra aziende.

- Jive Social Business Software(Jive SBS): E' un tool di knowledge management che monitora I Social Media e registra le discussioni definendo il sentiment di ogni post,stabilisce i trand e indica i vari Influencer.

2.5.4 CASI REALI DI PREDIZIONE ATTRAVERSO SOCIAL MEDIA E UGC

L'utilizzo degli user-generated content come oggetti utili ad analisi previsionali ebbe giá un notevole inizio intorno all'inizio dell'anno 2000 quando ancora non esistevano i social network come Facebook e Twitter. Di seguito vengono presentati alcuni casi di utilizzo dei contenuti del web per effettuare previsioni in ordine cronologico:

1. Il primo caso fu studiato da Gruhl che fece una analisi per ricercare una correlazione tra il materiale che si trovava su internet e i risultati di vendita dei libri. Non essendoci ancora i Social Network la base che c'era in questo esperimento erano dati non strutturati che si ricavavano dal web, come blog, forum e pareri espressi in giro per la rete e confrontati coi dati di vendita di Amazon. Si notó che per certi libri, soprattutto quelli in voga nel periodo, c'erano dei picchi di vendita che erano correlati e combaciavano con dei picchi di citazione o riferimenti online.

2. Altro caso piuttosto particolare é lo scenario che ha raccontato Hal Varian capo economista di Google dove affermó che attraverso Google Trends[15], strumento che confronta volumi di ricerche in base a vari fattori, era possibile prevedere la crisi finanziaria del 2008. Piú precisamente secondo l'economista americano, la crisi finanziaria del 2008 era scritta nelle ricerche effettuate dagli utenti: In quel periodo abbiamo riscontrato su Google

Trends prima un calo della fiducia dei consumatori, poi del commercio al dettaglio e infine una riduzione dei consumi. Attraverso l'analisi e la correlazione delle informazioni, avremmo dunque potuto capire cosa stava succedendo.

3. Il caso studiato da Mishne e Glance consiste nel portare avanti lo studio presentato nel primo caso di Gruhl ma aggiungendo il concetto di sentiment analysis, infatti effettuó l'analisi sui film al botteghino cercando una correlazione tra le vendite e il volume di riferimenti sul web relativi ai film. In piú introdusse il concetto di sentiment analysis perché andó a cercare anche una correlazione tra le vendite e i pareri positivi o negati che trasparivano dal web sui film. E scoprí che questa correlazione esisteva e che soprattutto il sentiment che si ricavava nel periodo dell'anteprima del film era significativo sull'andamento delle vendite totali.

4. Il caso piú recente riguarda le primarie politiche italiane del Partito Democratico [16]. Infatti la societá Voices From the Blog ha applicato la sentiment analysis sui Tweet rilevabili al Social Network Twitter ricavando e stilando delle percentuali di previsione di voto. Il risultato é stato quello di presentare i sondaggi pre voto piú vicini alla realtá di quelli svolti da altre societá secondo altri criteri. Di seguito, In Figura 2.13, sono mostrate le percentuali previste da Voices From the Blog e quelle reali di voto e si nota come la differenza sia minima

2.5.5 SENTIMENT ANALYSIS SUPERVISIONATO

In questa sezione viene presentato un approccio alla sentiment analysis alternativo, applicato a Twitter, pubblicato sul sito lospaziodellapolitica.com e che utilizza la societá, anche precedentemente citata, Voices From the Blog [16] nei suoi sondaggi. La sentiment Analysis supervisionata é una tecnica che coniuga l'efficacia di un'analisi manuale con la rapiditá e l'efficienza dei metodi automatizzati di analisi statistica e in questo modo permette di ottenere risultati molto piú accurati rispetto alle tradizionali tecniche di sentiment analysis. L'analisi cerca di analizzare le preferenze sull'intera popolazione di tweet relativi a un preciso argomento. Ad esempio si prenda in esame il tema della politica. é importante osservare peró che, nonostante molti utenti dei social network discutano di temi diversi rispetto alla politica, il numero di tweet che tratta di questioni politicamente o socialmente rilevanti risulta particolarmente alto. Per analizzare questi dati in modo supervisionato viene utilizzato un processo a due stadi. Nel primo stadio un gruppo di coder ha il compito di leggere e codificare manualmente un sottogruppo dei documenti scaricati dalla rete (da varie prove é stato osservato che codificare manualmente un numero di commenti, compreso ad esempio tra 600 e 1000 tweet, sia sufficiente per permettere all'algoritmo di

Figura 2.13: Previsioni Twitter Primarie

comprendere il linguaggio utilizzato dagli utenti per trattare quel determinato tema, dato che nella realtá il linguaggio con cui ci si esprime in rete é meno variopinto di quel che si potrebbe pensare). Questo sottogruppo rappresenta il training set che verrá utilizzato successivamente da un algoritmo statistico per classificare l'insieme dei documenti che non sono stati letti nella seconda fase dell'analisi. I codificatori umani sono ovviamente piú effettivi rispetto ai dizionari ontologici nel riconoscere le sopradiscusse specificitá del linguaggio e per comprendere meglio la posizione di chi scrive un tweet rispetto a un dato argomento (per tornare all'esempio di prima, un codificatore umano é in grado di capire che quando si scrive ma che bella fregatura! ovviamente si sta esprimendo un parere negativo). In secondo luogo, una codifica manuale é anche in condizioni migliori per identificare eventuale spam presente tra i messaggi postati. Un vantaggio decisivo dato l'impatto che un elevato volume di spam puó avere sull'accuratezza del risultato finale. Nel secondo stadio dell'analisi, l'algoritmo statistico estende questa accuratezza prodotta da una codifica manuale all'intero universo dei post scaricati dalla rete, permettendo di catturare l'opinione di

tutti quelli che scrivono, con un errore atteso molto contenuto (tra il 2 e il 3%). C'é da sottolineare che tale algoritmo non cerca di classificare ciascun singolo tweet o post, per poi aggregare tali post successivamente. Quello che fa é ricostruire direttamente la distribuzione aggregata di preferenze presente nella rete. Sembra una distinzione da poco ma é rilevante. Immaginiamo di voler stimare quanto mangime consumano i pesci rossi contenuti in un acquario. Possiamo tentare di osservare ogni singolo pesce nel momento in cui si nutre, misurando quando cibo consuma, ma cosí rischiamo di sovrastimare o sottostimare il dato di ciascuno. In alternativa, e piú semplicemente, possiamo guardare quanto mangime é rimasto nel barattolo. In questo senso l'approccio aggregato é in grado di dare, nel modo piú efficiente, la risposta alla nostra domanda quale sono le preferenze complessive della popolazione che scrive on-line?. Nel corso del 2012 Voices from the Blogs ha realizzato diverse previsioni, sui piú svariati argomenti, utilizzando Twitter e l'approccio di Sentiment Analysis appena descritto. E ogni volta queste previsioni si sono rivelate molto accurate. Si va dalle tracce del tema di italiano alla maturitá (crisi economica e nuove tecnologie erano infatti le tracce piú gettonate on-line e sono puntualmente state scelte come tema d'esame), al vincitore di Sanremo sia nella edizione 2012 (VfB ha pronosticato correttamente il podio finale) che in quella 2013 (previsti da VfB correttamente i primi due). Ancora piú interessanti sono i risultati ottenuti nelle previsioni elettorali. VfB ha infatti predetto la vittoria di Hollande in Francia, di Obama nelle presidenziali USA e di Bersani nelle primarie del centrosinistra. Ma il dato piú sorprendente é che spesso le stime di VfB sono state piú precise di quelle dei sondaggi. Emblematico il caso americano. Il giorno prima delle elezioni VfB ha annunciato il successo di Obama nei tre stati chiave (Florida, Ohio e Virginia) pronosticato una vittoria anche nel voto popolare con un margine del 3,5%. A conti fatti, il divario tra Obama e Romney é stato del 3,85%, vicinissimo alla previsione di VfB, e i nostri dati sono stati gli unici a prevedere correttamente la vittoria di Obama nei tre piú importanti swing states, risultando piú affidabili rispetto a quasi tutti i sondaggi e alle stime degli analisti (con l'unica eccezione di Nate Silver del New York Times) che mostravano invece una prevalenza di Romney in Florida ed un sostanziale pareggio in Virginia.

Capitolo 3

MODELLO DEL SISTEMA DI ANALISI

3.1 INTRODUZIONE

L'obiettivo principe di questo lavoro consta nel cercare di scovare una certa correlazione tra dati di vendita e i dati ricavabili dai Social Media, piú propriamente Twitter, sia per quanto riguarda i volumi in modo generico, sia per quanto riguarda il sentiment. Per fare tutto ció come prima cosa, é stato necessario andare a definire in quale ambito andare ad effettuare questo studio. Per far si che l'analisi risulti efficace é importante che l'ambito e il contesto in cui si effettua lo studio sia qualcosa di uso comune, che un pubblico molto vasto ne sia interessato e ne parli, e che permetta di poter avere senza troppi problemi, accesso ai dati effettivi di vendita da poter compararli con i dati estrapolati dalla rete sociale. Inizialmente i settori presi in esame sono stati:

- **auto motive**, perchè sui social media si parla molto di auto, si commentano, si giudicano e si creano anche gruppi di fan nei confronti di modelli o di marchi automobilistici. Il pregio di questo settore stava nella grande mole di interesse sociale che ha e quindi un numero di dati elevato estraibile dalla rete. Di contro, in questo settore risultava difficile ricavare in modo automatico, affidabile, veritiero e aggiornato nel breve periodo, i dati di vendita

- **abbigliamento moda**, settore molto "social", infatti molti internauti si interessano, seguono o commentano articoli di abbigliamento, calzature o brand relativi a stilisti o case di moda. Per questo in rete e nei social media esistono numerosi dati relativi a questi oggetti permettendo di avere volumi elevati e dati di sentiment abbastanza rappresentativi. I difetti di questo settore sono sempre l'impossibilitá di poter avere accesso ai dati di vendita effettivi in tempo reale, ma anche per colpa della cosiddetta stagionalitá del settore. Con stagionalitá s'intende la differenza di volumi

e di interessi di certi articoli o brand in base alla stagione climatica che si ha in quel momento, ad esempio articoli come giacconi e stivali sono acquistati e anche "twittati" con piú frequenza in stagioni fredde mentre oggetti come occhiali da sole o costumi da bagno, nelle stagioni calde

Infine, come successivo e settore d'analisi é stato considerato e definitivamente accettato e utilizzato per il nostro studio, il settore **cinematografico**, piú precisamente i **film** in uscita nelle sale. Questo settore si presta molto bene alla nostra analisi perchè é un settore che coinvolge qualsiasi categoria di persone, dai giovani ai vecchi, dalle donne agli uomini, é spesso oggetto di recensioni, opinioni e pareri sul web specialmente sui social media, ed é un settore frequentato tutto l'anno, ovviamente con i suoi picchi in determinati momenti, ad esempio nel periodo natalizio. In piú, oltre ad essere adatto per le sue qualitá sociali, é idoneo per quanto riguarda la possibilitá di cercare, ricavare ed estrapolare i dati di vendita reali ed effettivi delle pellicole cinematografiche perchè sono resi pubblici in tempi immediati. Inoltre questo settore, proprio a causa dei pregi precedentemente elencati, é giá stato motivo di studio come presentato nel capitolo 2.

3.2 DATI DI VENDITA

Come affermato nel precedente paragrafo, i dati di vendita relativi ai film al cinema sono pubblicati da vari enti e possono essere ricercati facilmente. In questo progetto la fonte dei dati di vendita dei film nei cinema italiani, ma anche la fonte dei film in uscita, é il sito **Mymovies.it**. MYmovies é il piú consultato database italiano online di cinema, nasce nel 2008, ed il servizio appartiene a Mo–Net S.r.l. (Multimedia–Online–Network), con direzione, redazione e uffici a Firenze e sede legale a Milano. Nella societá partecipa IBS (InternetBookShop) del gruppo Emmelibri. Con 50 milioni di pagine viste e oltre di 7 milioni di utenti unici al mese (Nielsen SiteCensus, gennaio 2011), MYmovies.it si attesta come il primo sito italiano di cinema per visitatori. Raccoglie circa 45.000 articoli tratti dalle maggiori testate giornalistiche e oltre 150.000 tra recensioni e commenti dei lettori. Nell'archivio dei film sono proposte in forma digitale le versioni cartacee di alcuni dizionari di cinema piú importanti:

- Il Farinotti Dizionario di tutti i Film: oltre 40.000 schede dal 1895 ad oggi
- Il Morandini Dizionario dei film 2007 (Zanichelli) circa 20.000 schede di film
- La Garzantina del Cinema
- Dictionnaire des films di Gorge Sadoul –Tra i piú grandi critici cinematografici di sempre

- Il Dizionario degli Attori: oltre 65.000 schede personali, molte delle quali corredate da biografia
- Il Dizionario dei Registi: oltre 20.000 schede, molte delle quali corredate da biografia
- Dizionario dei Professionisti del cinema: sceneggiatori, soggettisti, fotografi, scenografi ecc.
- Il Dizionario dei Premi: tutti i premiati nei maggiori festival mondiali dal 1928
- Il Dizionario delle Colonne Sonore: oltre 1.300 schede con informazioni tecniche e tracklist
- Il Dizionario dei Libri da cui é stato tratto un film
- Il Dizionario dei Telefilm (Garzanti) di Leopoldo Damerini e Fabrizio Margarina
- Dizionario della TV di Giorgio Carbone e Leo Pasqua
- Scegliere un film (edizioni '04 '05 '06 '07 '08) di Armando Fumagalli e Luisa Cotta Ramosino
- Lupu ululá e castello ululí. Le migliori battute del cinema di Daniele Soffiati (Comix Mondadori)

Nella sezione critica vengono inoltre riportate recensioni da testate come Corriere della Sera, la Repubblica, Le Monde, The New York Times ed altre. Sebbene si siano avvicendati diversi critici a dare il loro contributo alla recensione delle uscite cinematografiche, negli ultimi anni la redazione critica é stata ridotta e concentrata solo attorno ad alcuni piú selezionati nomi: Edoardo Becattini, Marianna Cappi, Gabriele Niola, Giancarlo Zappoli e Marzia Gandolfi. Fanno invece parte della squadra piú saltuariamente (in occasione di Festival o con qualche ruolo specifico) Tirza Bonifazi Tognazzi, Nicoletta Dose, Mattia Nicoletti e Emanuele Sacchi. All'interno del sito é presente la sezione *Box Office* che permette di accedere ai dati di vendita. Questi dati sono pubblicati settimanalmente e sono suddivisi in due sezioni:

- Classifica dei primi 20 film col maggior incasso ai botteghini in Italia
- Classifica dei primi 20 film col maggior incasso ai botteghini negli USA Ogni film presente in classifica, oltre a presentare la posizione in classifica (per noi ininfluente), ci fornisce i seguenti dati:
- Titolo Italiano
- Titolo Originale (solo se differente dal titolo italiano)

- Regista
- Genere
- Stato di produzione
- Anno
- Vendita, in euro o in dollari, del week–end, considerando quindi solo il venerdí, sabato e domenica della settimana corrente
- Vendita totale, in euro o in dollari, dall'uscita fino alla fine della settimana corrente

Essendo una classifica, sono presentati, per quella settimana, solo i primi 20 film che hanno incassato maggiormente e non tutte le vendite di tutti i film presenti al cinema. Peró questo non compromette l'utilitá di questi dati perchè i film che non rientrano in questa classifica sono pellicole che hanno venduto cosí poco da non influenzare il nostro studio, anche perchè essendo cosí poco visti, sarebbero praticamente ignorati anche dai Social Media. A causa di questa particolaritá avremo film di cui avremo dati di vendita per piú di 3 settimane, altri di una sola settimana e cosí via, e considereremo la scomparsa del film dalle classifiche come se il film non fosse piú nelle sale, proprio perchè anche se nella realtá potrebbe essere ancora in proiezione, i risultati di vendita sono irrisori e quindi non considerabili. Avendo questi dati a disposizione, attraverso semplici calcoli possiamo ricavare le seguenti informazioni:

- Vendita totale del film nell'arco della sua presenza in sala, ricavabile andando ad estrarre il campo "Vendita totale" dell'ultima settimana in cui appare in classifica
- Vendita totale della settimana, andando a sottrarre dal campo "Vendita totale" dall'uscita fino alla settimana corrente, il corrispettivo campo della settimana precedente
- Vendita del week–end, campo esplicitamente fornito da Mymovies
- Vendita dei giorni da lunedí a giovedí, ricavabile sottraendo alla vendita totale settimanale, la vendita del week–end.

3.3 ANTEPRIME FILM

Per poter andare ad estrarre i commenti e i pareri degli utenti del web é necessario conoscere, in anticipo rispetto all'uscita, i film presenti nei cinema italiani. E' importante saperli in anticipo perchè spesso il maggior volume di pubblicitá e di pareri si ha prima dell'uscita ed é anche il momento piú importante, dove si va a raccogliere il pubblico della prima settimana che é quella con maggior

affluenza. In piú, come per molti altri tipi di oggetti, piú persone guardano il film, piú persone sono attirate dal vederlo, quindi la fase anticipatoria é molto importante e significativa. Per conoscere i film in uscita si é utilizzato, per quanto riguarda i film in Italia, il sito Mymovies.it che presenta una sezione dal nome "Film Uscita" che é composta da 12 schede, ognuna rappresentante i mesi dell'anno,e selezionando il mese, presenta l'elenco dei film che escono in quella mensilitá in ordine cronologico. Per ogni film, Mymovies, indica in anteprima i seguenti dati:

- Titolo
- Regia
- Attori principali
- Genere
- Stato di produzione
- Anno
- Durata
- Data di uscita

In piú aggiunge una breve descrizione del film e il parere dei critici di Mymovies, piú precisamente se é consigliata o meno la visione. Per quanto riguarda i film americani Mymovies non fornisce informazioni sulle uscite, quindi per ricavare i dati dei film in uscita degli USA si é utilizzato Comingsoon.net che presenta sul proprio sito una sezione "Movies" all'interno del quale si ha la voce "Release Dates" dove selezionando il mese desiderato viene presentato l'elenco dei film, raggruppati per giorno d'uscita. I titoli dei film sono rigorosamente quelli originali e che sono utilizzati in America per poter appunto utilizzarli per la nostra ricerca dei giudizi. Comingsoon.net fornisce anche informazioni dettagliate sui film in uscita che sono:

- Titolo
- Data di uscita
- Studio
- Regista
- Sceneggiatori
- Attori
- Genere

- Trama

Inoltre sono presenti altri informazioni meno importanti come la valutazione MPAA del film e Sito Web ufficiale, ecc.

3.4 MODELLO QUERY PER L'ESTRAZIONE DEI TWEET

Il passo successivo é stato quello di pensare al modo con cui estrarre i tweet da Twitter e quindi ad un modello di query adatto per ogni film da poter utilizzare nella ricerca dei tweet a loro inerenti. Infatti, andare ad utilizzare il solo titolo del film, risultava poco esaustivo e in alcuni casi portava ad avere come risultato tweet per nulla attinenti con il contesto dei film e del cinema in generale. Infatti, prendendo ad esempio come film la pellicola "Lo Hobbit", utilizzando solo il titolo avremmo praticamente la totalitá dei tweet che hanno come soggetto il film, mentre se prendiamo in considerazione il film dal titolo "Colpi di fulmine" avremmo certamente tweet che trattano del lungometraggio ma anche molti altri che avrebbero come oggetto della questione il colpo di fulmine nel contesto amoroso. Inoltre é necessario considerare che le persone sono spesso attratte o meno da un film a seconda del regista che ha prodotto la pellicola oppure dagli attori presenti. Per questo motivo é stata pensata ed utilizzata una query un po' complessa ma che tiene conto di questi aspetti in modo tale da raccogliere il maggior numero di tweet attinenti al film in oggetto. Come prima cosa vediamo come é stata definita la query per il crawler dei tweet italiani. La generazione e la strutturazione di questa query é visibile nel diagramma seguente:

Come si puo' notare, la prima cosa che viene svolta é quello di controllare se il titolo del film contiene il carattere ":" o il carattere "–" perchè parecchi film, soprattutto quelli facente parte di trilogie o saghe, hanno un titolo costituito come prima_parte_titolo –seconda_parte_titolo, ad esempio "Rec 3 –La Genesi" oppure come prima_parte_titolo : seconda_parte_titolo, come "Cirque du Soleil 3D: Mondi Lontani". In questo caso, essendo un titolo lungo e articolato, ricercarlo direttamente per intero porterebbe a meno risultati di quelli che in realtá esistono e per questo, il titolo viene spezzato in due parti, come ad esempio in "Rec 3" e "La Genesi". Se il titolo non contiene nessuno dei due caratteri viene mantenuto nella sua forma originale. Successivamente si legano attraverso un AND logico il titolo o le parti del titolo con alcune parole chiave fisse in modo tale da avere le seguenti porzioni di query:

- TITOLO AND "cinema"
- TITOLO AND "film"
- TITOLO AND "visto"

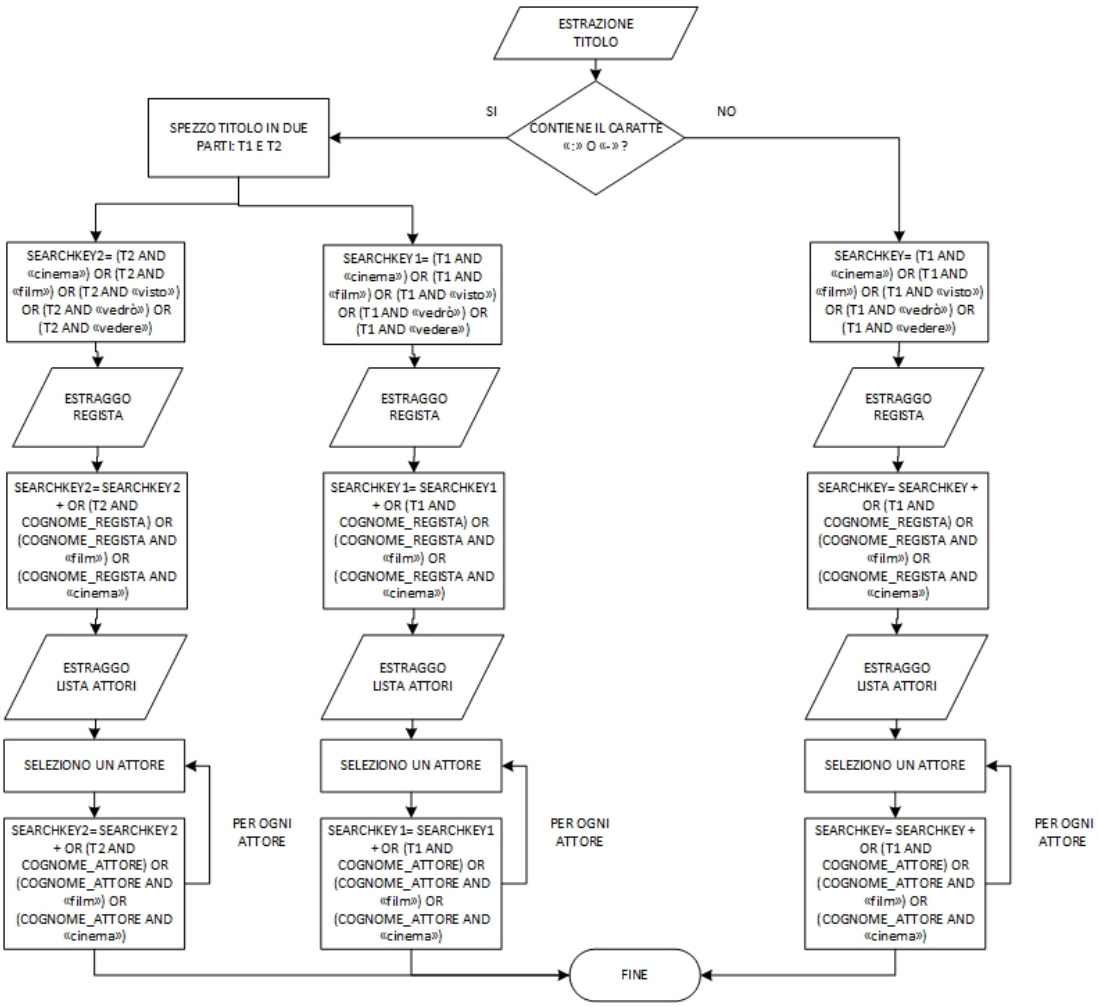


Figura 3.1: Organigramma Query di estrazione dei tweet da Twitter

- TITOLO AND “vedró”
- TITOLO AND “vedere”

Sono state scelte queste parole perchè, analizzando vari tweet presenti sulla rete, risultavano le parole piú frequenti e che maggiormente caratterizzavano i messaggi sulle pellicole cinematografiche. Le porzioni di query precedentemente presentate sono concatenate tra loro attraverso un OR logico. In seguito utilizziamo anche il regista del film, piú precisamente solo il cognome, perchè l’utilizzo anche del nome risultava maggiormente vincolante mentre il suo inutilizzo non comporta la perdita di nessun tweet (é molto difficile che una persona si riferisca ad un regista chiamandolo solo per nome). Quindi, andiamo ad aggiungere in OR con le altre porzioni di query, le combinazioni:

- TITOLO AND COGNOME_REGISTA
- COGNOME_REGISTA AND “film”

- COGNOME_REGISTA AND “cinema”

Infine, come ultima parte della query, andiamo a prendere gli attori che recitano nel film in oggetto, ne estraiamo solo il cognome per lo stesso motivo presentato per il regista, e lo concateniamo con le parole “cinema” , “film” e con il titolo o le porzioni del titolo del film in modo tale da avere, per ogni attore del film, come porzioni della query:

- COGNOME_ATTORE AND TITOLO
- COGNOME_ATTORE AND “film”
- COGNOME_ATTORE AND “cinema”

Tutto questo porta ad avere alla fine una query molto lunga che ad esempio, per il film di Giuseppe Tornatore, “La migliore offerta”, si presenta nel seguente modo: (La migliore offerta AND cinema) OR (La migliore offerta AND film) OR (La migliore offerta AND visto) OR (La migliore offerta AND vedró) OR (La migliore offerta AND vedere) OR (La migliore offerta AND Tornatore) OR (Tornatore AND film) OR (Tornatore AND cinema) OR (La migliore offerta AND Rush) OR (Rush AND cinema) OR (Rush AND film) OR (La migliore offerta AND Sturgess) OR (Sturgess AND cinema) OR (Sturgess AND film) OR (La migliore offerta AND Hoeks) OR (Hoeks AND cinema) OR (Hoeks AND film) OR (La migliore offerta AND Sutherland) OR (Sutherland AND cinema) OR (Sutherland AND film) OR (La migliore offerta AND Sutherland) OR (Sutherland AND cinema) OR (Sutherland AND film) OR (La migliore offerta AND Jackson) OR (Jackson AND cinema) OR (Jackson AND film) Invece, per film aventi il titolo suddiviso in due parti, viene duplicate la query, creandone prima una con la prima parte del titolo e poi una seconda con la restante parte. Per quanto riguarda il crawler relativo ai tweet americani, semplicemente si sono utilizzate delle parole chiave inglesi corrispondenti a quelle individuate nella lingua italiana e per questo la query in lingua inglese si presenta come la concatenazione attraverso l’OR logico delle seguenti parti:

- TITOLO AND “cinema”
- TITOLO AND “film”
- TITOLO AND “movie”
- TITOLO AND “movies”
- TITOLO AND “see”
- TITOLO AND “saw”
- TITOLO AND “see”

- TITOLO AND COGNOME_REGISTA
- COGNOME_REGISTA AND “film”
- COGNOME_REGISTA AND “cinema”
- COGNOME_REGISTA AND “movie”
- TITOLO AND COGNOME_ATTORE
- COGNOME_ATTORE AND “film”
- COGNOME_ATTORE AND “cinema”
- COGNOME_ATTORE AND “movie”

Ovviamente, le parti relative al cognome dell’attore, risultano replicate per ogni attore partecipante al film. Problematiche e casi particolari La query presentata riesce a coinvolgere gran parte dei tweet di nostro interesse ma capita di raccogliere tweet inutili o non includerne altri. Di seguito sono presentati le casistiche piú evidenti:

1. Estrarre tweet sintatticamente corretti ma che semanticamente non sono inerenti al film in oggetto di studio. Ad esempio prendendo come regista Quentin Tarantino, avremmo una parte della query costituita da “Tarantino AND film” e questo comporta l’estrazione anche di tweet del tipo “Questa sera resto a casa a vedermi un bel film in dvd di Tarantino”. Ovviamente questo post non appartiene ad un film presente nelle sale cinematografiche di Tarantino e quindi non andrebbe considerato. Questa problematica accade poco frequentemente grazie alla proprietá di twitter di non mostrare tweet piú vecchi di 6–9 giorni e quindi mostrare tweet molto attuali. Questo permette di avere tweet che sono inerenti ai film che sono presenti nelle sale in quel momento e non a film di quel regista usciti nelle sale mesi addietro.

2. Tweet che non hanno nulla a che fare con i film e il cinema. Questo accade con determinati titoli di film, piú precisamente quando il titolo é composto da singole parole o da brevi frasi di uso comune, ad esempio per il film “La madre”, attraverso la porzione di query TITOLO AND “film”, andiamo ad includere nella nostra ricerca tweet come “La madre di Luca non ci fa vedere i film horror”. Purtroppo questa casistica non é correggibile automaticamente perchè dipende strettamente dal titolo del film

3. Non estrazione di tweet che fanno riferimento al film utilizzando parole sintatticamente differenti al titolo del film ma che alludono ad esso. Ad esempio il tweet “Mi é piaciuto molto l’ultimo film di James Bond” é chiaramente indirizzato alla pellicola dal titolo “Skyfall” ma dalla nostra

query non sarebbe stato rilevato. Anche questa casistica é imprevedibile automaticamente.

4. Non estrazione di tweet che fanno riferimento al film utilizzando soltanto parte del titolo. Un esempio di questa casistica si puo' fare per il film dal titolo "Django Unchained", infatti molte persone nei loro tweet fanno riferimento al film utilizzando solo la parola "Django".

5. Altri casi particolari come ad esempio l'abitudine a scrivere nomi degli attori in modo scorretto, un esempio lampante riguarda Leonardo DiCaprio il cui cognome andrebbe scritto attaccato (infatti nell'estrazione dei dati da Mymovies sugli attori rileviamo il cognome scritto attaccato), ma moltissime persone pensano o erroneamente lo scrivono staccato.

Per cercare di ovviare in qualche modo alle ultime tre casistiche presentate, nel nostro crawler si dá la possibilitá di inserire manualmente delle query proprio per poter recuperare una certa quantitá di dati utili al nostro progetto.

3.5 DATABASE DEI DATI SCARICATI DAI CRAWLER

In questa sezione verrá presentato lo schema del database che andrá a contenere i dati che ricaveremo dai crawler che realizzeremo e che ci serviranno ad effettuare le analisi, sia del sentiment, sia delle correlazioni vendite/volumi e vendite/sentiment. Il diagramma ER del nostro database é mostrato nella Figura seguente:

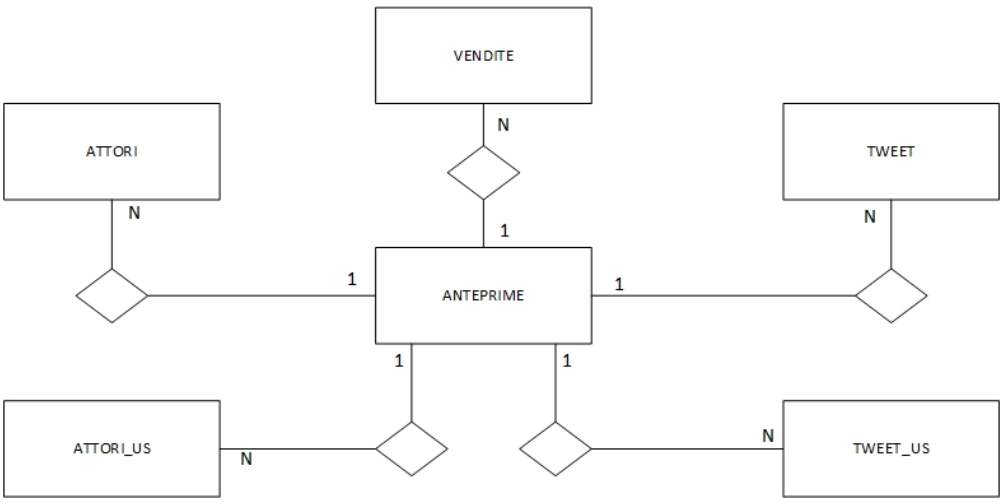


Figura 3.2: Diagramma del database contenente i dati del progetto

Come si puo' notare dal diagramma ER le tabelle significative sono Anteprime, Vendite, Tweet e Attori, le tabelle Attori_US e Tweet_US sono identiche

ad Attori e Tweet ma contenenti le informazioni relative ai dati americani. Dal diagramma si nota come tutte le relazioni tra le tabelle siano del tipo 1 a molti peró é necessario specificarne il perchè per la relazione Anterpime/Attori. Infatti la logica dice che questo collegamento debba essere rappresentato da una relazione molti a molti perchè un attore puo' recitare in piú film e in un film recitano piú attori, peró in questo caso, per questioni di semplicitá, é stata utilizzata una semplice relazione 1. In piú il nostro database contiene solo informazioni relativi a circa i primi 3 mesi dell'anno e in questo breve periodo sono molto rari i casi in cui ci siano al cinema piú film con gli stessi attori in ruoli tali da comparire negli attori principali. La tabella Anteprime rappresenta l'elenco dei film che verranno estratti attraverso il crawler anteprime. Insieme alla tabella vendite, contiene i dati sia relativi ai film in uscita in America sia ai film in uscita in Italia e contiene i seguenti campi:

- IdAnteprime: Chiave primaria e quindi identificativo univoco di ogni record all'interno della tabella
- Titolo: Titolo del film
- Mese: Mese in cui il film uscirá nelle sale cinematografiche
- Regista: Regista del film
- Nazione: Indica se il film in questione appartiene alle anteprime in Italia o a quelle statunitensi, infatti puo' assumere solo due valori, ITA o USA.

La tabella Vendite contiene i dati di vendita dei film che varranno estratti attraverso il crawler vendite dal sito Mymovies e racchiuderá sia i dati italiani che quelli americani. La tabella é costituita dai campi:

- IdVendite: Chiave primaria e quindi identificativo univoco di ogni record all'interno della tabella
- Titolo: Titolo del film
- Regista: Regista del film
- Genere: Genere attribuito al film
- Nazione: Nazione di produzione del film
- Anno: Anno di produzione
- Vendita_weekend: Quantitá in euro incassata dal film nel weekend
- Vendita_totale: Quantitá in euro incassata dal film in tutta la settimana
- Stato: indica se i dati di vendita sono relativi al botteghino italiano o a quello americano. Puo' assumere due valori, ITALIA o USA.

- Data: Rappresenta la data della domenica della settimana di cui sono indicati i dati di vendita. E' la data che viene passata come parametro all'url della pagina di Mymovies all'avvio del crawler
- Titolo_originale: Titolo originale del film. Se il titolo é uguale al titolo originale, il campo é vuoto.

La tabella Tweet racchiude tutti i tweet che saranno estratti grazie al crawler che verrá realizzato attraverso l'utilizzo delle API di Twitter. I tweet che conterrá sono quelli italiani mentre la corrispondente tabella Tweet_us conterrá quelli in lingua inglese. Queste tabelle contengono come campi:

- IdTweet: Chiave primaria e quindi identificativo univoco di ogni record all'interno della tabella
- Testo: Il messaggio di testo contenuto nel tweet
- Data: Timestamp che mi indica la data e l'instante in cui é stato pubblicato il tweet
- Titolo: Titolo del film al quale il tweet é riferito

La tabella Attori conterrá tutti i principali attori che hanno preso parte nei film e che saranno estratti attraverso il crawler anteprime. In questa tabella saranno contenuti gli attori che recitano nei film che troviamo nelle sale cinematografiche italiane, mentre la tabella Attori_us conterrá gli attori che hanno recitato nelle pellicole in uscita nei cinema americani. Queste tabelle possiedono i seguenti campi:

- IdAttore: Chiave primaria e quindi identificativo univoco di ogni record all'interno della tabella
- Nome: Indica il nome e il cognome dell'attore
- Titolo: titolo del film nel quale ha recitato

Capitolo 4

IMPLEMENTAZIONE DEL SISTEMA

4.1 INTRODUZIONE

In questo capitolo andremo a vedere come concretamente é stato realizzato il progetto, piú precisamente come sono stati realizzati i crawler che mi permettono di estrarre i dati utili alla nostra analisi e che sono stati presentati nel capitolo precedente. In piú, nell'ultima parte viene presentato lo strumento utilizzato per l'estrazione del sentiment dai tweet e come é stato settato ed utilizzato

4.2 CRAWLING

Un Crawler é un software che analizza i contenuti di una rete (o di un database) in un modo metodico e automatizzato, in genere per conto di un motore di ricerca [17]. I crawler solitamente acquisiscono una copia testuale di tutti i documenti visitati e le inseriscono in un indice. Un uso estremamente comune dei crawler é nel Web. Sul Web, il crawler si basa su una lista di URL da visitare fornita dal motore di ricerca (il quale, inizialmente, si basa sugli indirizzi suggeriti dagli utenti o su una lista precompilata dai programmatori stessi). Durante l'analisi di un URL, identifica tutti gli hyperlink presenti nel documento e li aggiunge alla lista di URL da visitare. Il processo puó essere concluso manualmente o dopo che un determinato numero di collegamenti é stato seguito. In Figura 4.1 viene presentata l'architettura base di un Crawler Web.

Di seguito viene presentata una tabella contenente i nomi dei crawler associati ai vari motori di ricerca:

Esistono principalmente due tipologie di Crawler:

1. Parser HTML: I crawler basati su parser HTML effettuano un'analisi della struttura della pagina Web attraverso i cosiddetti parser che permettono

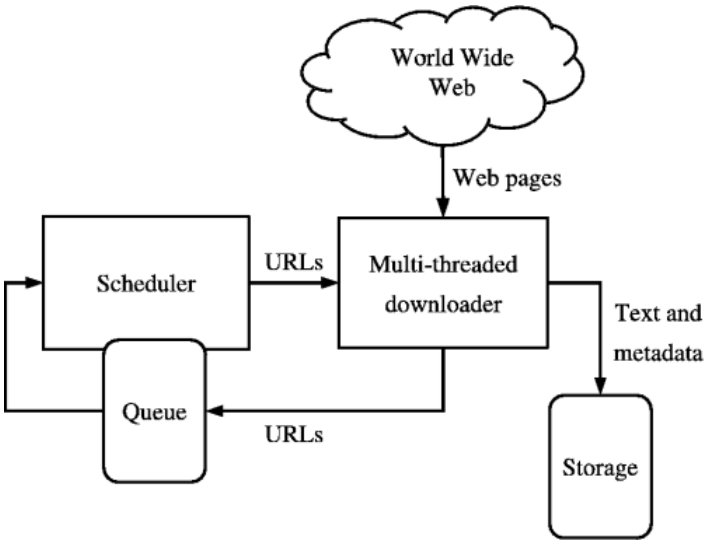


Figura 4.1: Architettura Web Crawler

l'identificazione dei tag HTML che si trovano all'interno della pagina, cosí da capire quali sono i contenuti e i link. I contenuti utili vengono salvati in un database mentre, una volta analizzata tutta la pagina, si passerá alla visita di tutte le pagine a cui fanno riferimento i link nella pagina corrente. Piú in breve,partendo dalla prima pagina, si visitano tutte le pagine connesse a questa e si estraggono i contenuti. Questa tecnica permette di effettuare ricerche molto complete sui vari siti Web, ma impiega piuttosto tempo per elaborare tutte le pagine interconnesse tra loro.

2. API - Application Programming Interface: si indica ogni insieme di procedure disponibili al programmatore, di solito raggruppate a formare un set di strumenti specifici per l'espletamento di un determinato compito all'interno di un certo programma. Spesso con tale termine si intendono le

CRAWLER	MOTORE DI RICERCA
Googlebot	Google
Fast	Fast - Alltheweb
Slurp	Inktomi - Yahoo!
Scooter	Altavista
Mercator	Altavista
Ask Jeeves	Ask Jeeves
Teoma agent	Teoma
Ia archiver	Alexa - Internet Archive
Yahoo! Slurp	Yahoo
Romilda	Facebook

Tabella 4.1: Tabella dei crawler con i rispettivi motori di ricerca

librerie software disponibili in un certo linguaggio di programmazione. La finalitá é ottenere un'astrazione, di solito tra l'hardware e il programmatore o tra software a basso e quello ad alto livello semplificando cosí il lavoro di programmazione. Le API permettono infatti di evitare ai programmatori di riscrivere ogni volta tutte le funzioni necessarie al programma dal nulla, ovvero dal basso livello, rientrando quindi nel piú vasto concetto di riuso di codice. Le API stesse rappresentano quindi un livello di astrazione intermedio. Solitamente é un servizio che é fornito dal proprietario della piattaforma o del sito proprio per garantire l'utilizzo al massimo delle potenzialitá e delle funzionalitá della piattaforma stessa. L'utilizzo di un crawler attraverso API é piú veloce ma i dati ricavabili possono subire limitazioni/restrizioni che sono imposti dal produttore. Esistono due linee di condotta per quanto riguarda la pubblicazione delle API:

(a) Alcune compagnie conservano gelosamente le loro API. Per esempio, la Sony e la Microsoft forniscono le API ufficiali solamente agli sviluppatori per rispettivamente PlayStation e X-Box registrati. Questo perché ambedue intendono restringere il numero di persone che possono scrivere giochi per le rispettive console in modo da ricavare il piú possibile. Questo atteggiamento é tipico delle compagnie che non traggono profitto dalla vendita dell'implementazione delle API.

(b) Altre compagnie distribuiscono le API pubblicamente. Per esempio la stessa Microsoft fornisce al pubblico informazioni sulle sue Windows API, tra cui le Microsoft Foundation Classes (MFC), per scrivere software per la piattaforma Windows. La vendita di applicazioni di terze parti fa vendere copie di Windows. Questo é tipico di compagnie che traggono profitto dalla vendita dell'implementazione delle API.

4.2.1 JSOUP

Jsoup é parser HTML5 (conforme alle specifiche WHATWG) le cui API consentono di estrarre dati e manipolare in maniera estremamente semplice documenti sfruttando le potenzialitá di DOM, CSS e metodi di accesso simili a quelli offerti da JQuery. In Jsoup un documento xml é rappresentato dalla classe *Document*, i nodi del documento da istanze della classe *Node* e i tag da istanze della classe *Element* Esistono tre diversi modi per effettuare il parsing di un documento xml ed ottenere cosí un *Document* che lo rappresenti:

- caricare il documento da un URL
- caricare il documento da un File
- effettuare il parsing di una String rappresentante il contenuto del documento xml o una parte di questo

Nel primo caso si utilizza il metodo connect della classe Jsoup:

Document doc = Jsoup.connect(http://html.it/).get();

Nel secondo caso si utilizza il metodo parse della classe Jsoup passando come argomenti il File di input, il charset e il base-uri con il quale risolvere gli indirizzi relativi:

File input = new File(documento.html);
Document doc = Jsoup.parse(input, UTF-8, http://html.it/);

Nel terzo caso si utilizza il metodo parse della classe Jsoup passando una String rappresentante il contenuto del documento, e il base-uri:

String html = < html > < head > < title > Documento di prova < /title > < /head > < body > Corpo del body < /body > < /html > ;
Document doc = Jsoup.parse(html, http://html.it/);

É possibile effettuare anche il parsing di un frammento di codice HTML mediante il metodo *parseBodyFragment*: *String html = DIV inserito ;*
Document doc = Jsoup.parseBodyFragment(html);
Element body = doc.body();

Una volta ottenuta l'istanza della classe *Document* é possibile procedere alla navigazione del documento sfruttando diversi metodi. Ad esempio se vogliamo accedere alla sezione *title* del nostro documento html basta utilizzare il metodo *title()* del *Document*, stessa cosa per il body e le altre sezioni. In piú per recuperare l'elemento-figlio i-esimo di un altro si usa il metodo *child(int index)*:

Element figlio = document.body().child(0);

per recuperare l'elemento adiacente successivo o precedente ad un dato elemento si usano i metodi *nextElementSibling()* e *previousElementSibling()* Per recuperare tutti gli elementi di un dato tag si usa *getElementsByTag()* passando come argomento il tag desiderato:

Elements elementidiv = document.getElementsByTag(div);

come risultato si ottiene un oggetto *Elements* che rappresenta una collezione di *Element* (tanti quanti sono gli elementi con dato tag presenti all'interno del documento). Per recuperare un elemento a partire dall'id si usa il metodo *getElementById()* passando come argomento l'id desiderato:

Element div2 = document.getElementById(div2);

Jsoup mette a disposizione anche due metodi per recuperare elementi che abbiano un dato attributo (a prescindere dal valore)

Elements elementi = documento.body().getElementsByAttribute(id);

o che abbiamo un dato attributo valorizzato in un modo ben preciso:

Elements elementi = documento.body().getElementsByAttributeValue(id,div2);

Per recuperare l'*Element* padre di un dato *Element* si usa il metodo *parent()*:

Element padre = figlio.parent();

Questi sono i principali modi per accedere agli *Element* presenti all'interno di un *Document*. Se si vuole accedere agli attributi di un Element si usano invece i metodi:

- attributes(): recupera tutti gli attributi di un *Element* restituendo un oggetto *Attributes* che é una collection di *Attribute*

  Attributes attributi = element.attributes();

- attr(): recupera l'attributo il cui nome é passato come argomento:

  Attribute attributo = element.attr(id);

Accanto ai metodi tradizionali per l'accesso agli elementi di un documento html, Jsoup introduce uno strumento molto potente per recuperare uno o piú elementi attraverso l'uso di selettori.

Un selettore é un'espressione che individua un sottoinsieme degli elementi presenti in un documento. Una volta definito il selettore é possibile utilizzare il metodo *select(String selettore)* di un *Element* per selezionare gli elementi individuati dal selettore ottenendo come risultato una collection *Elements*. Il selettore viene costruito considerando che:

- nometag: recupera tutti gli elementi che hanno il tag nometag

  body.select(div)

- num-id: recupera l'elemento che ha il dato id

  body.select(num-div2);

- .nomeclasse: recupera gli elementi che hanno come classe css nome classe

  body.select(.rosso);

- [nome attributo]: recupera gli elementi che hanno l'attributo specificato

  body.select([id])

- [nome-attributo=valore-attributo]: recupera gli elementi che hanno per l'attributo indicato il valore indicato

  body.select([id=div2]);

Ovviamente é possibile utilizzare un selettore composto da una o piú delle precedenti regole. Ad esempio supponendo di voler recuperare tutti gli elementi con tag *div* e classe css *rosso*, utilizzeremo il selettore *div.rosso*. Accanto ai metodi per selezionare e navigare gli elementi di un documento xml Jsoup mette a disposizione un insieme di metodi per manipolarne il contenuto:

- html(String html): modifica l'Element sostituendolo con il codice passato come argomento. element.html(< p > paragrafo < /p >);
- prepend(String html): aggiunge l'Element di cui viene passato il codice in testa agli elementi figlio dell'elemento dato element.prepend(< p > paragrafo iniziale < /p >);
- append(String html): aggiunge l'Element di cui viene passato il codice in coda agli elementi figlio dell'elemento dato element.append(< p > paragrafo finale < /p >);
- element.attr(class,rosso): modifica gli attributi di un Element

4.2.2 CRAWLER VENDITE

Per poter estrarre i dati di vendita dei film in modo automatico é stato realizzato un'applicazione che effettua il parsing della pagina html del sito di Mymovies contenente la classifica dei 20 film che hanno avuto il maggior incasso della settimana. Questa applicazione é stata realizzata in Java e sono state sfruttate le funzionalitá contenute nella libreria Jsoup, che come illustrato precedentemente,é una libreria che permette di effettuare in modo relativamente semplice il parsing di una pagina html. Il sito Mymovies pubblica le vendite settimanalmente e per questo motivo, l'applicazione chiede di inserire in input la data della domenica della settimana di cui si vogliono scaricare le vendite al botteghino.

Viene indicata la domenica perché Mymovies, nell'URL delle pagine che caratterizzano ogni classifica settimanale, vi é la data della domenica della settimana in esame, ad esempio per le vendite della settimana che va dall'11 marzo al 17 marzo, l'URL della pagina di Mymovies é:

http://www.mymovies.it/boxoffice/italia/top20/?weekend= **17/03/2013**

Avendo l'URL possiamo utilizzare la funzione parse di Jsoup per tramutare la pagina html in un Documet:

Document docita = Jsoup.parse(new URL(url-mymoves), 2000);

Una volta entrati in possesso dell'oggetto Document é necessario individuare le parti utili della pagina che contengono le informazioni e i dati di vendita dei film. Andando ad analizzare il codice html della pagina si é individuato, come si puo' notare dalla figura sottostante, il tag che racchiude le informazioni salienti. Questo tag, con i relativi attributi é:

div[style=margin-bottom:10px; margin-top:3px; padding-left:5px;]

Come si nota nella in Figura 4.2 , ogni Element mi rappresenta i dati di un

```
<div>
  <table border="0" cellpadding="0" cellspacing="0" style="font-size:95%;">

      <tr valign="top">
      <td align="right" valign="top" style="height:230px;">
      <strong style="color:#ff0066; font-size:30px">1 </strong>
      </td>
      <td style="width:100%; font-size:12px; color:#191919" class="linkblu">
      <div style="margin-bottom:10px; margin-top:3px; padding-left:5px;">
      <span style="font-size:17px; font-weight:bold">
      <a href="http://www.mymovies.it/film/2013/ozthegreatandpowerful/">Il grande e potente Oz</a></span>
      <br />(Oz: The Great and Powerful)<br />

          Un film di <strong>Sam Raimi</strong>.<br />
          Fantastico - USA, 2013<br />

              <strong>Week-end € 2.116.009</strong>

                 (totale: 5.701.372)

      </div>
```

Figura 4.2: Html pagina delle vendite di Mymovies

film e la stringa é strutturata:

Titolo (*Titolo Originale*) Un film di *Regista*. *Genere* - *Stato*, *Anno* Week-end *IncassoWeekend* (totale: *IncassoTotale*)

Proprio i campi specificati in corsivo nella stringa precedente, sono i dati che ci servono e che andremo a salvare in un database. Per estrapolarli dall'intera stringa si é semplicemente lavorato con i metodi delle stringhe tenendo presente

proprio come era strutturata e quindi individuando i confini dei parametri. Per quanto riguarda le vendite americane, anch'esse sono estratte da Mymovies e la struttura della pagina é identica a quella italiana e quindi l'unica cosa a cambiare é l'URL di riferimento di partenza, anch'esso dotato del parametro *data* che mi indica la settimana, e che in questo caso, ad esempio é il seguente:

http://www.mymovies.it/boxoffice/usa/top20/?weekend=17/03/2013

4.2.3 CRAWLER ANTEPRIME

L'altra serie di dati di cui abbiamo bisogno sono l'elenco dei film in uscita con le loro rispettive informazioni dettagliate. Come anticipato nelle sezioni precedenti, utilizziamo la sezione dedicata ai film in uscita di Mymovies per i film in Italia, mentre il sito Comingsoon.net per quelli americani. Entrambi i crawler si basano sul concetto utilizzato per il crawler delle vendite, cioé effettuare il parsing della pagina html contenente le informazioni utili attraverso l'uso della libreria Jsoup.

Crawler anteprime Italia

Questo crawler come prima cosa, attraverso il metodo parse di Jsoup, converte la pagina html relativa all'elenco dei film in uscita in un oggetto di tipo Document. Per fare ció é necessario dare in input il mese in cui usciranno i film che si vogliono scaricare, questo perché Mymovies struttura i film in uscita in pagine diverse organizzandoli mese per mese, infatti ad esempio l'URL della pagina relativi ai film in uscita a Marzo é:

http://www.mymovies.it/film/uscita/marzo/

Una volta convertita la pagina, é necessario individuare anche qui, i tag che racchiudono le informazioni relativi alle pellicole in uscita. Nella figura seguente viene mostrato l'aspetto della pagina del sito di Mymovies dove si puo' notare come i titoli abbiano uno stile a livello di font diverso dalla parte che caratterizza le informazioni del film, come regia, attori, produzione, ecc. Proprio per questo a livello di codice si é individuato due tipi di selettori e quindi si sono create due tipi di Elements, uno contenente l'elenco dei titoli dei film, l'altro contenente le informazioni relative ai film.

Figura 4.3: Anteprime Mymovies

Andando ad analizzare il codice sorgente della pagina html, come si vede nel seguente codice, si é individuato in $<$ h2 $>$ il tag che racchiude i titoli dei film. Quindi attraverso la linea di codice:

```
<table style="margin-top:7px" cellpadding="0" cellspacing="0" border="0">
<tr>
<td>
<h2><a href="http://www.mymovies.it/film/2012/anijalanave/" title="Anija - La nave"  >Anija - La nave</a></h2>
</td>
<td valign="top" style="width:85px; text-align:center;" >
<div style="height:8px; width:10px"></div>

        <img title="valutazione media tra critica e pubblico: 3,00 stelle" title="valutazione media tra critica e pubb
src="http://pad.mymovies.it/v7/img/stella.gif" />
```

Figura 4.4: Html relativo ai titoli pagina Anteprime di Mymovies

Elements el = d.select(h2);

Dove d rappresenta l'oggetto *Document*. Avendo nell'oggetto el tutti i titoli é bastato semplicemente inserirli in ordine in un vettore e poi passare all'estrazione delle informazioni in dettaglio dei film. Per questo é stato individuato il tag div[class=linkblu]. Peró in questo caso, questo tag non racchiudeva solo i dati a noi utili peró, le informazioni del singolo film avevano una struttura ben precisa e quindi é bastato filtrare i dati a seconda della presenza o meno di questa struttura. Piú precisamente questa struttura é:

Regia di *NomeRegista*. Con *ElencoAttori*. Genere *Generefilm*, produzione *Statoproduzione*, *Anno*. Durata *Minutidurata*. Da *Datauscita* al cinema. *Trama* Di questi dati che sono forniti da Mymovies, estraiamo soltanto:

Figura 4.5: Html relativo ai dati estratti dalla pagina Anteprime di Mymovies

- Titolo
- Regista
- Elenco degli attori

Questo perché, come vedremo poi nelle sezioni successive, saranno gli unici dati che utilizzeremo nella query di interrogazione al database di Twitter. **Crawler anteprime USA** Per questo crawler, come anticipato, utilizziamo Comingsoon.net ma la filosofia di estrazione dei dati risulta essere la stessa dei crawler precedentemente illustrati. Proprio per questo anche in questo caso prendiamo l'URL della pagina html dove Comingsoon mostra le anteprime:

www.comingsoon.net/movies.php?year=2013 & month=03 Anche qui le uscite sono strutturate per mese e nell'URL vi é il parametro rappresentante il numero del mese (nel link precedente vediamo ad esempio indicato il mese di marzo), e per questo, nell'applicazione creata, in input viene chiesto il mese di cui si vogliono scaricare le anteprime. Di seguito viene mostrata la pagina di comingsoon.net perché, a differenza di Mymovies, presenta solo l'elenco dei titoli e solo cliccando sul titolo viene aperta una seconda pagina contenente il dettaglio del film. Per questo, da questa pagina estraiamo il titolo e il link sul quale a sua volta effettueremo il parsing ed estrarremo i dati a noi utili. Per fare

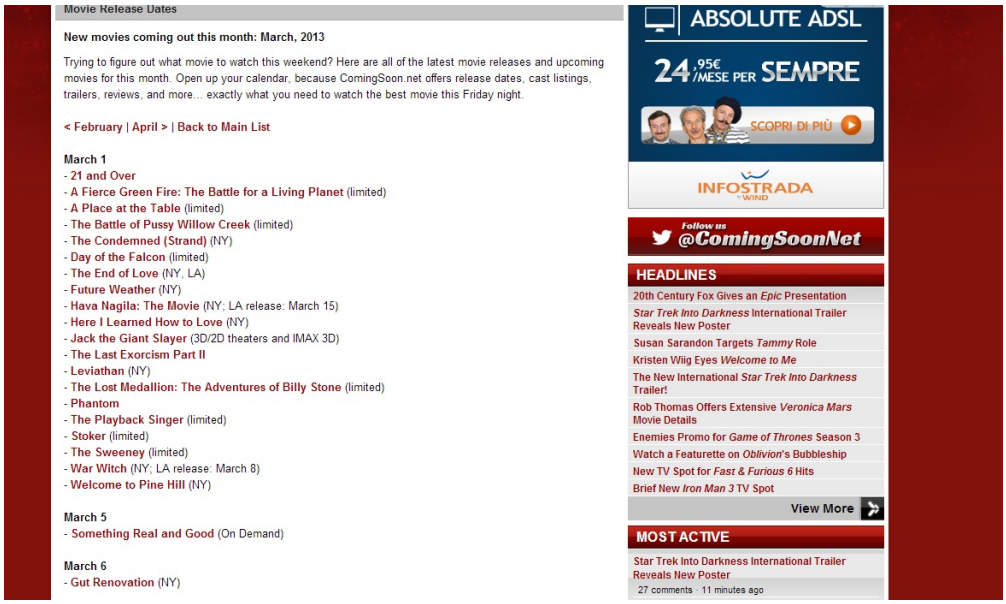


Figura 4.6: Elenco film in uscita nel mese Comingsoon

questo utilizziamo le potenzialitá della libreria Jsoup e come primo passo convertiamo la pagina in un oggetto Document fornendo l'URL con il metodo parse

Document d = Jsoup.parse(new URL(url), 2000); Come secondo passo definiamo gli Elements che conterranno i titoli dei film e per far ció, analizzando il codice sorgente della pagina, abbiamo individuato il tag font > a[href] che ci permette di selezionare solo l'elenco dei titoli.

Elements el = d.select(font > a[href]); Successivamente attraverso un semplice ciclo scorriamo tutti gli element e, ed attraverso due metodi, *text()* e *attr(abs:href)*, estraiamo rispettivamente il titolo del film e il link alla pagina di dettaglio del film e li salviamo in una matrice.

Algorithm 4.1 Algoritmo utilizzato per estrarre il titolo del film e il link alla pagina di dettaglio del film

```
1: for Element e: el do
2:     if i > 2 && i <n - 3 then
3:
4:         filmlink[j][0] = e2.text();
5:
6:         filmlink[j][1] = e2.attr( abs:href );
7:
8:         j++;
9:
```

Figura 4.7: Pagina di dettaglio del film su Comingsoon

Il tag nel quale é racchiuso il testo del corpo della pagina é div[id=subPageContent]. Come si puo' notare dall'immagine precedente, le informazioni sono strutturate come:

- Release Date
- Studio
- Director
- Screenwriter
- Starring
- MPAA Rating
- Official Website
- Review
- DVD Review
- DVD
- Movie Poster
- Production Stills
- Plot Summary

Sfruttando questa strutturazione del testo e vari metodi applicabili alle stringhe riusciamo ad estrarre i contenuti utili. Come per Mymovies, non andiamo a salvare nel database tutti questi dati ma solo quelli che ci saranno utili in seguito che sono sempre:

- Director (Regista)
- Starring (Elenco degli attori)

Oltre che al titolo estratto dalla pagina precedente.

4.3 TWITTER

Twitter é stato creato nel marzo 2006 [19] dalla Obvious Corporation di San Francisco ed é un servizio di social network e microblogging che fornisce agli utenti una pagina personale aggiornabile tramite messaggi di testo con una lunghezza massima di 140 caratteri. Gli aggiornamenti possono essere effettuati tramite il sito stesso, via SMS, con programmi di messaggistica istantanea, posta elettronica, oppure tramite varie applicazioni basate sulle API di Twitter. Il nome Twitter deriva dal verbo inglese to tweet che significa cinguettare. Tweet é anche il termine tecnico degli aggiornamenti del servizio. Gli aggiornamenti sono mostrati nella pagina di profilo dell'utente e comunicati agli utenti che si sono registrati per riceverli. é anche possibile limitare la visibilitá dei propri messaggi oppure renderli visibili a chiunque. Il servizio é diventato estremamente popolare grazie alla semplicitá ed immediatezza di utilizzo. Esistono diversi esempi in cui Twitter é stato usato dagli utenti per diffondere notizie, come strumento di giornalismo partecipativo. Ad esempio, nel caso del terremoto in Abruzzo del 6 aprile 2009, gli utenti Twitter hanno segnalato la notizia prima dei media tradizionali. Twitter deve la sua semplicitá anche alla mancanza di alcune funzioni tipicamente riscontrabili in social-network come Facebook, tra cui la condivisione di fotografie. Tuttavia esistono molti servizi esterni che possono aiutare a potenziare Twitter e ad implementarlo con funzionalitá prossime agli altri Social Network. L'insieme degli status message pubblicati su Twitter dagli utenti costituisce un'enorme quantitá di materiale, che puó essere utilizzata anche dalle aziende: ad esempio Dell ha aperto un canale di comunicazione con i propri clienti su Twitter e molti servizi offrono il monitoraggio della reputazione dei brand su Twitter. Anche in Italia alcune aziende, universitá, scuole e pubbliche amministrazioni utilizzano Twitter a scopi didattici.

Profili

Un profilo twitter rappresenta un individuo all'interno del social network ed é identificato da uno username. Ogni profilo possiede una serie di informazioni che sono:

- Tweet: Rappresenta l'elenco dei messaggi che l'utente ha pubblicato sul social network. Viene anche indicato quanti tweet sono stati pubblicati
- Following: Mi indica il numero di profili che l'utente in questione ha deciso di seguire e dei quali vuole entrare a conoscenza dei messaggi (tweet) che

esso pubblica. Oltre al numero sono indicati anche i piú recenti tweet dei profili al quali ci si é legati

- Follower: Indica quanti e quali profili hanno la volontá di seguire ció che l'utente in questione pubblica su Twitter.

- Preferiti: Sezione in cui vengono visualizzati i tweet che sono stati marcati dall'utente come Preferiti cioé messaggi che sono particolarmente piaciuti o che hanno una certa importanza. Per marcare un tweet come preferito esiste l'opzione sottostante ad ogni singolo tweet per aggiungerlo a questa sezione.

- Liste: sezione che mi mostra in quanti e in quali gruppi il profilo in questione é incluso. Una lista é semplicemente un insieme di contatti che hanno qualcosa in comune. Servono per organizzare i propri contatti

Ad un profilo sono associate anche una foto e delle informazioni personali sulla persona. Non sempre dietro ad un profilo vi é una persona fisica ma a volte sono profili gestiti in automatico da elaboratori solitamente con scopi pubblicitari o di divulgazione di certe notizie. Attraverso Twitter, oltre a scrivere pensieri, opinioni e messaggi generici, rivolti alla rete, é possibile instaurare delle conversazioni con uno o piú utenti. Esistono piú varianti di conversazioni che si possono creare su Twitter:

- Risposta: Consiste nell'inviare un messaggio in risposta ad un tweet, é caratterizzato dall'avere in testa al messaggio la dicitura @username, che rappresenta lo user della persona al quale si sta inviando il messaggio. Questo é visibile a chi é specificato, al mittente e a coloro che essi hanno in comune come following.

- Menzione: Quando viene scritto un tweet é possibile fare riferimento (menzionare) una persona specificando sempre la dicitura @username ma all'interno del messaggio. In questo caso il tweet sará inviato a chi é stato menzionato ma sará visibile anche da tutti gli altri utenti.

- Retweet: Consiste nel ripubblicare un tweet che é stato pubblicato da un altro utente. Questo viene fatto quando si vuole divulgare il contenuto di questo messaggio o perché il contenuto piace particolarmente all'utente che vuole riproporlo.

All'interno di un tweet é possibile menzionare piú di un utente ma anche definire l'argomento del messaggio. Questo viene fatto con il cosiddetto Hashtag, che si specifica con la dicitura # parolachiave; spesso rappresenta il modo migliore per cercare o scoprire di cosa stanno parlando gli utenti di Twitter. Infatti Twitter fornisce un servizio di ricerca all'interno del quale si possono ricercare tweet, di cui si ha il permesso di visualizzazione, nel quale sono contenute certe parole o

frasi oppure che contengono certi Hashtag # parolachiave o certi @username. Twitter pubblica i suoi messaggi in tempo reale e quando si effettuano delle ricerche vengono restituiti sempre i tweet piú recenti in modo tale da avere contenuti sempre aggiornati e di attualitá. Tendenzialmente Twitter non mostra nei risultati delle ricerche, messaggi piú vecchi di una decina di giorni e permette di dire che le informazioni sono altamente volatili. Inoltre é possibile definire anche alcune tipologie di utenti che si trovano su Twitter:

- Amici: coloro che rappresentano persone conosciute nella realtá, non solo amici e conoscenti ma anche famigliari oppure persone conosciute sul social media

- Creatori d'informazione: sono coloro che hanno pochi Following ma tanti Follower questo perché preferiscono scrivere e divulgare informazione anziché riceverla e ascoltare

- Ascoltatori: solo coloro che hanno pochi Follower ma tanti Following questo perché sono molto piú propensi all'ascoltare e captare le notizie sul social media anziché produrre informazione o proporre dibattiti

Tra i creatori d'informazione si possono identificare gli Influencer, essi sono coloro che hanno il potere di influenzare le opinioni altrui. Essi, sfruttando il potere che da Twitter, stanno creandosi una rete di Followers che si fida di loro, ne segue i consigli e le segnalazioni. Questi personaggi hanno un grande potere soprattutto dal punto di vista del marketing. Infatti avere da loro un giudizio positivo o negativo per il proprio brand o prodotto puo' risultare a volte determinanti per il successo/insuccesso del marchio o prodotto.

4.3.1 TWITTER API

Le API di Twitter sono delle Interfacce per la programmazione di applicazioni che permettono di usufruire della grande quantitá di informazioni di cui dispone il Social Network [20]. Esse si suddividono in alcuni tipi:

- Twitter for Websites: una suite per integrare Twitter molto semplicemente nei vari siti web, queste Api sono utilizzabili anche dai meno esperti, ed é molto utile nel caso si volesse inserire il famoso bottone per il following o per mostrare i propri tweet senza dover possedere grandi capacitá di programmazione.

- Search API: queste Api sono create appositamente per eseguire query, cioé delle ricerche all' interno del contenuto di Twitter. Servono per trovare tweet con specifiche keyword, tweets di uno specifico utente o riferiti all' utente stesso. Oltre a definire le keyword é possibile settare anche alcuni parametri come la lingua e la posizione geografica.

- REST API: permettono l' accesso al core interno di Twitter, sono le piú utili nel caso in cui si ha l'intenzione di sviluppare un'applicazione per il social network. Piú precisamente permettono di interagire con i server inviando richieste sia di tipo Get (per ottenere dati), sia di tipo Post (per inviare dati); esse vengono, ad esempio, utilizzate per creare i cosiddetti bot, programmi in grado di svolgere operazioni automatiche (es. retwittare tutti i tweet contenenti determinate parole chiave). Di interesse per la ricerca é che tramite la REST é possibile accedere ai dati relativi agli utenti, quali numero ed elenco di follower, ultimi tweet inviati, informazioni riportate nei profili, ecc. Sono molto numerose e non sempre molto semplici da utilizzare, proprio per questo sono fornite guide dettagliate per l'utilizzo.

- Streaming API: queste Api permettono lo streaming in realtime di Twitter. Sono utili nel caso venga richiesto un intensivo utilizzo dei dati. Consentono di stabilire una connessione permanente con i server, in grado di raccogliere il flusso di tweet in tempo reale fintanto che la connessione é mantenuta; la richiesta puó specificare, ovviamente, determinate parole chiave, ma anche utenti, lingua, posizione; é inoltre possibile raccogliere un campione rappresentativo del flusso, non filtrato per alcun parametro. Anche queste non sono di facile utilizzo e viene raccomandato l'utilizzo delle guide.

Purtroppo, sia per questioni prettamente tecniche (collegate alla sostenibilitá ingegneristica del flusso di dati), sia per questioni di strategie di business, esistono diverse limitazioni nella raccolta di dati Twitter; limitazioni di tipo diverso, variabili nel tempo che sono riassumibili grossolanamente in:

- Limitazioni di accesso. Twitter possiede tutti i tweet pubblicati all'interno di un database ma non si puo' avervi accesso. Questo tesoro di inestimabile valore é custodito dalla Library of Congress americana. Inoltre fino a Marzo 2011 esistevano diverse iniziative volte a raccogliere archivi tematici di tweet, da rendere disponibili per il download, mentre gli odierni termini del servizio non consentono piú di mettere a disposizione di terzi i suddetti archivi.

- Limitazioni di richieste. Non é possibile inviare quante richieste vogliamo. Questa limitazione riguarda in particolare i metodi che si basano sulle REST API e rappresenta un grosso inconveniente per quanto riguarda studi di una certa dimensione: é infatti possibile inviare un massimo di 350 richieste all'ora (es. si possono raccogliere liste di follower per un massimo di 350 utenti all'ora). Tuttavia la violazione di questi rate limits é rilevata sulla base dell'indirizzo IP da cui si effettua la richiesta, quindi é possibile aggirare la limitazione utilizzando software di mascheramento dell'indirizzo IP. La SEARCH API, ufficialmente, sarebbe anch'essa soggetta a

dei rate limits, i cui parametri non sono tuttavia resi noti; in realtá alcuni utenti riportano che un suo impiego continuativo ad elevata frequenza non ha creato alcun problema, mettendo in dubbio l'esistenza di questi limiti. La STREAMING API, invece, consistendo in una connessione permanente e non in un certo numero di richieste, non é soggetta a nessun limite di questo tipo.

- Limiti di volume. Non é possibile nemmeno raccogliere quanti tweet vogliamo. Questa limitazione chiama in causa soprattutto le raccolte dati attraverso il metodo SEARCH, in quanto essa ha un dominio di applicazione estremamente limitato: é in grado di raccogliere solo i 1500 tweet piú recenti, risalendo all'indietro nel tempo per non piú di 7 giorni circa; inoltre, i risultati di questo metodo tendono ad escludere una piccola porzione di tweet, privilegiando, entro la popolazione che riesce a raccogliere, quelli emessi da utenti piú popolari o escludendo quelli il cui contenuto risulta di bassa qualitá (es. composti da una sola parola). Nemmeno il metodo STREAMING é esente da problemi di esaustivitá della raccolta: qualora si vogliano archiviare flussi molto consistenti, é possibile che parte dei tweet vada persa ma Twitter stesso invita a non preoccuparsi, garantendo che quanto concede é sufficiente per i possibili scopi a cui possono servire i suoi dati.

Output API

Utilizzare le API consiste nell'effettuare delle chiamate a queste ed esse restituiscono i dati con formati che possono essere di tipi diversi, a seconda della chiamata e del dato da restituire. Questi tipi possono essere:

- XML: acronimo di di eXtensible Markup Language, é un linguaggio di markup, ovvero un linguaggio marcatore basato su un meccanismo sintattico che consente di definire e controllare il significato degli elementi contenuti in un documento o in un testo. Costituisce il tentativo di produrre una versione semplificata di Standard Generalized Markup Language (SGML) che consenta di definire in modo semplice nuovi linguaggi di markup da usare in ambito web. Rispetto all'HTML, l'XML ha uno scopo ben diverso: mentre il primo definisce una grammatica per la descrizione e la formattazione di pagine web (layout) e, in generale, di ipertesti, il secondo é un metalinguaggio utilizzato per creare nuovi linguaggi, atti a descrivere documenti strutturati. Mentre l'HTML ha un insieme ben definito e ristretto di tag, con l'XML é invece possibile definirne di propri a seconda delle esigenze.

- JSON: acronimo di JavaScript Object Notation, é un formato adatto per lo scambio dei dati in applicazioni client-server. é basato sul linguaggio

JavaScript Standard ECMA-262 3Âª edizione dicembre 1999, ma ne é indipendente. Viene usato spesso in AJAX come alternativa a XML/XSLT a causa della sua semplicitá. JSON non é un linguaggio di maratura ma un formato di interscambio di dati.

- RSS: acronimo di RDF Site Summary, spesso riportato come Really Simple Syndication. é uno dei piú popolari formati per la distribuzione di contenuti Web; é basato su XML, da cui ha ereditato la semplicitá, l'estensibilitá e la flessibilitá. L'applicazione principale per cui é noto sono i flussi RSS che permettono di essere aggiornati su nuovi articoli o commenti pubblicati nei siti di interesse senza doverli visitare manualmente uno a uno. RSS definisce una struttura adatta a contenere un insieme di notizie, ciascuna delle quali sará composta da vari campi (nome autore, titolo, testo...). Quando si pubblicano delle notizie in formato RSS, la struttura viene aggiornata con i nuovi dati; visto che il formato é predefinito, un qualunque lettore RSS potrá presentare in una maniera omogenea notizie provenienti dalle fonti piú diverse.

- ATOM: Con il termine Atom ci si riferisce a due standard distinti. Atom Syndication Format é un formato di documento basato su XML per la sottoscrizione di contenuti web, come blog o testate giornalistiche. Atom é basato sull'esperienza delle varie versioni del protocollo lanciato da Netscape, RSS. In passato atom é stato conosciuto (anche se solo per un breve periodo) come Pie e poi come Echo.

Search API

In questa sezione vediamo in dettaglio cosa sono, quali risultati forniscono e come si usano le Search API, questo perché sono la categoria usata nel nostro progetto. Come giá detto le Search API servono per effettuare delle ricerche sui tweet pubblicati dagli utenti, peró queste API possiedono delle limitazioni:

- L'elenco dei tweet restituito non é completo ma é costituito da tweet recenti, piú precisamente non sono piú vecchi di 6-9 giorni.

- La query di ricerca é soggetta a limitazioni di complessitá. Nel caso in cui la query sia ritenuta troppo complessa, verrá restituito il messaggio d'errore: error : Sorry, your query is too complex. Please reduce complexity and try again.

- La ricerca non supporta l'autenticazione, cioé tutte le query sono effettuate in forma anonima

- La ricerca é focalizzata in rilevanza e non completezza. Ció significa che alcuni tweet e gli utenti possono essere mancanti dai risultati di ricerca

- Non puo' essere utilizzato l'operatore NEAR. In alternativa vi é il parametro geocode

- Le query sono limitate a 1000 caratteri di lunghezza, compresi gli eventuali operatori.

- Esiste un limite al rate di richieste che,diversamente dalle REST API, non é definito in modo fisso. Infatti il limite non é dato da un certo numero di richieste all'ora ma dipende dalla complessitá e dalla frequenza delle richieste. Per questo il rate puo' essere limitato e in questo caso la Search API risponderá con: HTTP 420 error: error: You have been rate limited. Enhance your calm.

Per effettuare una richiesta alla Search API si utilizza questo indirizzo:

http://search.twitter.com/search.format

dove format mi indica il formato della richiesta. Seguito da una serie di parametri opzionali che permettono di affinare la ricerca. Questi parametri sono:

- q: é l'unico parametro non opzionale che rappresenta la query di ricerca. In questa query é necessario rispettare i limiti precedentemente elencati. Se non contiene nessuna parola chiave si riceverá un errore HTTP 403 con il messaggio error : You must enter a query. .

- callback: disponibile solo per il formato JSON. Se in dotazione, per la risposta verrá utilizzato il formato JSONP con un callback del nome dato.

- geocode: restituisce tweet di utenti che si trovano all'interno di un determinato raggio di latitudine / longitudine data. La posizione é preferenzialmente presa dalle API Geotagging, ma spesso ci si riduce alla posizione specificata dal profilo Twitter. Il valore del parametro viene specificato da latitudine, longitudine, raggio , in cui l'unitá di misura del raggio deve essere specificata come mi (miglia) o km (chilometri). Si noti che non é possibile utilizzare l'operatore di prossimitá. Un massimo di 1.000 distinte sotto regioni saranno prese in considerazione quando si utilizza il raggio.

- lang: limita tweets al linguaggio dato, dato da un codice ISO 639-1.

- locale: usato per specificare il linguaggio della query che si sta inviando (solo ja é attualmente attivo). Questo é inteso per specifici linguaggi client e di default dovrebbe funzionare nella maggior parte dei casi.

- rpp: il numero di tweet da restituire per ogni pagina, fino ad un massimo di 100.

- page: il numero di pagine (a partire da 1) in cui restituire i tweet, fino ad un massimo di circa 1500 risultati (page * rpp non deve superare i 1500).
- result-type: Specifica il tipo di risultati di ricerca che si preferisce ricevere. Il valore predefinito corrente é mixed. I valori validi sono:
  - mixed: Include sia i risultati in tempo reale sia quelli popolari nella risposta.
  - recent: restituisce solo i risultati piú recenti nella risposta
  - popular: restituisce solo i risultati piú popolari nella risposta.
- show-user: Se true , antepone : all'inizio del tweet. Questo é utile per i lettori che non visualizzano i campi autore Atom. Il valore predefinito é false.
- until: Restituisce tweet generati prima della data stabilita. La data deve essere formattata come AAAA-MM-GG. Non é possibile settare una data futura altrimenti in questo caso verrá restituito un HTTP 403 error con il messaggio : error:You cannot use an 'until:' date in the future
- since-id: Restituisce i risultati con un ID superiore (cioé, piú recente), all'ID specificato. Ci sono dei limiti al numero di tweet a cui si accede tramite l'API. Se il limite di Tweets é verificato dopo il since-id, allora il since-id sará forzato al piú vecchio ID disponibile. Invece se il since-id ha un valore futuro restituirá un HTTP 403 error con il messaggio: error:since-id too recent, poll less frequently.
- max-id: restituisce i risultati con un ID inferiore (cioé, piú vecchio) o uguale all'ID specificato.
- include-entities: Quando impostata su true , t o 1 , ogni tweet includerá un nodo denominato entitá. Questo nodo offre una varietá di metadati relativi ai tweet in una struttura discreta, tra cui: URL, media e hashtag. Le entitá sono disponibili solo per le risposte col formato JSON nelle Search API.

Ad esempio una richiesta come: *http://search.twitter.com/search.json?q=twitterapi & rpp=1* restituisce un solo tweet, perché rpp=1, che soddisfa la ricerca della parola chiave twitterapi. Il campo q che rappresenta la query, come accennato precedentemente, oltre che alle parole chiave puo' contenere degli operatori. Di seguito, attraverso degli esempi, vediamo quali sono questi operatori e che risultati ci forniscono. Se il campo q é uguale:

- Twitter Search, ci verranno restituiti tutti i tweet che contengono sia la parola twitter sia la parola search. Questo é l'operatore predefinito
- Happy hour, ci verranno restituiti i tweet che contengono esattamente la frase Happy hour

- Love OR hate, avremo in risposta i tweet che contengono o la parola Love o la parola hate o entrambi
- Beer -root, restituirá i tweet con la parola beer e senza la parola root
- # Haiku, restituirá i tweet con l'hashtag Haiku
- from:twitterapi, restituirá i tweet inviati dall'utente @twitterapi
- to:twitterapi, avremo in risposta i tweet inviati all'utente @twitterapi
- place:opentable:2, avremo i tweet con con il posto ID opentable 2
- luogo: 247f43d441defc03, avremo i tweet con circa il posto con id 247f43d441defc03
- @twitterapi, ci verranno restituiti i tweet in cui é menzionato l'utente twitterapi
- superhero since:2011-05-09, avremo i tweet con la parola superhero inviati dal 09/05/2011 in poi
- twitterapi until:2011-05-09, avremo i tweet con la parola twitterapi inviati prima del 09/05/2011
- movie -scary :), avremo i tweet contenenti film, ma non paura, e con un'attitudine positiva
- flight :(, tweet conteneti la parola flight ma con un'attitudine negativa
- traffic ?, tweet contenenti traffic e che pongono una domanda
- hilarious filter:links, avremo l'elenco dei tweet con la parola hilarious e contenenti un URL
- news source:tweet-button, tweet contenenti news e inserito attraverso il Twitter Button

4.3.2 CRAWLER DEI TWEET INERENTI AI FILM

In questo progetto, dopo essere riusciti a ricavare i dati di vendita e le uscite in anteprima dei film, abbiamo la necessitá di ricavare dalla rete i tweet inerenti a questi film per poi poterne analizzare il sentiment. Per realizzare ció é stato realizzato un crawler che attraverso le Search API di Twitter effettua delle query ed estrae i tweet utili per i nostri scopi. Questo crawler consiste in un applicazione java che sfrutta le potenzialitá offerte dalla libreria twitter4j, una libreria Java non ufficiale per l' API di Twitter. Con questa libreria si puo' facilmente integrare l'applicazione Java con i servizi che mette a disposizione Twitter attraverso le proprie API. In questo crawler andremo a costruire

inizialmente la query nella forma presentata nel capitolo precedente e poi imposteremo questa query come argomento della Search API ed estrarremo i tweet a noi utili. Peró sorge una problematica data dalla complessitá e lunghezza della query, infatti come é stato detto, la Search API pone dei vincoli di lunghezza e complessitá sulla query. Per poter superare questo problema é stato necessario spezzare in 4 parti la query per i tweet italiani e in 5 parti quella per i tweet americani. Quindi, come verrá mostrato piú avanti nel codice, vedremo 4 stringhe chiamate SEARCHKEY1, SEARCHKEY2, SEARCHKEYATTORI1, SEARCHKEYATTORI2 e quindi 4 interrogazioni differenti alla API di Twitter. Questo suddivisione in 4 parti, porta nel complesso a moltiplicare le chiamate all'API fino al raggiungimento del limite e a stoppare e richieste. In soluzione a quest'ultimo problema, alla fine di tutte le richieste effettuate per un film, il processo viene messo a riposo per due secondi, in modo tale da distanziare tra loro le richieste e non superare il limite. Implementazione Crawler dei tweet La nostra applicazione come prima cosa, chiede in input il mese a cui appartiene l'uscita dei film di cui si vogliono scaricare i tweet. Questo significa che il nostro programma effettuerá la ricerca e la memorizzazione di tutti i tweet dei film di un certo mese indipendentemente dal giorno in cui essi usciranno perché, ai fini della nostra ricerca, risultano importanti anche i tweet precedenti all'uscita nelle sale. Successivamente il programma effettua delle interrogazioni al database contenente i dati relativi ai film immagazzinati con i crawler mostrati nelle precedenti sezioni e struttura 4 stringhe che comporranno il testo delle query da inviare all'API di Twitter. Una volta ottenuti i testi delle query, per ognuna viene proposto il seguente codice che, sfruttando le funzionalitá che la libreria twitter4j mette a disposizione, effettuano la query sull'API: Nel codice notiamo come la libreria consenta attraverso l'oggetto Query di poter impostare oltre al testo della query e quindi al parametro q della richiesta, anche tutti gli altri parametri. Nel nostro caso i parametri settati sono la lingua, posta ad italiano (it) con il metodo setLang, il tipo di tweet, posto a recenti attraverso il metodo setResultType e il numero di tweet da restituire per pagina, posto a 100 con setRpp. Con il metodo search() sull'istanza dell'oggetto Twitter andiamo a interrogare il database di Twitter dalla cui risposta riusciamo ad avere una lista di tweet dal quale prendiamo 2 parametri, il testo, con il metodo .text() e il timestap di pubblicazione, con il metodo .getCreatedAt(). Tutte queste informazioni, assieme al riferimento del titolo del film, vengono salvate all'interno del database assicurandoci di non averlo giá memorizzato. Per quanto riguarda il crawler per i film americani, l'unica differenza consta nel parametro Lang che viene posto uguale a en, cioé in lingua inglese.

Algorithm 4.2 Algoritmo utilizzato per estrarre il titolo del film e il link alla pagina di dettaglio del film

```
1: Twitter twitter = new TwitterFactory().getInstance();
2: Query search1 = new Query(searchkeyattori1);
3: search1.setLang("it");
4: search1.setResultType("recent");
5: search1.setRpp(100);
6: QueryResult result1= twitter.search(search1);
7: List < Tweet > tweets1 = result1.getTweets();
8: String testo1;
9: String tito;
10: for (Tweet tweet : tweets1)
11: testo1 = tweet.getText();
12: tito=" SELECT * FROM tweet WHERE testo=' " + testo1 +" ' AND
    data=' "+ tweet.getCreatedAt() + " ' AND titolo=' "+ titoli[t] + " ';";
13: ResultSet rs1 = stmt.executeQuery(tito);
14: if(!rs1.next())
15:
16: int val = stmt.executeUpdate("INSERT INTO tweet(testo, data, titolo) VA-
    LUES (' + testo1 + ','+ tweet.getCreatedAt() + " ',' " + titoli[t] + " '); "
    );
17:
18:
```

4.4 ESTRAZIONE DEL SENTIMENT

Come ultimo passo prima di effettuare le analisi sui dati e vedere se le nostre ipotesi sono fondate, andiamo ad arricchire i nostri dati aggiungendo le informazioni relative al sentiment estraibile dall'insieme di tweet relative ai lungometraggi ricavati dal crawler precedentemente presentato. Questa parte di estrazione del sentiment verrá applicata solo ai tweet italiani mentre non sará calcolato il sentiment per quelli in lingua inglese su cui si fará solo una valutazione di correlazione sui volumi. Per la realizzazione di questa parte é stato utilizzato uno strumento realizzato da Caludio Carcaci, studente del Politecnico di Milano, che come progetto di tesi magistrale ha realizzato questo tool per l'estrazione del sentiment facendo un'analisi di tipo sintattico.

4.4.1 TOOL ESTRAZIONE SENTIMENT DI CLAUDIO CARCACI

Il tool realizzato da Claudio Carcaci ha come obiettivo quello di fornire un'analisi guidata totalmente dalla sintassi che permetta di classificare messaggi, post, tweet provenienti dai social network. Per fare ció non utilizza algoritmi giá presenti in letteratura ma ha progettato un algoritmo che possa avere prestazioni ancora migliori degli algoritmi classici. Di seguito presentiamo in modo molto

sintetico in cosa consiste questo algoritmo e come dal punto di vista pratico si presenta questo tool nell'utilizzo.

Algoritmo ExhaustiveSets+meta

L'algoritmo si basa per prima cosa sulla definizione di tre entitá: testi, termini e classi, ma soprattutto sulle relazioni tra queste entitá:

- Relazione tra testi e termini: rappresentata da una matrice di rilevanza i cui valori rappresentano l'importanza di un termine all'interno del testo. Il valore si basa sul numero di comparse di un termine nei testi del nostro dataset e sul numero di comparse nel singolo testo.
- Relazione tra testi e classi: é la ricostruzione che l'algoritmo di classificazione si propone di fare. A livello teorico la relazione é di tipo funzionale, quindi ad ogni testo viene associata al massimo una classe e possibilmente non vi siano classi a cui non sono assegnati testi.
- Relazione tra termini e classi: questa relazione é definibile in due modi, tramite metadata forniti preventivamente circa i termini che identificano una classe oppure dopo la classificazione tramite i termini che appartengono a testi classificati.

Queste tre relazioni sono rappresentate da tre matrici, la prima dalla matrice di rilevanza, la seconda da una matrice di attinenza, che mi dice tramite dei valori per quel testo qual é la classe che é piú rappresentativa, e la terza dalla matrice dei pesi quindi l'importanza di quel termine per quella classe. Quindi partendo da una situazione iniziale data dalla matrice di rilevanza e da un file contenente dei metadati che permettono di definire le classi e i termini che appartengono alla classe, l'algoritmo, che é di tipo greedy, effettua in sintesi i seguenti passaggi:

- inizializzando la matrice dei pesi associando il valore iniziale a tutti i termini afferenti ad una data classe passati come metadata
- per ogni classe estrae l'elenco dei documenti che contengono almeno un termine associato ad essa e aggiorna la matrice di attinenza secondo la rilevanza di un termine all'interno della classe
- effettua l'aggiornamento della matrice dei pesi secondo tre modalitá:
  - Un termine rilevante per una classe é contenuto nel testo, questo comporta un rinforzo positivo
  - Un termine é contenuto nel testo ma non é indicato come rilevante per la classe, esso comporta un rinforzo positivo perché il termine puó essere utile per la classe ma non é specificato nei metadata relativi ad essa

  - Un termine specificato come rilevante non é contenuto nel testo, questo comporta un rinforzo negativo poiché é facile che se non compare in molti testi dove sono contenuti gli altri termini avrá una bassa rilevanza per quella classe e dovrá essere escluso.

- In base al superamento di alcuni valore soglia di inclusione, si includono nuovi termini e i relativi testi o si escludono i termini e i testi che contengono solo i termini esclusi che quindi hanno una rilevanza per la classe inferiore alla soglia di esclusione

Dopo aver eseguito l'algoritmo si ottengono come risultati le associazioni tra testi e classi. Si ottiene quindi un insieme di valori di attinenza tra queste entitá che é possibile esprimere in una matrice dove sulle righe vi sono i testi e sulle colonne le classi. Infine si effettua una scrematura parziale delle classi in modo da eliminare i testi che hanno una attinenza bassa con la classe di riferimento attraverso la definizione di una soglia al di sotto della quale l'associazione classe/testo non viene considerata. Definizione dei metadati Per il funzionamento di questa applicazione risulta fondamentale la definizione dei metadati che permettono di definire due questioni fondamentali:

- Le classi
- I termini associati ad ogni singola classe

In concreto questi metadati consistono in un semplice file CSV (Comma-Separated Values) di nome *classesterms.csv* avente come struttura:

Nomeclasse1;Termine1, Termine2, Termine3, ..., TermineN;
Nomeclasse2;Termine1, Termine2, Termine3, ..., TermineN;
...
NomeclasseN;Termine1, Termine2, Termine3, ..., TermineN;

Nel nostro progetto le classi da individuare sono semplicemente tre:

- Positivi: rappresenta la classe di termini che mi permette di identificare i tweet che danno un parere positivo sul film in questione.
- Negativi: rappresenta la classe di termini che mi permette di identificare i tweet che danno un parere negativo sul film in questione.
- Neutri: rappresenta la classe di termini che mi permette di identificare i tweet che non danno né parere positivo né parere negativo ma che sono attinenti al film

I tweet che verranno catalogati in Positivi o Negativi oltre che esprimere esplicitamente un parere sul film, possono essere anche messaggi di apprezzamento

o di disprezzo relativi al regista o agli attori presenti nel film che comunque vanno ad inficiare sul giudizio generale del film. Infatti molte persone decidono o meno di andare a vedere un film in base al regista o agli attori che fanno parte del cast. La categoria Neutri é utile soprattutto per effettuare una sorta di filtro ai tweet estratti dalla rete attraverso la nostra query. Infatti, come giá presentato nel capitolo precedente, possono essere stati estratti tweet che non hanno nulla a che vedere col settore cinematografico che potrebbero alterare i risultati della nostra analisi. Quindi, attraverso proprio questa classe, andiamo ad individuare solo i tweet che sono attinenti ai film e al cinema. Ogni tweet puo' essere contenuto in piú classi, anche ad esempio, in Positivi e in Negativi perché il singolo tweet puo' contenere porzioni di testo che danno un parere positivo al film ed altri un parere negativo, come ad esempio: Flight é noioso peró Danzel Washington é super!!! Per identificare i termini da associare alle classi, é stato ispezionato manualmente, un campione di piú di 3000 tweet presi casualmente a gruppi da quelli estratti per i vari film. Di questi tweet, sempre manualmente, si é cercato di capire le parole chiave per identificare il sentiment dei tweet e una volta individuati, sono stati inseriti nel file classesterms.csv nella corrispettiva classe. Di seguito verrá presentata una parte dei termini che sono stati inclusi nelle singole classi per far capire concretamente quali termini sono stati considerati importanti ai fini dell'estrazione del sentiment. Per la classe Positivi sono stati individuati i seguenti termini:

gran film, idolo, andate a vedere, andatelo a vedere, adoro, capolavoro, grande, il piú grande, piaciuto, migliore, miglior, miglior film, merita, spacca, magnifico, grandiosamente, bello, belli, troppo forte, figo, magistrale, stupendo, grandioso, miglior, molto bene, mi piace, consigliato, gran, toccante, spettacolare, spassoso, notevole, notevoli, spettacolo, bellissimo, bravissimo, da premiare, una bomba, grandissimo, bel film, preferito, preferiti, da vedere, magistralmente, meritatissimo, meritatissima, divertimento, incredibile, favoloso, splendido, immenso, imperdibile, bravura, perfezione, ottimo, top, spassoso, fenomenale, bella, bei, amo, entusiasmante, ben fatto, carino, piú belli, tanta roba, bomba, consiglio, meravigliosi, commuoventi, meraviglioso, ecc.

Per la classe Negativi sono stati individuati i seguenti termini:

scarso, scarsi, molto scarsi, sconsiglio, pessimo, lento, angoscia, ansia, tristezza, mezzo film, pacchiano, vergogna, deluso, brutto, brutti, noioso, noiosi, fatto male, male, punti morti, moralista, fasullo, niente di che, piaciuto meno, non é piaciuto, non mi é piaciuto, deludente, troppo serio, palloso, crollo, crolla, di meglio, non mi ha entusiasmato, tremendo, aspettavo di piú, delude, banale, peggio, sbadiglio, sonno, bruttissimo, retorico, scontato, prolisso, addormentarsi, delusione, vomitare, evitate, evitare, inconsistente, operetta, patetico, pate-

tiche, pesante, noiosissima, americanata, americanate, stufato, pretenzioso, non dei migliori, non andate, addormentata, addormentato, dormito, dormita, non eccezionale, mediocre, peggiore, ecc.

Per la classe Neutri sono stati individuati i seguenti termini:

trailer, programmazione, proiezione, cinema, film, girare, girato, recensione, sala, scena, scene, colonna sonora, storia, attori, regia, sceneggiatura, schermo, attrice, attore, attori, incassi, botteghino, pellicola, vedere, visto, vedró ecc.

Utilizzo operativo del tool

Qui vedremo ora come si utilizza il tool e come si presenta a livello d'interfaccia nella funzionalitá da noi utilizzata in questo progetto, questo perché il tool possiede altre funzionalitá che peró non verranno presentate. All'avvio dell'applicazione abbiamo un'interfaccia che si presenta come in Figura 4.8: Attraverso

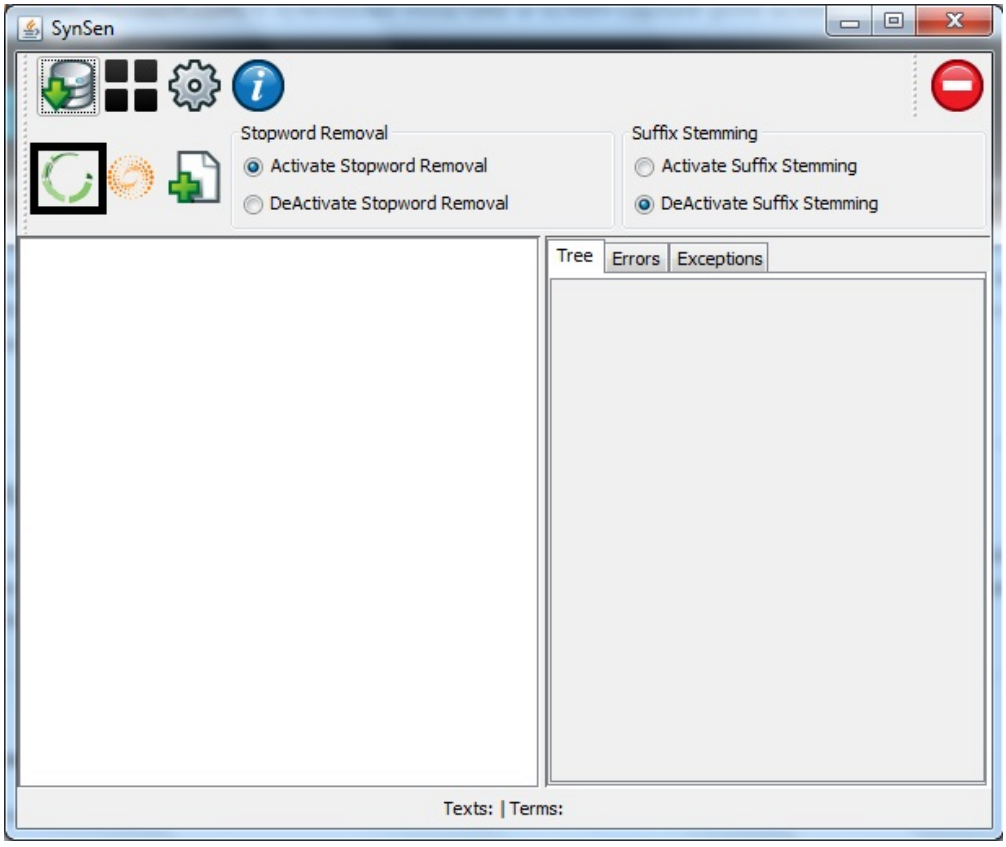


Figura 4.8: Interfaccia iniziale tool Carcaci

il pulsante che si puo' vedere in figura evidenziato dal riquadro nero, andiamo a selezionare la cartella in cui é necessario che siano contenuti due file:

- classesterms.csv: file che contiene i metadati contenenti le informazioni relative alle classi ed ai termini che appartengono alle classi. Questo file deve rispettare il formato definito nella precedente sezione
- textsclasses.csv: file contenente su ogni riga il testo di un tweet da analizzare. Non sono ammessi tweet scritti su piú righe.

Una volta selezionata la cartella, nella parte destra compariranno i testi dei tweet contenuti nel file textsclasses.csv, mentre a destra, alla selezione di un testo, comparirá l'albero sintattico della frase in questione, ma questa é una funzionalitá non utile al nostro progetto (4.9). Cliccando sul pulsante eviden-

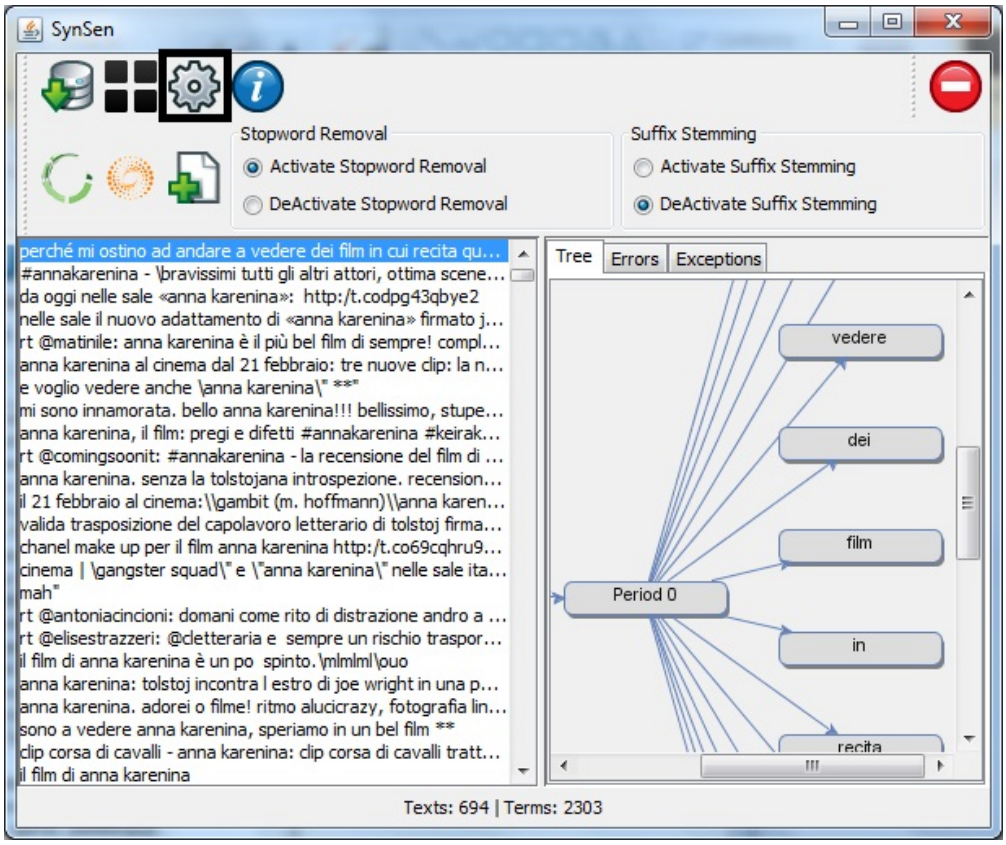


Figura 4.9: Interfaccia dopo il caricamento dei testi del tool Carcaci

ziato nella precedente figura andiamo a scegliere quale algoritmo utilizzare per effettuare l'associazione dei testi alle varie classi. Ovviamente noi andiamo a specificare l'algoritmo *ExhaustiveSets+meta*, che permette di sfruttare il nostro file contenente le classi, e avviamo l'esecuzione. Nel corpo del form compariranno le classi, il numero di tweet associato ad ogni classe, e l'elenco dei tweet, non completo, per ogni classe, come in Figura 4.10

In figura abbiamo analizzato ad esempio, un campione di tweet relative alla settimana compresa tra il 18 e il 24 Febbraio sul film Anna Karenina. Come si puo' notare dei 713 tweet dati al tool, esso ha classificato ben 615 di questi

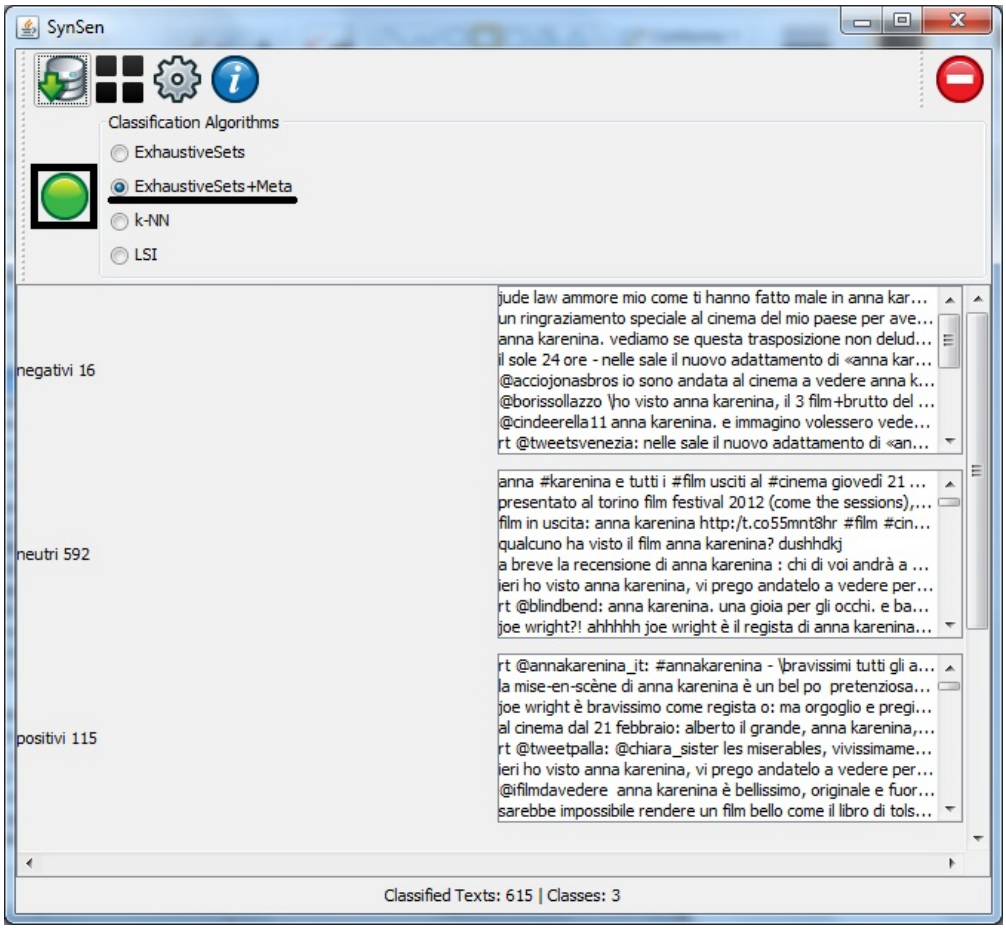


Figura 4.10: Interfaccia che mostra i risutati del sentiment del tool Carcaci

dividendoli in: 115 Positivi, 16 Negativi, 592 Neutri. Calcolando il sentiment sui tweet di tutti i film presi in esame notiamo la poca percentuale di tweet positivi o negativi rispetto al totale, questo perché spesso questi rappresentano semplice pubblicitá al film, informazioni sulla programmazione dei cinema o link a recensioni e interviste dei protagonisti del film. In piú parecchi utenti scrivono la loro intenzione di andare a vedere il film o informano di aver visto il film senza peró dare un parere su di esso. Da notare anche la bassissima percentuale dei tweet negativi, sia rispetto al totale, sia rispetto ai positivi, questo potrebbe indicare semplicemente che coloro che assistono ad uno spettacolo cinematografico ne escono con un parere positivo perché avevano effettuato una scelta della pellicola consona con i propri gusti.

4.5 STATA 9

In questa sezione andiamo a presentare lo strumento che ci permetterá nel capitolo successivo di verificare o meno se esiste un rapporto di correlazione lineare

tra le variabili da noi individuate. Questo strumento é Stata, un software statistico che permette di organizzare ed elaborare dati, di produrre statistiche e grafici e di stimare una grande varietá di modelli econometrici. E' disponibile sul mercato in varie versioni ma quella utilizzata in questo progetto é la versione 9. Essa non é l'ultima versione disponibile ma le differenze fra queste versioni sono molto ridotte e limitate soprattutto a funzionalitá complesse. All'avvio stata presenta 4 finestre principali:

- Command: sezione in cui i comandi vengono inseriti tramite tastiera e mandati in esecuzione premendo il tasto Invio
- Review: finestra in cui sono mostrati tutti i comandi memorizzati in un buffer di memoria e dal quale posso essere richiamati per essere rieseguiti.
- Results: mostra le informazioni sulla sessione di lavoro, i risultati dei comandi, compresa la segnalazione di eventuali errori.
- Variables: mostra la lista delle variabili contenute nel dataset attivo.

Con poche eccezioni, la sintassi base del linguaggio di STATA é:

[by *varlist*:] *command* [*varlist*][= *exp*][if *exp*][in *range*][*weight*][, *options*]

dove le parentesi quadre denotano componenti opzionali. In questo schema, *varlist* indica una lista di nomi di variabili, *command* indica un comando, *exp* indica un'espressione algebrica, *range* indica un intervallo di osservazioni, *weight* indica un'espressione per attribuire un peso alle osservazioni e *options* indica una lista di opzioni.

varlist : la maggior parte dei comandi che accettano una lista di variabili non necessitano che questa sia esplicitamente indicata. Se non appare alcuna lista, tali comandi assumono che la lista sia *all*, che é un'abbreviazione per indicare tutte le variabili del dataset caricato. Per i comandi che alterano o distruggono dati, STATA richiede che la lista di variabili sia esplicitamente indicata

by varlist : questo prefisso fa in modo che STATA ripeta il comando specificato per ogni sottoinsieme individuato dalle modalitá della/e variabile/i indicata/e. I dati devono essere giá ordinati secondo tale/i variabile/i

if exp : si restringe l'esecuzione di un comando a quelle osservazioni per le quali il valore dell'espressione specificata é vera

in range : si restringe l'esecuzione del comando ad uno specifico intervallo di osservazioni. La specificazione dell'intervallo ha la seguente forma: # 1[# 2],

dove # 1 e # 2 sono numeri che indicano la prima e l'ultima osservazione coinvolte Stata mette a disposizione moltissimi comandi, di seguito saranno mostrati solo quelli utilizzati per effettuare le nostre analisi:

- Insheet: comando utile per acquisire dati dall'esterno, piú precisamente per fogli elettronici salvati come .csv o tab-delimited da programmi come excel. Nel nostro caso utilizziamo anche l'opzione delimiter che permette di specificare il carattere che delimita le colonne di dati, per noi il ;. Per esteso il comando utilizzato é:

  insheet using percorsofile, delimiter(;)

- Tsset: comando che permette di settare la variabile che scandisce il tempo.

  tsset var1

- Regress: Il comando che permette di ottenere un'analisi di regressione lineare. La sua sintassi base é:

  regress vardipendente listadivarindipendenti, [opzioni] Tra le opzioni ci sono: la possibilitá di sopprimere il termine costante (nocostant), di ottenere i coefficienti standardizzati (beta), di cambiare il livello degli intervalli di confidenza (level(#)) e utilizzare il costrutto if per poter filtrare i dati. In output questo comando restituisce una serie di valori che vengono calcolati che sono:

  - RSS: somma dei quadrati spiegati dalla regressione
  - ESS: somma degli errori al quadrato
  - TSS: somma totale dei quadrati
  - Gradi di libertá usati
  - Valore dell'F-test
  - R-squared = RSS/TSS, coefficiente di determinazione, é la percentuale di variabilitá in un insieme di dati che viene valutata con il modello statistico
  - Adj R-squared, é l'R-squared corretto
  - Root MSE, radice dell'errore quadratico medio, indica la discrepanza quadratica media fra i valori dei dati osservati ed i valori dei dati stimati
  - Intevallo di confidenza al 95

  Nel nostro contesto, fra tutti questi output analizzeremo R-squared, il cui valore é compreso tra 0 e 1 e se esso tende ad 1 significa che l'andamento lineare si sposa bene coi dati, se tende a 0 indica che la retta di regressione non si adatta molto bene ai dati.

- twoway (scatter var1 var2) (lfit var1 var2): comando che permette di tracciare un grafico avente l'andamento a dispersione dei valori delle variabili specificate e la retta che rappresenta l'andamento lineare perfetto per quei dati.

Capitolo 5

ANALISI E RISULTATI

In questo capitolo andiamo ad utilizzare tutti i dati raccolti per effettuare delle analisi che ci permettono di verificare o meno la nostra ipotesi, ossia la presenza di correlazione lineare tra le vendite e i volumi o tra le vendite e il sentiment web estraibile da Twitter. Queste analisi verranno effettuae attraverso uno strumento e software statistico per analisi dei dati per i professionisti, Stata 9. Per regressione lineare tra due o piú variabili s'intende quella relazione in cui si puo' identificare una variabile indipendente e una o piú variabili dipendenti e la relazione tra queste é esprimibile attraverso un'equazione lineare, o retta della forma:

$Y = \alpha + \beta X$

Dove,α detta *intercetta*, é il valore dell'equazione quando interseca l'asse y, mentre β detto *coefficiente di regressione o pendenza della retta*, e indica di quanto aumenta la y per l'aumento di una unitá delle x.

5.1 VERIFICA CORRELAZIONE PER SINGOLO FILM SU DATI ITALIANI

Nel nostro database abbiamo tutti i dati che sono stati estratti attraverso i vari crawler e ora é necessario ricavare i dati utili alla nostra analisi. Questi dati sono stati raccolti in modo completo e affidabile in un range temporale che va dal 24 Gennaio fino al 24 Marzo e si é deciso di considerare i film che sono usciti nelle sale e che hanno completato la loro "vita all'interno di questo intervallo. Come precisato nel capitolo 3 consideriamo un film ancora "vivo se esso é presente nella classifica delle 20 vendite di Mymovies anche perché essere fuori da questa classifica vuol dire aver avuto poco successo al botteghino in quella settimana e quindi avere delle vendite trascurabili. Attraverso delle semplici interrogazioni al database abbiamo estratto i dati di vendita relativi ai film che soddisfano la condizione precedentemente esposta e che rappresentano i dati relativi ai

film usciti in italia, ed estratto i volumi di tweet relativi a questi film, nelle settimane in cui risultano nella classifica di vendita. Tutti questi dati assieme ai valori relativi al sentiment estratti con il tool Carcaci, sono stati riassunti in un unico file csv che sará la nostra sorgente dati da date in input a stata. Vediamo nel dettaglio questo file:

film	num_sett	titolo	data	volume_s	vendite_sett	volume_c	vendite_c	positivi	negativi	neutri
1	7	Amiche da morire	10/03/2013	438	748119	438	748119	77	2	263
1	8	Amiche da morire	17/03/2013	136	640560	574	1388679	36	3	75
1	9	Amiche da morire	24/03/2013	51	315938	625	1704617	9	0	36
2	5	Anna Karenina	24/02/2013	713	825401	713	825401	115	16	471
2	6	Anna Karenina	03/03/2013	565	875324	1278	1700725	142	19	307
2	7	Anna Karenina	10/03/2013	250	405871	1528	2106596	69	7	136
2	8	Anna Karenina	17/03/2013	178	155938	1706	2262534	32	4	107
3	5	Beautiful Creatures	24/02/2013	353	419303	353	419303	51	18	248
4	3	Broken City	10/02/2013	260	604989	260	604989	25	4	158
4	4	Broken City	17/02/2013	130	364228	390	969217	24	5	102
5	8	Buongiorno Papà	17/03/2013	510	1059740	510	1059740	86	7	366
5	9	Buongiorno Papà	24/03/2013	229	831045	739	1890785	72	4	169

Figura 5.1: File di input passato a stata

- Film: rappresenta un valore numerico che é stato assegnato al film utile per filtrare i dati all'interno di stata piú facilmente rispetto allo specificare il titolo del film
- Numsett: rappresenta un numero che va da 1 a 9 e mi indica la settimana a cui si riferiscono i dati. Alla prima settimana in esame, dal 21 al 27 gennaio, é associato il numero 1, e via dicendo fino ad arrivare all'ultima settimana, dal 18 al 24 Marzo, a cui é associato il numero 9
- Titolo: titolo del film in oggetto
- Data: indica la data della domenica della settimana a cui sono riferiti i dati del film
- Volume_s: numero di tweet relativi al film corrente presenti in rete nella settimana specificata
- Vendite_s: cifra in euro che rappresenta le vendite di quel singolo film nella settimana indicata
- Volume_c: volume cumulato, cioé il numero di tweet estratti dalla rete a partire dalla settimana di uscita del film fino alla settimana indicata. Ovviamente nella prima settimana volumes e volumec sono identici
- Positivi: tweet che attraverso l'analisi del sentiment, risultano contenere un parere positivo per film corrente
- Negativi: tweet che attraverso l'analisi del sentiment, risultano contenere un parere negativo per il film corrente

- Neutri: tweet che attraverso l'analisi del sentiment, risultano essere associati alla categoria neutri per il film corrente

In questa nostra tabella risultano esserci i dati relativi a 40 film con 123 record. Peró analizzandola notiamo che alcuni film sono stati presenti in classifica per 1 o 2 settimane e con questi risulta inutile effettuare un'analisi di correlazione per singolo film a causa dell'insufficienza di dati. Tralasciando questi film, per gli altri, di cui abbiamo per ciascuno dai 3 ai 7 record di dati, andiamo a verificare se esiste una correlazione lineare tra:

- Vendite cumulate e volumi cumulati
- Vendite settimanali e volumi settimanali
- Vendite settimanali e sentiment settimanale (positivi e negativi)
- Vendite settimanali e volumi associati alla classe Neutri

5.1.1 Vendite cumulate e volumi cumulati

Applicare la regressione lineare tra le vendite cumulate e i volumi cumulati per singolo film, vuol dire andare a verificare se l'andamento dei nostri dati é prossimo a quello di una retta di equazione:

$vendite\_c = \alpha + \beta volumi\_c$

per verificare questo si é eseguito su stata il comando:

regress vendite_ c volume_ c if(film==N)

dove N mi indica il numero identificativo di un film. Applicando la regressione lineare per tutti i film con questi dati, notiamo che abbiamo dei Rsquared tutti sopra lo 0.94, quindi tendenti fortemente ad 1. Questo comporta un'evidente correlazione lineare fra queste due metriche. Per completezza presentiamo nella Tabella 5.1 tutti i valori dell'R–Squared per i film aventi 3 o piú record.
Mentre in Figura 5.2 vediamo un grafico relativo al film "Il principe abusivo" dove si puo' apprezzare l'andamento lineare dei dati.

5.1.2 Vendite settimanali e volumi settimanali

Applicare la regressione lineare tra le vendite settimanali e i volumi settimanali per singolo film, vuol dire andare a verificare se l'andamento dei nostri dati é prossimo a quello di una retta di equazione:

$vendite\_s = \alpha + \beta volumi\_s$ per verificare questo si é eseguito su stata il comando:

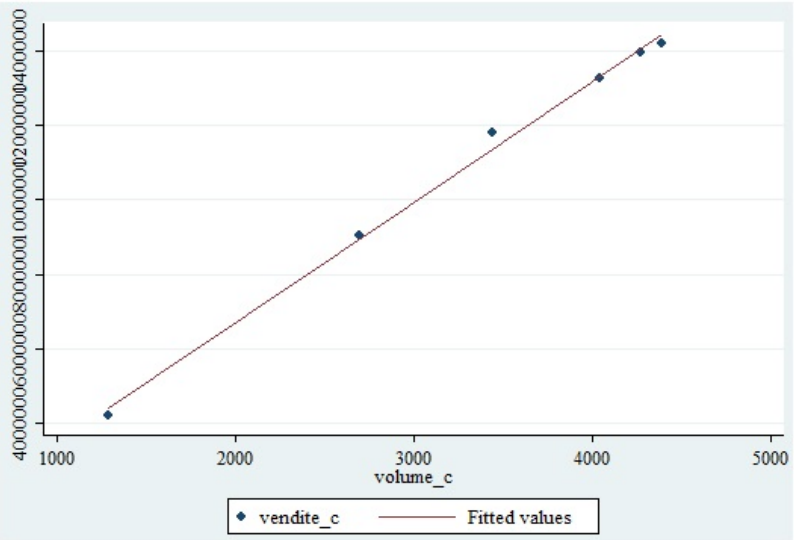


Figura 5.2: Regressione Lineare film "Il principe abusivo"

regress vendite_s volume_s if(film==N)

dove N indica il numero identificativo di un film.

In questo caso, la maggior parte dei valori dello Rsquared sono superiori allo 0.7 che si puo' considerare una buona soglia per poter affermare che esiste una correlazione lineare. Peró é importante segnalare anche che per alcuni film abbiamo addirittura valori tendenti allo 0 e che quindi presentano un andamento totalmente non lineare. In Tabella 5.2 sono riportati tutti gli Rsquared calcolati sui film per la correlazione tra le vendite settimanali e i volumi settimanali:
Come si puo' notare per 8 film su 24 considerati abbiamo un Rsquared inferiore allo 0.7 di cui 4 addirittura molto sotto lo 0.5. Peró per ben 2/3 dei film presi in esame abbiamo una correlazione lineare molto evidente, ad esempio vediamo in Figura 5.3 il grafico relativo al film "Les Misérables" notando l'andamento molto prossimo alla retta lineare:

5.1.3 Vendite settimanali e sentiment settimanale

Applicare la regressione lineare tra le vendite settimanali e il sentiment settimanale, ovvero considerando sia il volume dei tweet con sentiment positivo come fattore additivo sia il volume dei tweet con sentiment negativo come fattore sottrattivo, per singolo film. Questo significa andare a verificare se l'andamento dei nostri dati é prossimo all'equazione:

$vendite\_s = \alpha + \beta sentiment\_pos - \gamma sentiment\_neg$ per verificare questo

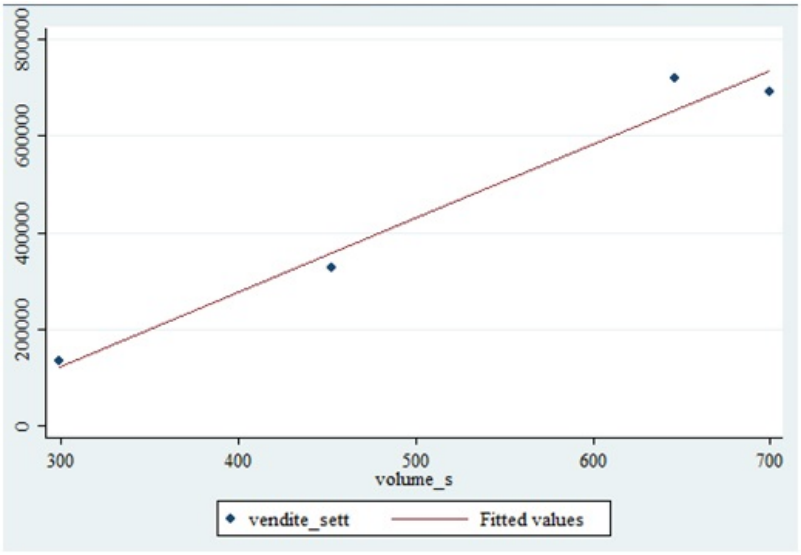


Figura 5.3: Regressione Lineare film "Les Misérables"

si é eseguito su stata il comando:

regress vendite_ s positivi negativi if(film==N)

dove N indica il numero identificativo di un film.

Avendo questa volta 2 variabili in gioco é necessario che esistano almeno 4 record per ogni singolo film, quindi il numero di film diminuisce perché pochi film sono rimasti nella top 20 delle vendite per almeno 4 settimane consecutive. Degli 11 film che superano questa limitazione solo 3 di questi presentano un Rsquared inferiore allo 0.7 quindi possiamo dire che per la maggior parte dei film vediamo che l'andamento dei tweet positivi alto, porta effettivamente ad avere buone vendite e che quindi il parere sul web risulta abbastanza indicativo sul successo o meno di un film. In tabella 5.3 sono mostrati i dati relativi all'Rsquared calcolato su tutti i film.

5.1.4 Vendite settimanali e volumi associati alla classe Neutri

Ora, come in precedenza, andiamo a definire se esiste la regressione lineare, e in questo caso tra le vendite settimanali e i tweet relativi al film filtrati attraverso il tool Carcaci, piú precisamente associati alla categoria dei neutri. Questo significa vedere se l'andamento é rappresentabile dall'equazione:

$vendites = \alpha + \beta neutri$

Andando ad analizzare questi dati vediamo che sui 24 film a disposizione 15

possiedono un R-squared superiore allo 0.7, 4 compresi tra lo 0.6 e lo 0.7 e 5 inferiori allo 0.6. Questo ci permette di notare che nonostante il filtro applicato ai tweet utilizzando il tool Carcaci, non abbiamo rilevato grosse differenze rispetto ai volumi settimanali non filtrati. Questo risultato infatti é dovuto al fatto che non sono molti i tweet che sono stati tagliati da questo filtraggio anche perché giá dalla query di estrazione da Twitter avevamo dei tweet validi per la nostra analisi. Come per gli altri casi, presentiamo in Tabella 5.4 di tutti gli R-squared per film.

5.2 VERIFICA CORRELAZIONE COMPLESSIVA SU DATI ITALIANI

Dopo aver analizzato l'andamento dei dati per singolo film e notato che per la maggior parte dei film c'é una correlazione lineare tra vendite e volumi di tweet in rete, ora andiamo a vedere se considerando complessivamente tutti i film assieme, sempre nel contesto italiano, si puo' dire che esiste questa correlazione in modo piú generale. Anche qui facciamo le nostre analisi verificando che se esiste o meno una correlazione tra:

- Vendite cumulate e volumi cumulati
- Vendite settimanali e volumi settimanali
- Vendite settimanali e sentiment settimanale (positivi e negativi)
- Vendite settimanali e volumi associati alla classe Neutri

5.2.1 Vendite cumulate e volumi cumulati

Nell'analisi effettuata per singolo film avevamo degli ottimi risultati per quanto riguarda la relazione vendita/volumi per valori cumulati, invece se consideriamo tutti i dati di tutti film assieme abbiamo un valore dell'R-squared molto basso, piú precisamente pari allo 0.3820. Peró sarebbe banale trarre conclusioni affrettate dicendo che non esiste alcuna relazione di tipo lineare. Andando ad analizzare i dati e soprattutto guardando il grafico, che viene mostrato di seguito, abbiamo la maggior parte dei valori che si concentrano lungo l'andamento lineare, mentre un numero limitato di punti si trova nettamente sotto o nettamente sopra questo andamento. Andando piú nel dettaglio i punti divergenti appartengono sostanzialmente a quattro film:

- Noi siamo infinito
- Il principe abusivo

- Spring Breakers - Una vacanza da sballo
- Viva la libertá

Per questi film sono stati osservati manualmente i tweet che la query ha estratto da Twitter e abbiamo notato che per 3 di questi 4 film, abbiamo un numero non trascurabile di tweet non attinenti in senso stretto alla pellicola cinematografica. Questo é successo perché i titoli dei film, "Noi siamo infinito, "Una vacanza da sballo, "Viva la libertá, sono composti da termini o da frasi anche di uso comune e che sono state incluse erroneamente nei nostri dati di analisi. Infatti era abbastanza eclatante la discrepanza dei dati per il film "Noi siamo infinito dove si sono registrati dati di vendita modesti, dell'ordine dei 400000 euro settimanali di punta, contro volumi su Twitter di oltre 2000 tweet settimanali, quando un film come "Lincoln con incassi settimanali arrivati ad oltrepassare i 2 milioni di euro, si ritrovava un volume inferiore alla metá. Per quanto riguarda invece il film "Il principe abusivo non si sono trovate irregolaritá nei tweet, l'unica cosa che si puo' osservare é che é stato un film "fuori norma, infatti ha avuto una pubblicitá massiccia, denotata da volumi altissimi per la maggior parte neutri e non esternanti un parere, accompagnato da un risultato di vendite al botteghino molto alto, infatti nel nostro periodo d'osservazione risulta essere il film che ha incassato totalmente di piú con oltre 14 milioni di euro seguito da "Il grande e potente Oz che ha incassato poco piú di 7 milioni (Figura 5.4). Per questo, si

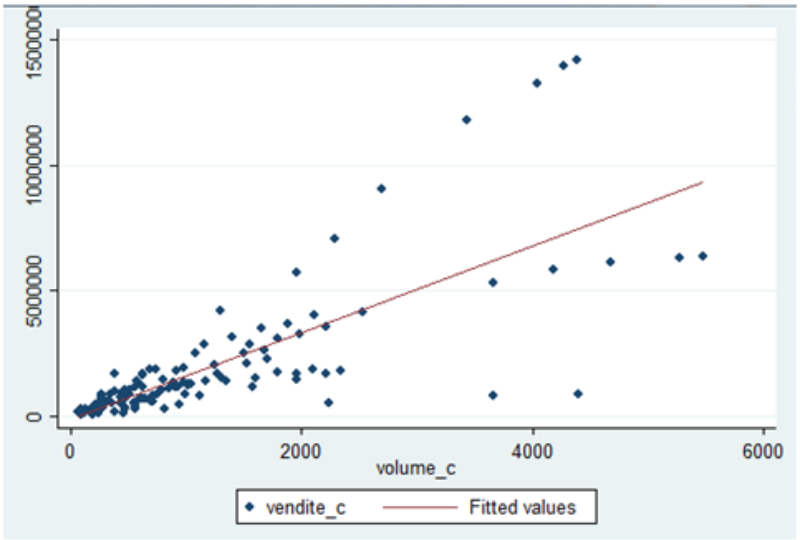


Figura 5.4: Grafico vendite/volumi cumulati con tutti i dati

é deciso di eliminare dalla nostra analisi i 4 film precedentemente citati visto la loro "irregolaritá. Applicando questo filtro notiamo che l'R-squared aumenta notevolmente arrivando al valore di 0.8792. A sostegno del valore dell'R-squared viene anche il grafico ricavato dai dati, dove si nota come l'andamento segue quello di una linea retta (Figura 5.5). Vendite settimanali e volumi settimanali

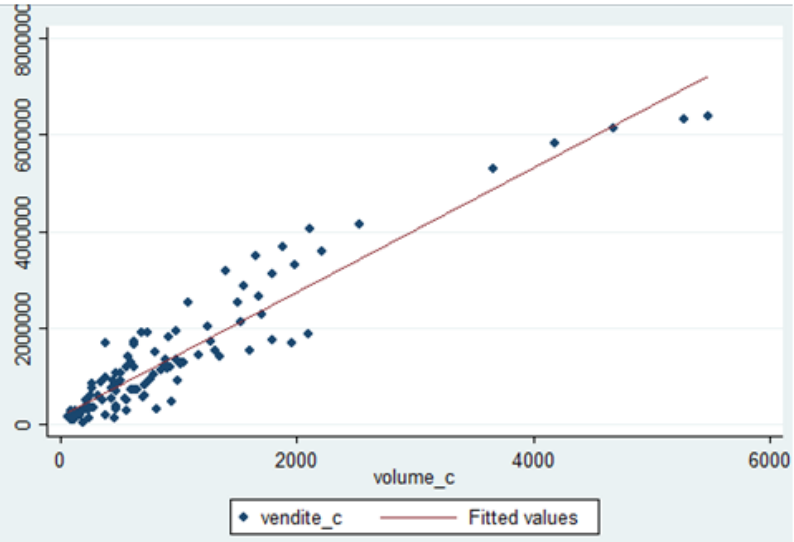


Figura 5.5: Grafico vendite/volumi cumulati con i dati filtrati

Per quanto riguarda l'analisi di correlazione dei dati settimanali é stata fatta una prima verifica circa a metá del nostro periodo di osservazione, intorno alla fine di febbraio, e considerando solo i film che avevano completato il loro ciclo di vita, per la precisione 14 film, abbiamo notato un R-squared buono, pari allo 0.7622 comprovato in Figura 5.6. Analizzando tutti i film peró notiamo, come

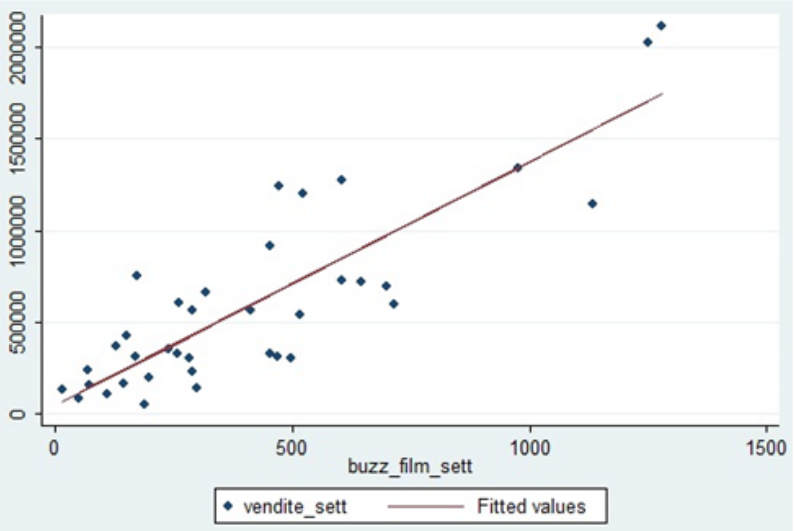


Figura 5.6: Grafico vendite/volumi settimanali con i dati parziali

per i dati cumulati, un risultato piuttosto deludente con un R-squared pari a 0.2201, nettamente inferiore allo 0.7. Peró compiendo gli stessi tagli effettuati nella precedente analisi, abbiamo anche qui un risultato molto piú soddisfacente, infatti ci ritroviamo con un valore di R-squared pari a 0.7606 e con un grafico

dell'andamento dei dati meno pregevole rispetto all'analisi con i dati cumulati, ma comunque accettabile (Figura 5.7).

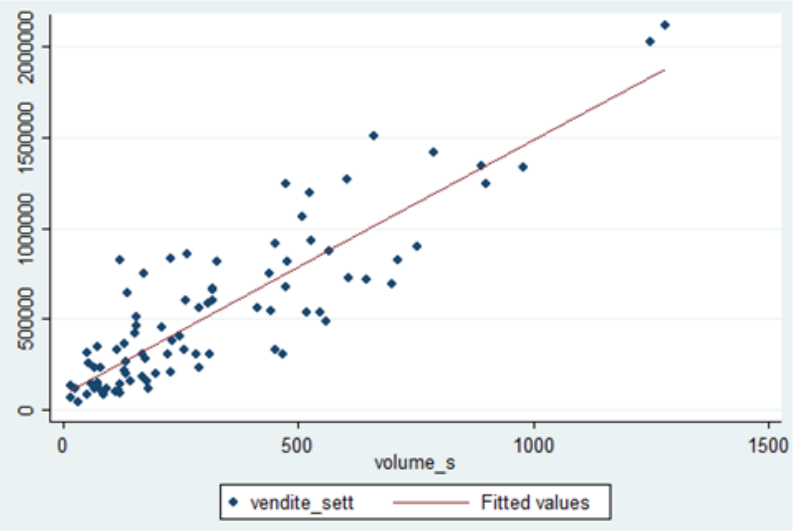


Figura 5.7: vendite/volumi settimanali con i dati totali filtrati

5.2.2 Vendite settimanali e sentiment settimanale (positivi e negativi)

L'analisi presa in esame ora é piuttosto interessante perché ci permette di capire che effettivamente il sentiment ricavabile dai vari tweet puo' essere indicatore di un andamento positivo o negativo di vendita. Il dato curioso di questo caso é che utilizzando i dati al completo abbiamo giá un andamento quasi lineare, infatti ci troviamo un R-squared pari a 0.6302, di poco inferiore al livello di accettabilitá. Questo puo' essere motivato dal fatto che i tweet sono stati filtrati dal tool Carcaci che ha permesso di eliminare da gran parte dei tweet fuori contesto che l'estrazione da Twitter aveva incluso. Applicando ai dati il taglio dei 4 film contenenti valori fuori norma, notiamo che il valore dell'R-squared cresce al valore di 0.7385.

5.2.3 Vendite settimanali e volumi associati alla classe Neutri

Come ultima casistica di studio, vediamo la correlazione tra le vendite settimanali e i tweet associati dal tool Carcaci alla cosiddetta classe Neutri che mi rappresenta una scrematura dei tweet con l'intento di eliminare quelli che erroneamente erano stati inclusi nell'atto in cui sono stati estratti dalla rete. I risultati, anche per questa casistica, si presentano come nei casi precedenti, nella fattispecie, l'analisi eseguita su tutti i dati di tutti i film presenta un andamento poco prossimo ad una funzione lineare con un valore di R-squared pari allo

0.3904. Peró filtrando i nostri dati nelle modalitá e con le motivazioni specificate nell'analisi delle vendite cumulate, vediamo che il fattore R-squared assume un valore pari allo 0.7260 che risulta accettabile. Il grafico di questa analisi é visibile in Figura 5.8: In Tabella 5.5 successiva viene presentato un riepilogo

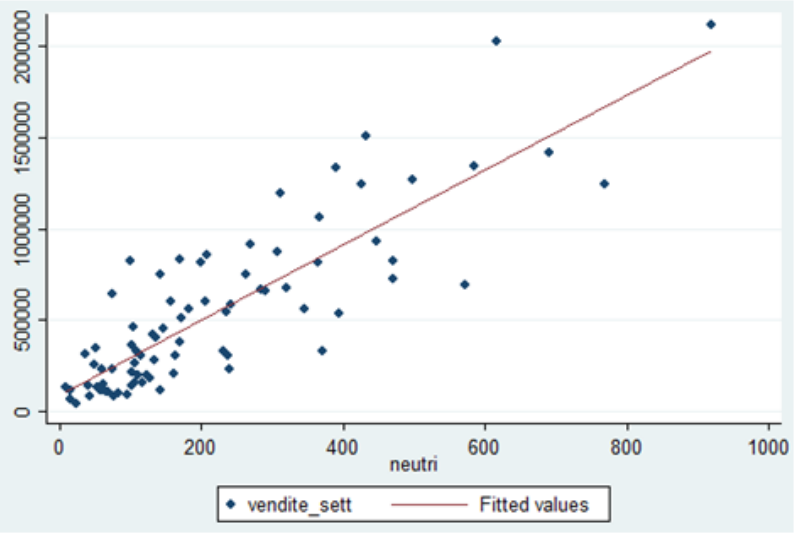


Figura 5.8: vendite settimanali/neutri con i dati totali filtrati

dei risultati dell'R-squared appena illustrati estratti dall'analisi di correlazione fatta sui dati relativi a tutti i film.

5.3 VERIFICA CORRELAZIONE SU DATI AMERICANI

In questa sezione, come ultimo step, andiamo ad effettuare la verifica di una possibile relazione tra vendite e volumi per quanto riguarda i dati relativi alle pellicole uscite negli Stati Uniti. Come giá anticipato, per questo gruppo di dati non effettueremo un'analisi tra le vendite e il sentiment, quindi le relazioni che verifichiamo saranno solo tra:

- Vendite cumulate e volumi cumulati per singolo film
- Vendite settimanali e volumi settimanali per singolo film
- Vendite cumulate e volumi cumulati sull'intero set di dati
- Vendite settimanali e volumi settimanali sull'intero set di dati

Prima di mostrare i risultati delle analisi é importante specificare alcune caratteristiche dei dati. A differenza dei dati relativi ai film italiani, i tweet relativi ai film americani sono stati estratti a partire dal 18 di Febbraio fino al 24 Marzo, questo implica che essendo piú stretto l'asse temporale, sono presenti meno film

che hanno compiuto un intero "ciclo di vita. Quindi le nostre analisi si basano su meno film e meno record di dati rispetto alle analisi precedenti, infatti se prima contavamo sui dati di 40 film, ora abbiamo i dati relativi a 16 film con 58 record totali di dati. Inoltre i dati di vendita e i volumi americani risultano essere in proporzione, rispetto a quelli italiani, molto piú elevati, infatti se prima il film col maggior incasso in Italia aveva guadagnato 14 milioni di euro, negli Stati Uniti abbiamo facilmente film che si aggirano sui 60 milioni di dollari e il film col maggior incasso nel nostro periodo di osservazione, dal titolo "Identity Thief, ha ricavato quasi 128 milioni di dollari. In aggiunta é da segnalare anche il fatto che i film negli Stati Uniti escono in date leggermente differenti tra Stati diversi e questo comporta periodi molto lunghi in cui i film si presentano nelle classifiche dei botteghini.

5.3.1 Vendite cumulate e volumi cumulati per singolo film

A questo punto siamo andati a verificare se esiste una regressione lineare tra la variabile indipendente rappresentata dalle vendite cumulate e la variabile dipendente basata sui volumi cumulati, ovviamente per singolo film. Questo é stato effettuato eseguendo il seguente comando stata:

regress venditec volumec if(film==N)

Per poter avere dei risultati é necessario che un singolo film abbia almeno 3 record di dati e quindi sia stato presente nella classifica dei 20 film con i maggiori incassi per almeno 3 settimane. Per questo motivo abbiamo solo 11 film. Nella tabella seguente elenchiamo i valori degli R-squared per ciascuno di questi film:

Come si puo' notare vediamo che a parte il risultato relativo al film "Jack the Giant Slayer, tutti i valori dell'R-squared sono superiori allo 0.7 e quindi si puo' affermare che presi i film singolarmente, tra i dati di vendita cumulati e i volumi cumulati esiste una correlazione lineare. Questa conclusione si puo' notare anche dal grafico (Figura **??**) che mostra l'andamento per il film "A Good Day To Die Hard.

5.3.2 Vendite settimanali e volumi settimanali per singolo film

Ora andremo a verificare sempre la presenza di regressione lineare ma tra le variabili vendita e volumi espressi nei valori per settimana. Come valeva per i valori cumulati, anche qui abbiamo 11 film su cui poter effettuare questo tipo di analisi, e questo a portato ad avere i valori di R-squared riportati in Tabella 5.6.

Anche in questo a caso possiamo constatare che tutti tranne uno, "Identity Thief, presentano un andamento prossimo a quello lineare. Ovviamente va

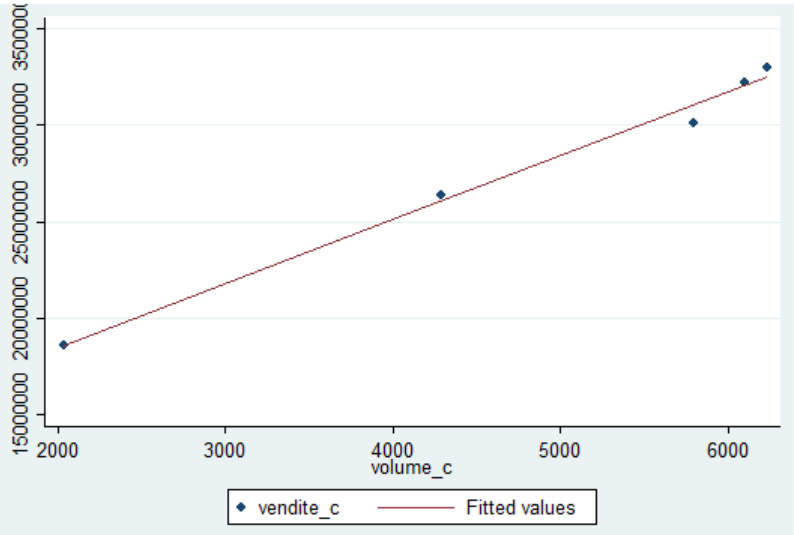


Figura 5.9: Andamento vendite/volumi cumulati per il film "A Good Day To Die Hard

constatato che, rispetto alla precedente analisi, in media, i valori dell'R-squared risultano piú bassi ma sempre superiori o prossimi allo 0.7. Vendite cumulate e volumi cumulati sull'intero set di dati L'analisi seguente riguarda invece tutti i dati di vendita e i volumi delle varie pellicole considerate in modo globale con le variabili espresse in modo cumulato. Eseguendo il semplice comando:

regress venditec volumec

abbiamo rilevato un R-squared pari allo 0.7525 superiore alla soglia di accettabilitá della situazione di correlazione lineare. Questa situazione possiamo visualizzarla Figura 5.10 dove si nota un andamento piú scostante dalla retta rispetto a quello riscontrato con la stessa analisi con i film italiani

Vendite settimanali e volumi settimanali sull'intero set di dati L'ultima analisi che andiamo ad affrontare per i film usciti nelle sale americane é la verifica dell'esistenza della regressione lineare tra le vendite e i volumi presi settimanalmente e quindi verificare se l'andamento delle variabili si discosta di poco dalla retta di equazione:

$vendites = \alpha + \beta volumis$

Eseguendo il software statistico troviamo un R-squared pari allo 0.6598 inferiore alla soglia di accettazione dello 0.7. Quindi attraverso questi dati non possiamo affermare con certezza che la funzione sia approssimabile all'andamento lineare. Comunque il valore dell'R-squared non é molto lontano dal valore soglia e per questo il grafico non risulta troppo distante dall'andamento lineare anche se é abbastanza visibile la dispersione dei punti (Figura 5.11).

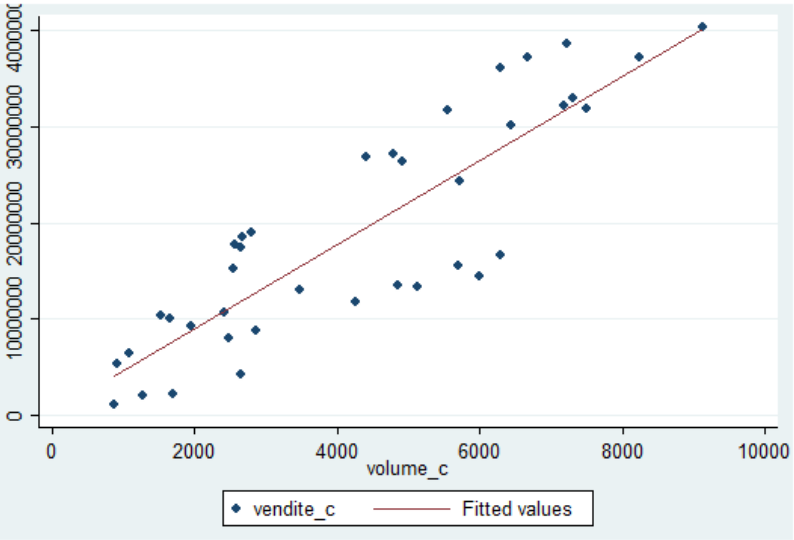


Figura 5.10: Andamento vendite/volumi cumulati su tutti i dati usa

TITOLO	**R-Squared**
Amiche da morire	0.9960
Anna Karenina	0.9947
Educazione Siberiana	0.9975
Flight	0.9859
Gambit	0.9679
Il grande e potente Oz	0.9980
Il lato positivo –Silver Linings Playbook	1.0000
Il principe abusivo	0.9957
La cuoca del presidente	0.9972
Les Misèrables	0.9817
Lincoln	0.9458
Looper	0.9810
Noi siamo infinito	0.9672
Pazze di me	0.9976
Pinocchio	0.9725
Quartet	0.9552
Re della terra selvaggia	0.9585
Spring Breakers –Una vacanza da sballo	0.9740
The Impossible	0.9823
Upside Down	0.9968
Viva la libertá	0.9992
Warm Bodies	0.9943
Zambezia	0.9607
Zero Dark Thirty	0.9661

Tabella 5.1: R–Squared per film 3 con piú record.

TITOLO	R-Squared
Amiche da morire	0.6855
Anna Karenina	0.8830
Educazione Siberiana	0.7779
Flight	0.7898
Gambit	0.7948
Il grande e potente Oz	0.8661
Il lato positivo - Silver Linings Playbook	0.6260
Il principe abusivo	0.9781
La cuoca del presidente	0.8485
Les Misèrables	0.9687
Lincoln	0.8803
Looper	0.8970
Noi siamo infinito	0.5032
Pazze di me	0.7312
Pinocchio	0.5828
Quartet	0.1028
Re della terra selvaggia	0.0033
Spring Breakers –Una vacanza da sballo	0.9959
The Impossible	0.8871
Upside Down	0.9466
Viva la libertá	0.9005
Warm Bodies	0.7002
Zambezia	0.0738
Zero Dark Thirty	0.1273

Tabella 5.2: Correlazione tra vendite settimanali e volumi settimanali.

TITOLO	R-Squared
Anna Karenina	0.9775
Educazione Siberiana	0.9808
Flight	1.000
Il principe abusivo	0.9855
Lincoln	0.7921
Pinocchio	0.5762
Quartet	0.6522
The Impossible	0.9985
Upside Down	0.9584
Viva la libertá	0.9762
Zambezia	0.4372

Tabella 5.3: Correlazione tra vendite settimanali e sentiment settimanale.

TITOLO	R-Squared
Amiche da morire	0.6391
Anna Karenina	0.7567
Educazione Siberiana	0.7432
Flight	0.9744
Gambit	0.8170
Il grande e potente Oz	0.8170
Il lato positivo - Silver Linings Playbook	0.8743
Il principe abusivo	0.6840
La cuoca del presidente	0.9840
Les MisÃ©rables	0.8226
Lincoln	0.9941
Looper	0.7460
Noi siamo infinito	0.9402
Pazze di me	0.8378
Pinocchio	0.6117
Quartet	0.1670
Re della terra selvaggia	0.0170
Spring Breakers - Una vacanza da sballo	0.0176
The Impossible	0.9466
Upside Down	0.9776
Viva la libertÃ	0.8863
Warm Bodies	0.6464
Zambezia	0.0975
Zero Dark Thirty	0.1535

Tabella 5.4: Correlazione tra vendite settimanali e volumi associati alla classe Neutri.

	Dati complessivi	Dati filtrati
Vendite cumulate/volumi cumulati	0.3820	0.8792
Vendite settimanali/volumi settimanali	0.2201	0.7606
Vendite settimanali/sentiment	0.6302	0.7385
Vendite settimanali/neutri	0.3904	0.7260

Tabella 5.5: Riepilogo R–Squared

TITOLO	R-Squared
A Good Day To Die Hard	0.7638
Dark Skies	0.9703
Dead Man Down	0.7272
Escape From Planet Earth	0.9008
Identity Thief	0.2447
Jack the Giant Slayer	0.9825
Quartet	0.6925
Safe Haven	0.8893
Side Effects	0.9757
Snitch	0.8712
Warm Bodies	0.9471

Tabella 5.6: Correlazione tra vendite settimanali e volumi settimanali per singolo film.

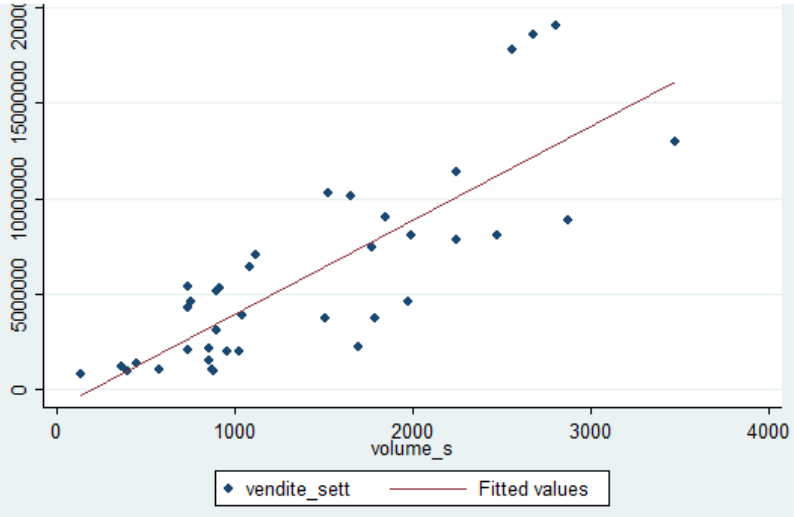


Figura 5.11: Andamento vendite/volumi settimanali su tutti i dati usa

Capitolo 6

CONCLUSIONI

In questo elaborato di tesi siamo andati ad analizzare e a dimostrare con dati concreti, l'effettiva importanza e utilitá dal punto di vista previsionale, nell'ambito delle vendite, della mole di dati che si possono ricavare dai Social Media, nel nostro caso dalla piattaforma di microblogging Twitter. Abbiamo individuato le vendite ai botteghini delle pellicole cinematografiche, il contesto adatto a questo tipo di analisi, che ci ha permesso di accedere facilmente ai dati di vendita e ed inoltre é un contesto piuttosto popolare e discusso su Twitter. Proprio perché si parla molto di film nei Social Media, era necessario pensare ed ideare un modo piuttosto valido e preciso per estrarre dalla rete solo i dati utili all'analisi. Una volta riusciti ad entrare in possesso di tutti i dati attraverso vari crawler software, si é passata all'analisi vera e propria dei dati.

Analizzando i film all'interno del contesto italiano, possiamo concludere che l'andamento delle vendite ai botteghini di ogni singola pellicola é seguito di pari passo dal volume di discussione presente in rete su di esso e questo volume, é spesso caratterizzato da semplici citazioni del film. Tutto ció nonostante i film in Italia non stiano nelle sale a lungo, solitamente non piú di 4 settimane. Per questo si puo' tranquillamente affermare che il parlare di un film su Twitter influenza i vari utenti del Social Media portando una sorta di curiositá che si traduce in maggiore visione per quel film. Invece per pellicole poco chiacchierate vediamo che gli incassi sono inferiori. Inoltre, a prescindere dal successo o meno di un film, nelle dovute proporzioni, tutti i film presentano volumi di citazione su di esso molto elevato in prossimitá dell'uscita nelle sale o delle anteprime, e si protrae per una settimana (massimo due per le grandi pellicole), per poi calare inesorabilmente sempre di piú fino ad azzerarsi quasi completamente. Questo tipo di andamento rispecchia pienamente quello delle vendite dove per le pellicole molto pubblicizzate o in cui sono presenti grandi interpreti, abbiamo grossi dati di vendita nelle prime due settimane per poi calare vistosamente fino al termine della loro presenza nelle sale. Tutto questo ci porta a dire che, nel contesto da noi analizzato, il quanto si parla di un film é indicatore di successo o meno della pellicola in questione e puo' essere analizzato in modo tale da

poter prevedere con un certo anticipo il successo o fallimento di una produzione cinematografica.

Analizzando i film in modo complessivo e non prendendoli uno ad uno, si é comunque rilevato, anche se con meno evidenza, una correlazione tra le vendite e i volumi dei tweet. Questo avvalora ancor di piú quanto precedentemente affermato in quanto anche considerando film di diverso genere, con piú o meno attori di fama mondiale che differiscono sensibilmente anche nei periodi di uscita nelle sale, questi presentano un'evoluzione nel tempo nell'ambito delle vendite simile tra loro e che é ben descrivibile anche attraverso i volumi dei tweet. Inoltre attraverso l'analisi del sentiment dei tweet estratti, si é notato una particolare evidenza di tweet positivi per le pellicole importanti, che sono stati premiati o canditati al premio Oscar e che hanno avuto incassi notevoli. Questo perché i giudizi e i pareri presenti sui nuovi media, in questo caso Twitter, rappresentano la piú vasta bacheca di recensioni del mondo che ha il potere di essere altamente influenzante. Quindi il giudicare positivamente un film, un attore o il regista del lungometraggio, porta a tutti coloro che entrano in contatto con questi messaggi, a un'idea positiva sull'oggetto in questione con evidenti conseguenze sulle vendite. E' stato interessante notare la notevole quantitá di tweet positivi presenti contro una bassissima percentuale di tweet negativi. Per pellicole con poco successo sono presenti piú tweet con parere negativo rispetto a coloro che hanno ricevuto i maggiori incassi, ma la percentuale di questi rispetto al totale é abbastanza ridotta. Questo ci porta a due possibili deduzioni, che anche pochi pareri negativi risultano efficaci in questo tipo di media per screditare una pellicola, oppure che é importante il parere su un film, ma lo é ancor di piú il quanto se ne parla. Con quest'ultima assunzione vogliamo intendere che come indice di fallimento di un film i tweet a sentiment negativo sono importanti ma é significativo considerare anche l'aver parlato poco di quel film.

Per quanto riguarda l'analisi dei dati relativi ai film americani abbiamo effettuato delle considerazioni nell'ambito dei volumi e questo a portato a dei risultati meno brillanti rispetto al contesto italiano ma pur sempre positivi. Questo potrebbe essere dovuto particolarmente al breve periodo di osservazione che abbiamo dedicato all'andamento del mercato americano che essendo in proporzioni piú elevato, andava monitorato per piú tempo.Nell'analisi effettuata sull'andamento dei singoli film abbiamo notato una forte correlazione tra le vendite e i volumi e lo stesso si puo' dire sull'andamento delle vendite cumulate calcolato sull'intera serie di dati.

Come possibili sviluppi a questo tipo di analisi sarebbe importante andare sicuramente ad analizzare meglio il mercato americano che differisce da quello italiano soprattutto nelle modalitá d'uscita nelle sale delle pellicole, infatti spesso capita che alcuni film escano con una o due settimane di ritardo in certi stati rispetto ad altri all'interno degli Stati Uniti. Un'ulteriore studio potrebbe essere quello di analizzare in dettaglio le uscite in anteprima dei film e quin-

di vedere se la presenza e il successo di un'anteprima é fortemente indicatrice dell'andamento delle vendite del film una volta uscito definitivamente nelle sale.

Bibliografia

[1] http://oreilly.com/pub/a/web2/archive/what-is-web-20.html?page=1

[2] http://www.thegoodmorning.com/2-0/decalogo-del-web-2-0.html

[3] http://it.wikipedia.org/wiki/Socialmedia

[4] Kaplan and Haenlein, 2009

[5] http://www.alittlebit.it/chi-sono-gli-utenti-dei-social-media/

[6] http://nielsenfeaturedinsight.mag-news.it/nl/nielsenlink4496.mn

[7] http://peracchiafloriana.wordpress.com/web-20-media-sociali-vs-media-tradizionali/

[8] http://wearesocial.it/blog/2011/01/lutilizzo-dei-social-network-italia-dati-statistiche/

[9] http://it.wikipedia.org/wiki/Socialmediamarketing

[10] http://socialmediamarketing.nextep.it/2010/01/20/strategie-di-social-media-marketing/more-2117

[11] http://blog.tagliaerbe.com/2009/05/brand-reputation-management-20.html

[12] http://www.slideshare.net/gzarantonello/social-media-marketing-strategia-e-casi-concreti

[13] http://www.adamwesterski.com/wp-content/files/docsCursos/sentimentAdocTLAW.pdf

[14] http://www.slideshare.net/marketingarena/brand-reputation-tools

[15] http://onlinelibrary.wiley.com/doi/10.1111/j.1475-4932.2012.00809.x/pdf

[16] http://www.lospaziodellapolitica.com/2013/02/prevedere-il-presente-e-limmediato-futuro-con-twitter-adelante-con-juicio/

[17] http://it.wikipedia.org/wiki/Crawler

[18] http://jsoup.org/

[19] http://it.wikipedia.org/wiki/Twitter

[20] https://dev.twitter.com/

Elenco delle figure

Elenco delle tabelle