UNIVERSITÀ DEGLI STUDI DI
MODENA E REGGIO EMILIA

Dipartimento di Scienze Fisiche, Informatiche e Matematiche

CORSO DI LAUREA IN INFORMATICA

Analisi e Valutazione Sperimentale di Tecniche di Sentiment Analysis Basate su Machine Learning e Dizionari

Daniela Conti

Tesi di Laurea

Relatore:
Ing. Riccardo Martoglia

Anno Accademico 20014-2015

A tutti quelli che

non si sono mai arresi

RINGRAZIAMENTI

Ringrazio l'Ing. Riccardo Martoglia per la continua disponibilità e pazienza dimostrata durante questi mesi di lavoro.

Un ringraziamento speciale va ai miei genitori e a mia sorella, che hanno sempre appoggiato le mie scelte e non hanno mai smesso di credere in me.

Ringrazio Luca con il quale ho condiviso le ansie degli ultimi esami e per l'aiuto tempestivo nel momento del bisogno.

Infine, e non per ordine di importanza, ringrazio Salvatore che mi ha sostenuto e con il quale ho condiviso gioie e dolori in questi anni universitari.

PAROLE CHIAVE

Base di Dati

Sentiment Analysis

WordNet

SentiWordNet

Naïve Bayes

Indice

Elenco delle figure

Introduzione

La nascita e l'evoluzione del web ha cambiato radicalmente la vita dell'uomo semplificandola e trasformando totalmente il modo di comunicare e di esprimere opinioni. La filosofia del web 2.0 si basa su tre concetti fondamentali: il *software come servizio,* dove i software non sono più elementi statici installati sui personal computer ma diventano gestibili tramite apposite interfacce web permettendo quindi di lavorare in mobilità, i concetti *condivisione* e *partecipazione* diventano termini fondamentale nella nuova rete sociale dove ogni utente si sente libero di condividere con il resto del mondo, o con una cerchia ristretta di utenti, i propri pensieri e le proprie opinioni in merito ad un prodotto o ad un servizio.
Questa proliferazione di informazione libera e gratuita causa l'*Information Overload* cioè il fenomeno che si manifesta quando la quantità di dati che vengono generati quotidianamente dal web è superiore rispetto a quella che si è in grado di analizzare e gestire. Tutto questo ha portato alla creazione di sistemi complessi in grado di raccogliere questi dati grezzi ed elaborarli con l'obbiettivo di ottenere delle informazioni potenzialmente utili. Negli ultimi anni si è intensificato lo studio e la ricerca sulla Sentiment Analysis, ovvero Trattamento Automatico del Linguaggio, il cui scopo è quello di identificare e classificare informazioni di tipo soggettivo.
Nella presente tesi sarà discusso ed approfondito il ruolo della Sentiment Analysis e le tecniche utilizzate per estrapolare le informazioni dei dati presenti in alcuni siti internet più cliccati del web.
La presente tesi è suddivisa in due parti ed è strutturata in cinque capitoli. Nella prima parte sono spiegati gli argomenti e gli strumenti utilizzati durante la fase di

ricerca; mentre la seconda parte è relativa all'implementazione dell'applicazione. Il documento è così strutturato:

- *Parte I: Il caso di studio*
  - **Capitolo 1: Introduzione al Sentiment Analysis.** Il seguente capitolo ha lo scopo di illustrare cosa è il Sentiment Analysis, quali sono gli obiettivi e le tecniche utilizzate.
  - **Capitolo 2: Analisi Testuale.** In questo capitolo sono presentati due database lessicali, quali WordNet e SentiWordNet. È presentata, inoltre, la libreria NLTK e il fenomeno lessicale del Disambiguation.
- *Parte II: Progetto e valutazione*
  - **Capitolo 3: Naïve Bayes e Dictionary Based Approach.** Questo capitolo presenta l'approccio machine learning basato sul Naïve Bayes e l'approccio basato sui dizionari. Sono presentati inoltre, i siti e le categorie da cui sono stati raccolti i dati sui quali saranno eseguiti i test.
  - **Capitolo 4: Valutazione efficacia dei due metodi.** E' il capitolo più importante dell'intero progetto perché sono presentati e analizzati i risultati dei test che sono stati eseguiti sui due approcci esposti nel Capitolo 3.
  - **Capitolo 5: Considerazioni finali.** Descrive le conclusioni finali cui si è giunti alla fine dei test svolti, con aggiunta dei possibili sviluppi futuri.

Parte I

Il caso di studio

Capitolo 1

Introduzione al Sentiment Analysis

Negli ultimi anni si è registrata una crescita esponenziale dell'uso dei social media, termine generico utilizzato per indicare il mezzo tramite il quale più persone condividono contenuti testuali, immagini, video, e audio. Oggi giorno, sempre più aziende, utilizzano i social media come mezzo per il loro processo decisionale e strategie di mercato anche se, a causa della proliferazione di migliaia di siti, risulta essere un problema molto arduo. Tale problema è intensificato anche a causa dell'enorme quantità di testo non sempre decifrabile rafforzato da un'analisi molto soggettiva dovuta alle proprie attitudini e preferenze con conseguente alterazione dell'opinione stessa.

Scopo del Sentiment Analysis è estrarre, classificare, capire e valutare le opinioni espresse in varie fonti quali notizie online, commenti dei social media e altri contenuti creati dagli utenti.

Il SA è nato in ambito pubblicitario per sondare il gradimento di nuovi prodotti immessi sul mercato e attualmente trova la sua evoluzione e specifica identità in attività di intelligence multimediale a livello sociologico, politico, economico, pubblicitario e di sicurezza nazionale. Fino a qualche tempo fa gli strumenti per la rilevazione dei consensi e delle opinioni avvenivano tramite sondaggi e indagine statistiche, ma oggi, grazie alle tecniche di Opinion Mining si hanno costi di rilevazione nettamente inferiori e in molti casi molta più autenticità informativa:

gli utenti non sono vincolati ad esprimere opinioni così come succedeva in passato, ma al contrario, queste fluiscono liberamente senza alcuna costrizione.

1.1. Opinion o Sentiment?

Il Sentiment Analysis (SA) è conosciuto anche con il nome di Opinion Mining (OM).

Queste due espressioni sono considerate ambivalenti e vengono utilizzati in maniera indifferente: il termine Opinion Mining è più comune nel mondo accademico, mentre il termine Sentiment Analysis è più comune nelle organizzazioni. Tuttavia, alcuni ricercatori affermano che OM e SA sono nozioni leggermente diverse. Per Opinion Mining si intende l'opinione espressa da più utenti su un soggetto; mentre per Sentiment Analysis si identifica il sentimento espresso in un testo. Per capire se esiste realmente una differenza tra Opinione e Sentimento, è utile far riferimento alla definizione formale presente all'interno della lingua italiana:

- **Opinione**: *concetto che una o più persone si formano riguardo a particolari fatti, fenomeni, manifestazioni, quando, mancando un criterio di certezza assoluta per giudicare della loro natura (o delle loro cause, delle loro qualità, ecc.), si propone un'interpretazione personale che si ritiene esatta e a cui si dà perciò il proprio assenso*[1].
- **Sentimento**: *ogni forma di affetto, di impulso dell'animo, di movimento psichico, di emozione, sia che rimangano chiusi entro l'animo della persona stessa, sia che si rivolgano e proiettino verso gli altri, verso il mondo esterno; modo di pensare e di sentire, considerato come parte del carattere di una persona, come complesso delle inclinazioni al bene o al male, come guida del comportamento morale*[2].

[1] http://www.treccani.it/vocabolario/opinione/
[2] http://www.treccani.it/vocabolario/sentimento/

Definizioni di questo genere, sono perfette dal punto di vista linguistico ma troppo astratte e poco precise per una trattazione computazionale del problema. Inoltre le due definizioni sopra differiscono per molti aspetti, ma dal punto di vista del Sentiment Analysis, il termine opinione ed il termine sentimento, indicano la stessa identica cosa poiché permane l'obiettivo di trovare opinioni, identificare i sentimenti che vengono espressi e quindi classificare la loro rispettiva polarità. [6]

1.2. Obiettivo del Sentiment Analysis

L'obiettivo del Sentiment Analysis non è identificare il soggetto di un documento ma classificare le opinioni espresse sul medesimo. Questo processo di classificazione può essere riassunto nella seguente figura:

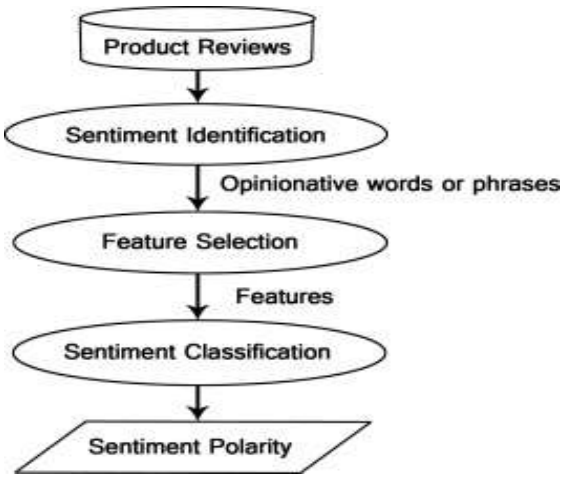

Figura 1: Sentiment analysis process on product reviews

Il Sentiment Analysis, come descritto in [2], può essere eseguito secondo tre principali livelli di classificazione:

- a *livello di documento*, si occupa di estrarre l'opinione generale espressa in un documento secondo un parere positivo o negativo. In questo livello è importante che l'opinione espressa affronti una sola entità o argomento;
- a *livello di frase*, dove il primo passo è identificare se la frase è di tipo soggettivo o oggettivo. Generalmente i fatti sono affermazioni oggettive

mentre le opinioni sono affermazioni soggettive. Se la frase è soggettiva il SA determinerà la polarità positiva o negativa.

- a *livello di aspetto*, mira a identificare le entità e i loro aspetti identificando quali caratteristiche sono stati apprezzati e criticati dal soggetto per la relativa entità.

In generale non vi è alcuna differenza fondamentale tra la classificazione a livello di documenti e a livello di frase, perché le frasi possono essere considerati documenti brevi. Inoltre, le opinioni sono espresse in testi non strutturati e questo complica l'analisi. Un modo per risolvere, almeno in parte il problema, è partire da un modello strutturato da elaborare in modo computazionale.

1.3. Tecniche di Classificazione

Negli ultimi anni sono state introdotte diverse tecniche atte allo studio del Sentiment Analysis. Queste tecniche, anno per anno, sono state raccolte e classificate. Per ogni nuovo argomento proposto sono state illustrate, mediante l'ausilio di tabelle e grafici, le definizioni, le problematiche legate allo sviluppo, eventuali tematiche non ancora trattate e argomenti non ancora risolti al fine di migliorare i nuovi obiettivi raggiunti.
Le tecniche di classificazione sul Sentiment Analysis possono essere approssimativamente suddivise in tre tipi di approcci:

- Machine Learning
- Basato sul Lessico
- Approccio Ibrido.

Questi approcci possono essere utilizzati singolarmente oppure possono essere utilizzati combinandoli tra di loro. Ogni approccio può essere suddiviso in due o più sottoapprocci. Ognuno di essi, può essere implementato con uno schema algoritmico che riprende i concetti statistici e del calcolo della probabilità oppure legati alla sintassi e semantica della lingua.

La figura 2 riassume le tecniche utilizzate:

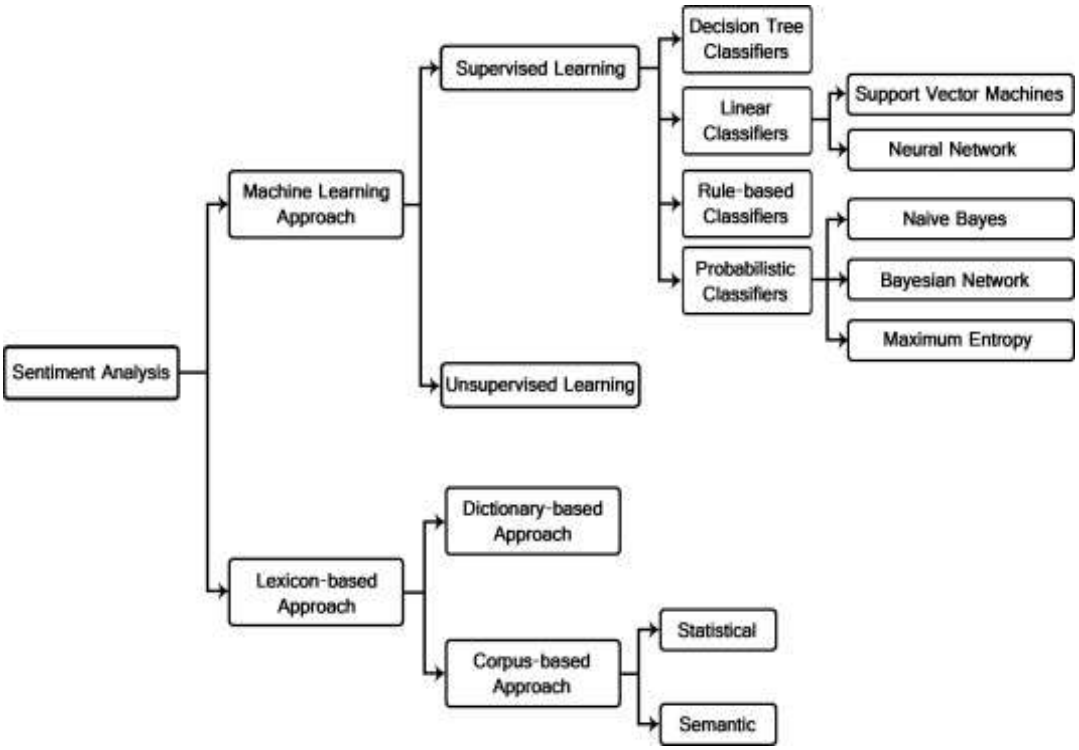

Figura 2: Schema delle Tecniche di Sentiment Analysis

Lo schema si divide in due grandi filoni: l'approccio Machine Learning e l'approccio lessicale.

L'approccio Machine Learning utilizza tecniche di intelligenza artificiale che, partendo da una raccolta di esempi pre-etichettati, permette di generalizzare la polarità di altri contenuti testuali generalmente rappresentati come vettori di features.

L'approccio lessicale utilizza un dizionario con informazioni riguardanti la positività, negatività, oggettività di parole o frasi. Questi dizionari possono essere creati manualmente o automaticamente.

L'approccio ibrido combina entrambi gli approcci. [2] [6]

1.3.1. Approccio Machine Learning

L'approccio Machine Learning (ML) o Apprendimento Automatico[3] è una delle aree fondamentali dell'intelligenza artificiale e si occupa della realizzazione di sistemi e algoritmi basati sull'osservazione dei dati per la sintesi di nuova conoscenza [2] [6]. Uno dei tanti obiettivi della ricerca sull'apprendimento automatico è imparare a riconoscere automaticamente modelli complessi e restituire informazioni corrette partendo da tecniche di addestramento dati.
I metodi di classificazione dei dati che utilizzano l'approccio ML possono essere suddivisi in due categorie:

- Supervised Learning;
- Unsupervised Learning.

La differenza sostanziale tra i due metodi consiste nel fatto che i metodi supervisionati (Supervised Learning) fanno uso di un insieme di documenti etichettati, mentre; i metodi non supervisionati (Unsupervised Learning), non dispongono di documenti etichettati e consiste nel fornire al sistema una serie di input che il sistema classificherà ed organizzerà sulla base di caratteristiche comuni per cercare di effettuare ragionamenti e previsioni sugli input successivi.
Tuttavia, ancora oggi, non si è ancora giunti alla costruzione di un sistema di apprendimento automatico paragonabile a quello umano; ma esistono algoritmi efficaci per alcuni modelli di apprendimento.
I metodi Supervisionati si possono ulteriormente suddividere in:

- *Decision Tree Classifier*: i classificatori di decisione ad albero sono costruiti utilizzando una serie di esempi di addestramento per i quali sono note le etichette di classe. La loro struttura consiste in una separazione gerarchica dello spazio dei dati di apprendimento in cui viene utilizzata una condizione sul valore dell'attributo per dividere i dati: la condizione

[3] http://www.tesionline.it/default/glossario.jsp?GlossarioID=5155

è la presenza o l'assenza di una o più parole. La divisione dello spazio dei dati avviene ricorsivamente fino ai nodi foglia, contenenti un determinato numero minimo di record che sono utilizzati a scopo di classificazione.

- *Linear Classifier*: l'obiettivo della classificazione statistica consiste nell'utilizzare le caratteristiche degli oggetti per identificare a quale cluster appartengono. Questo scopo è raggiunto basandosi sul valore di una combinazione lineare delle varie caratteristiche. Alla famiglia dei classificatori lineari appartengono il Support Vector Machines e il Neural Network.
    - Il *Support Vector Machine* (SVM), chiamato anche classificatore a massimo margine, poiché minimizza l'errore empirico di classificazione e massimizza il margine geometrico; consiste in un insieme di metodi per la regressione e classificazione di pattern. Il principio fondamentale del SVM è determinare i separatori lineari nello spazio di ricerca, in grado di separare meglio le classi diverse. Come è possibile osservare dalla figura sottostante, ci sono 2 classi *x* e *o*, e 3 iperpiani *A*, *B* e *C*. L'iperpiano *A* fornisce la migliore separazione tra le classi, perché la distanza normalizzata di uno dei punti di dati è la più grande, in modo da rappresentare il margine massimo di separazione.

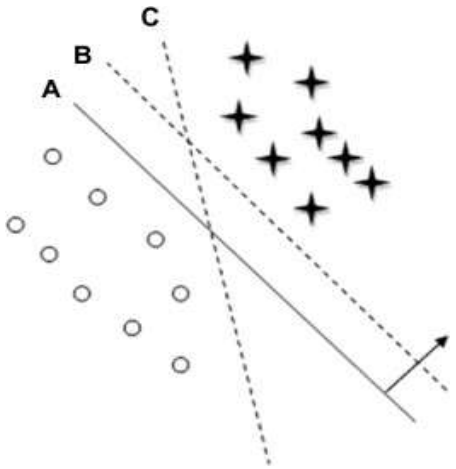

Figura 3: Schema esempio separatori lineari

- Il *Neural Network* (NN) è una tecnica appartenente alla famiglia di modelli di apprendimento statistici ispirate alle reti neurali biologiche come ad esempio il sistema nervoso centrale degli animali, in particolare del cervello. Questi, sono utilizzati per stimare e approssimare funzioni che possono dipendere da un gran numero di ingressi. In questo tipo di rete, l'unità di base è il neurone; mentre gli ingressi ai neuroni sono indicati con il vettore *overline* X_i che indica le frequenze delle parole contenute nell'i-esimo documento. A ciascun neurone sono associati una serie di pesi A utilizzati per il calcolo della funzione degli ingressi $f(\cdot)$. La funzione lineare della rete neurale è:

$$p_i = A * X_i$$

- *Rule-based Classifier*: sono classificatori in cui lo spazio per i dati è modellato con una serie di regole. La parte sinistra della regola rappresenta una condizione sul set di funzionalità, espresse in forma normale disgiuntiva, mentre il lato destro è l'etichetta della classe. La condizione è la presenza di un termine. La regola dell'assenza di un termine è usata raramente perché non è informativa in un contesto di dati sparsi.
  La differenza principale tra i classificatori di decisione ad albero e quelli basati su regole di decisione è che nei primi vi è una rigorosa suddivisione gerarchica dello spazio dati, mentre nei secondi classificatori sono consentite sovrapposizioni nello spazio di decisione.

- *Probabilistic classifier*: sono classificatori probabilistici che utilizzano una miscela di modelli per la classificazione. Il modello composto assume che ogni classe è un componente della miscela; ogni componente della miscela è un modello generativo che fornisce la probabilità di campionamento di un termine particolare per quel componente. Questi tipi di classificatori

sono chiamati anche classificatori generativi. Tre dei più famosi classificatori sono:

- *Naïve Bayes Classifier* (NB): è un classificatore probabilistico molto semplice basato sull'applicazione del teorema di Bayes. Il modello Naïve Bayes calcola la probabilità a posteriori di una classe e si basa sulla distribuzione delle parole nel documento. Il modello lavora con funzioni di estrazione che ignora la posizione della parola nel documento. La probabilità, secondo il teorema, che la caratteristica di un insieme appartiene a una particolare etichetta è data dall'equazione:

$$P(label|features) = \frac{P(label) * P(features|label)}{P(features)}$$

  dove *P(label)* è la probabilità a priori di una etichetta o la probabilità che una caratteristica casuale imposti l'etichetta; *P(features | label)* è la probabilità a priori che un dato set di funzionalità sia stato classificato come etichetta; *P(features)* è la probabilità a priori che si sia verificato un determinato set di funzionalità.

- *Bayesian Network* (BN): l'assunzione principale del classificatore BN è l'indipendenza delle caratteristiche; l'altra ipotesi estrema è di assumere che tutte le funzioni siano completamente dipendenti. Il modello di rete bayesiana è un grafo orientato aciclico in cui nodi rappresentano variabili casuali, e gli archi rappresentano le dipendenze condizionali. BN è considerato un modello completo per le variabili e le loro relazioni; pertanto, è specificata una completa distribuzione di probabilità congiunta su tutte le variabili. Nel Sentiment Analysis, la complessità di calcolo di BN è molto costosa, per questo, non è spesso utilizzato.

- *Maximum Entropy Classifier* (ME): questo classificatore (noto come classificatore esponenziale condizionale) converte un set etichettato di funzionalità in vettori che utilizzano la codifica. Questo vettore codificato è quindi utilizzato per calcolare i pesi per ogni caratteristica che possono poi essere combinati per determinare la probabile etichetta per un set di funzionalità. Questo classificatore è parametrizzato da un insieme *X{pesi}*, che è utilizzato per combinare le caratteristiche comuni che sono generate da un insieme di caratteristiche *X{codifiche}*. In particolare, la codifica mappa ogni coppia *{C(caratteristiche-set, etichetta)}* ad un vettore. La probabilità di ciascuna etichetta viene calcolata utilizzando la seguente equazione:

$$P(fs|label) = \frac{dotprod(weights,\ encode(fs, label))}{sum(dotprod(weights, encode(fs, l))forlinlabels)}$$

1.3.2. Approccio basato sul Lessico

Lo scopo dell'approccio Basato sul Lessico consiste nel trovare l'opinione espressa nel lessico per analizzarne il testo. Al fine di giungere allo scopo occorre utilizzare raccolte catalogate ed etichettate. Queste raccolte possono essere create direttamente dall'utente, ma ciò richiederebbe molto tempo e comunque per una questione di accuratezza è opportuno che siano affiancate da approcci automatizzati. Gli approcci automatizzati possono essere divisi in due metodi:

- Dictionary-based Approach;
- Corpus-based Approach.

I due metodi presentano caratteristiche comuni perché entrambi partono da un concetto chiave cioè, le parole e il loro modo di inserirsi all'interno di una frase o di un testo. Però; mentre il Corpus-based Approach una volta individuata

la lista di parole da cui ricavare le informazioni le ricerca in un ampio corpus con orientamenti specifici legati al contesto, il Dictionary-based Approach ricerca le parole all'interno di un dizionario dove ogni termine, oltre al significato, contiene anche informazioni relative agli aspetti ortografici, alla classe lessicale, eventuali flessioni, sinonimi e contrari, e fornisce anche esempi che illustrano come queste possono combinarsi tra loro e con altri elementi linguistici con cui possono co-occorrere[4].

Di seguito sono discussi i due metodi automatizzati nel dettaglio:

- *Dictionary-based Approach*: in questo approccio vengono selezionate le Opinion-words da cui è possibile ricavare informazioni significative e successivamente queste vengono ricercate all'interno di noti database a cui ad ogni voce è stato assegnato un punteggio. Questo processo iterativo si arresta nel momento in cui non sono rilevate le Opinion-words da analizzare. Questo approccio ha un limite significativo: ogni parola assume un significato specifico in base alla posizione o al contesto in cui viene espressa. Tali problematiche sono legate al problema del disambiguation.

- *Corpus-based Approach*: questo approccio aiuta a risolvere il problema di trovare Opinion-words con orientamenti specifici di contesto. L'utilizzo del Corpus-based Approach da solo non è così efficace come l'approccio basato sul dizionario, e questa inefficienza è dovuta alla difficoltà di avere un enorme Corpus in grado di coprire tutte le parole con le loro specifiche combinazioni. Questo approccio ha però un vantaggio: facendo uso di metodi statistici o semantici riesce a trovare il contesto e il dominio di alcuni Opinion-words.
    - *Statistical approach*: è un metodo che ricerca in un insieme di documenti indicizzati (come ad esempio l'intero web) le occorrenze delle Opinion-words: se la parola è contenuta maggiormente nei

[4] http://www.uniba.it/docenti/cardona-mario/attivita-didattica/Lapprocciolessicale.pdf

testi positivi allora la sua polarità sarà positiva; se si verifica più frequentemente nei testi negativi avrà polarità negativa; se ha frequenze uguali allora è neutra. Questo metodo si basa fondamentalmente sulla presenza di Opinion-words simili: se all'interno di uno stesso contesto due parole compaiono insieme, esse hanno la stessa polarità. Di conseguenza, la polarità di una parola sconosciuta può essere determinata calcolando la frequenza relativa di co-occorrenza con un'altra parola.

- *Semantic approach*: questo approccio fornisce direttamente i valori dei Sentiment e si basa su diversi principi per calcolare la somiglianza tra le parole. Questo approccio è molto diffuso in ambito del SA poiché permette di costruire un modello lessicale che permette di ricavare iterativamente la polarità dei Sentiment. Questo approccio è alla base dei maggiori lessici presenti online come ad esempio WordNet, SentiWordNet.

L'approccio semantico è utilizzato in molte applicazioni per costruire un modello lessicale per la descrizione di verbi, sostantivi, aggettivi e avverbi da utilizzare in Sentiment Analysis. [6] [2]

Capitolo 2

Analisi Testuale

L'analisi testuale è il processo che sta alla base del SA. In questo capitolo saranno presentati due database semantico-lessicale: WordNet e SentiWordNet. L'utilizzo di questi due database renderà possibile l'analisi testuale basata sul metodo Lexicon Based-Approach. Sarà presentato inoltre, la libreria NLTK, una piattaforma leader per la creazione di programmi in Python per estrarre automaticamente informazioni da testi o da altri media in lingua naturale.

2.1. La nascita di WordNet

WordNet [3] nasce nel 1985 come risultato di un progetto su cui hanno partecipato linguisti e psicologi dell'Università di Princeton, New Jersey USA. L'idea iniziale era fornire un'ulteriore risorsa on-line, rispetto ad una semplice ricerca di tipo alfabetico, ma con il tempo si è giunti ad un vero e proprio dizionario basato sui principi della psicolinguistica. Progettato inizialmente in lingua inglese, è successivamente tradotto in molteplici lingue e distribuito gratuitamente dal sito dell'università di Princeton con libera licenza di utilizzo previa citazione degli autori[5]; è disponibile inoltre un tool online, liberamente utilizzabile per testarne le funzionalità[6].

[5] http://wordnet.princeton.edu/
[6] http://wordnetweb.princeton.edu/perl/webwn

Ciò che contraddistingue WordNet dal classico dizionario è che esso divide il lessico in quattro categorie sintattiche: *sostantivi*, *verbi*, *aggettivi* ed *avverbi*; ed ogni categoria è ulteriormente raggruppata in insiemi di sinonimi detti *synset*, termine derivante dalla contrazione di "synonym set". Ogni insieme di sinonimi si riferisce ad un particolare concetto ed è posto in relazione con altri synsets tramite relazioni lessicali. Concettualmente può essere definito come un enorme grafo contenente diversi termini (i nodi del grafo) collegati fra loro attraverso varie relazioni (gli archi tra i nodi).

2.1.1. La matrice lessicale

In WordNet le informazioni sono memorizzate in base al loro significato e alle loro categorie sintattiche e sono legate tra loro tramite diversi tipi di relazioni. WordNet divide il significato di una parola in due concetti: *Word Form* che sta ad indicare la forma scritta e *Word Meaning* si riferisce al concetto espresso da tale parola. Il punto di inizio della classificazione delle parole sono le relazioni che intercorrono fra il lemma e il significato. Le associazioni tra lemma e synset possono essere descritti nella matrice lessicale.

Word Meanings	Word Forms				
	F_1	F_2	F_3	. . .	F_n
M_1	$E_{1,1}$	$E_{1,2}$			
M_2		$E_{2,2}$			
M_3			$E_{3,3}$		
⋮				⋱	
M_m					$E_{m,n}$

Figura 4: Matrice lessicale

Come è possibile osservare dalla Figura 4, le righe della matrice indicano i synset; mentre nelle colonne si indicano i lemmi. Quello che è possibile osservare inoltre dalla matrice sono le varie relazioni lessicali che prendono il nome di sinonimia o polisemia [3].

Concettualmente si parla di:

- **sinonimia**: quando in una riga ci sono più entry ciò implica che i due termini associati si riferiscono allo stesso significato come nel caso della prima riga: $E_{1,1}$ e $E_{1,2}$;
- **polisemia**: quando in una colonna ci sono più entry ciò implica che lo stesso termine è associato a più significati come nel caso della seconda colonna: $E_{1,2}$ e $E_{2,2.}$

2.2. Le relazioni di WordNet

In WordNet esistono due tipi di relazioni: le relazioni semantiche e quelle lessicali. Mentre le prime sussistono fra significati, le seconde sussistono fra parole. WordNet è quindi una rete di relazioni semantiche e lessicali, ognuna delle quali è rappresentata da un puntatore. La regola generale che i puntatori devono seguire prevede che non possano esistere relazioni fra due diverse categorie sintattiche, a meno di casi eccezionali.

2.2.1. Relazioni Lessicali

Le relazioni lessicali coinvolgono due o più forme di parola, i così detti lemmi e non comprendono dunque le relazioni tra synset. Le relazioni lessicali spiegano i fenomeni di:

- **Sinonimia**: proprietà di un concetto di avere due o più parole in grado di esprimerlo;
- **Antonimia**: proprietà di una parola di avere un significato opposto
  [A opposto di B non implica che A = not(B)];
- **Polisemia**: proprietà di una parola di avere due o più significati.

Questi concetti torneranno molto utili successivamente, poiché, in fase di analisi la sinonimia e la polisemia inducono al problema del disambiguation.

2.2.2. Relazioni Semantiche

Le relazioni semantiche, coinvolgono sempre due concetti quindi due significati (due synset) e variano in funzione del tipo di parola e includono[7]:

- per i sostantivi:
  - **iperonimia** (hyperonyms): Y è un iperonimo di X se ogni X è "una specie di" Y;
  - **iponimia** (hyponyms): Y è un iponimo di X se ogni Y è (una specie di) X;
  - **coordinazione**: Y è un termine coordinato di X se X e Y hanno un iperonimo in comune;
  - **olonimia** (holonym): Y è un olonimo di X se X è parte Y;
  - **meronimia** (meronym): Y è un meronimo di X se Y è parte X;
- per i verbi:
  - **iperonimia** (hypernyms): il verbo Y è un iperonimo del verbo X se l'attività X è (una specie di) Y (come viaggio rispetto a movimento);
  - **troponimia** (troponyms): il verbo Y è un troponimo del verbo X se nel fare l'attività Y si fa anche la X (come mormorare rispetto a parlare);
  - **implicazione** (entailment): il verbo Y è un'implicazione del verbo X se nel fare X uno deve per forza fare Y (come russare rispetto a dormire);
  - **coordinazione**: Y è un termine coordinato di X se X e Y hanno un iperonimo in comune.

[7] https://it.wikipedia.org/wiki/WordNet

- gli aggettivi sono classificati come:
  - **aggettivi descrittivi**: associa un valore a un attributo del nome cui esso è associato;
  - **aggettivi relazionali**: sono aggettivi che derivano dai nomi;
  - **aggettivi Reference-Modifying** sono in un insieme piuttosto piccolo e costituiscono una classe chiusa che viene mantenuta separata dagli altri aggettivi.
- gli **avverbi** seguono la classificazione dell'aggettivo da cui derivano

2.3. SentiWordNet

Sviluppato da Andrea Esuli e Fabrizio Sebastiani SentiWordNet [1] è un database, una risorsa lessicale, che si basa su WordNet alla versione 3.0.

L'aggiunta apportata a SentiWordNet consiste nel fatto che esso assegna ad ogni synset, che come descritto nei paragrafi precedenti indica un insieme di sinonimi, tre punteggi sentimentali: *positività*, *negatività* e *oggettività*. L'assunzione di fondo che permette di assegnare i punteggi ai synset e non ai termini è che ogni termine ha diversi significati le quali conducono a opinioni differenti.

La costruzione di SentiWordNet parte dalla classificazione dei synset. A tale scopo sono stati predisposti vari classificatori ternari, differenti per il training set utilizzato per il loro allenamento. Ogni synset è stato quindi sottoposto ad analisi da tutti i classificatori addestrati e i punteggi ottenuti sono proporzionali ai risultati dei classificatori[8]. Ogni synset ha un punteggio compreso tra 0.0 e 1.0.

La figura 5, mostra come SentiWordNet rappresenta e classifica un termine.

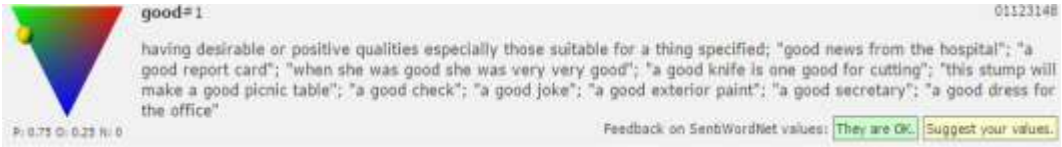

Figura 5: Termine "good" in SentiWordNet

[8] http://www.di.unipi.it/~cappelli/seminari/baccianella.pdf

Analizzando la figura [9]è possibile notare un triangolo i cui angoli rappresentano le tre etichette (in alto a sinistra la classe positiva, in alto a destra la classe negativa, in basso la classe oggettiva). Il cerchio di colore giallo, permette di avere un'idea visiva del punto dello spazio in cui il termine è collocato. In basso è possibile notare il punteggio positivo, oggettivo e negativo associato a tale termine; ed infine, sulla destra è possibile leggere le caratteristiche associato al termine good#1. Uno sguardo al sito permette di vedere che il termine "good" come aggettivo ha ben 21 utilizzi diversi; come nome 4 utilizzi diversi; come avverbio 2. Tutto questo deriva dal fatto che la lingua umana, utilizza lo stesso termine in svariati modi diversi. Questo, come sarà discusso successivamente al Capitolo 5, rappresenta uno dei motivi per cui, ancora oggi, non è stato implementato un programma informatico in grado di analizzare e comprendere esattamente il linguaggio umano.

2.4. Word Sense Disambiguation

La Word Sense Disambiguation (WSD) è l'operazione con la quale si precisa il significato di una parola o di un insieme di parole, contenute dunque in frasi, che denotano significati diversi a seconda dei contesti e che quindi risultano ambigue. Questo problema è forse uno dei più ostici degli studi sull'intelligenza artificiale.
Il linguaggio umano risulta essere molto ambiguo poiché molte parole possono essere interpretate in diversi modi a seconda del contesto in cui si verificano.
Per esempio, si consideri il seguente esempio:

(A) *Ho acquistato un <u>rombo</u> fresco al mercato*
(B) *Il <u>rombo</u> è una figura geometrica*

[9] http://sentiwordnet.isti.cnr.it/search.php?q=good

Sebbene per un essere umano sia ovvio che la prima frase si riferisce al significato di pesce e al secondo ad una figura geometrica, sviluppare algoritmi in grado di replicare questa capacità umana è tipicamente difficile.
I metodi algoritmici attualmente proposti per aggirare in parte il problema, sono stati riadattati per poter sfruttare WordNet e le funzioni della libreria NLTK: l'**algoritmo di Lesk** è uno di questi algoritmi. Questo algoritmo, inventato da Michael E. Lesk nel 1986, si basa sul presupposto che le parole in una data "zona" del testo tenderanno a condividere un argomento comune. La semplificazione dell'algoritmo, adattata per usare WordNet consiste nel confrontare la definizione del dizionario di una parola ambigua con i termini contenuti nel suo quartiere. L'algoritmo esegue una fase di analisi testuale in attesa di un termine che potrebbe essere ambiguo. Non appena trovato, esamina tra tutti i termini quelli che sono entrambi quartiere di quella parola e se la definizione di quel senso combacia. In caso di più termini, occorre scegliere quello con il maggior numero di conteggio. [5] [7]

2.5. Libreria NLTK

NLTK, acronimo di Natural Language Toolkit, è un progetto Open Source iniziato nel 2002 per mano di Steve Bird, Edward Loper, Ewan Klein. NLTK è un insieme di moduli Python che possono essere importati all'interno di altri programmi. I moduli presentati in NLTK permettono un facile sviluppo di algoritmi per il trattamento del linguaggio naturale basati su metodi simbolici, statistici, di apprendimento automatico, etc. NLTK viene distribuito insieme a diversi corpora annotati molto utilizzati all'interno della comunità NLP.
Questa libreria fornisce interfacce facili da usare per più di 50 corpi per le risorse lessicali, quali WordNet, unitamente a una suite di librerie di elaborazione del testo per la classificazione, tokenization, stemming, tagging, parsing, etc.
Python è particolarmente adatto allo sviluppo di sistemi per il trattamento di dati testuali poiché molte funzioni di base sono già presenti nelle librerie standard del

linguaggio. NLTK potenzia ed estende le funzioni di analisi e trattamento dei testi di Python, così da rendere facile anche l'implementazione di moduli di analisi linguistica più complessi. NLTK ha inoltre lo scopo di sostenere la ricerca e l'insegnamento in PNL o settori strettamente connessi, tra cui: scienze cognitive, intelligenza artificiale, information retriva e apprendimento automatico.

NLTK è stato usato con successo come strumento didattico, come uno strumento di studio individuale e come piattaforma per la prototipazione e costruzione di sistemi di ricerca. [4] [7] [8] [9] [10]

Parte II

Progetto e valutazione

Capitolo 3

Naïve Bayes e Dictionary Based Approach

Per approfondire le conoscenze acquisite durante il caso di studio sono state raccolte 350 recensioni al fine di testare due degli algoritmi esposti nel Capitolo 1: il primo appartenente agli approcci Machine Learning e il secondo appartenente al Lexicon Based-Approach. Lo scopo dei due algoritmi è classificare un tweet dato da input e restituire come output la sentenza positiva o negativa.

3.1. Raccolta dei dati

Per riuscire a testare perfettamente l'efficienza degli algoritmi, sono stati selezionati 350 recensioni dai siti più frequentati dagli utenti del web come: Amazon, TripAdivisor, MyMovie, Facebook e YouTube. La scelta di questi siti è dovuta anche al fatto che generalmente gli utenti si ritrovano ad esprimere liberamente le proprie opinioni dando spazio ai propri sentimenti e ai propri stati d'animo senza essere influenzati da nessun preconcetto. La raccolta manuale dei commenti ha permesso di preferire recensioni molto più significativi in termini di espressione di linguaggio e di aggettivi utilizzati. I prodotti recensiti spaziano tra le varie categorie: musica, film e tv; elettronica e informatica; casa, giardino e fai da te; abbigliamento, scarpe e gioielli; hotel; ristoranti; casa vacanza; thriller e film d'azione. Questo set di dati comprende esclusivamente recensioni positive e

recensioni negative. Ogni recensione, oltre ad avere espressioni e frasi diverse, spazia da un massimo di 90 parole ad un minimo di 2 in modo da verificare, in fase di test, se il numero di parole contenute in documento possa influire nel risultato finale.

3.2. Naïve Bayes

Un'applicazione creata con algoritmi di machine learning necessita di una fase di addestramento. Il classificatore è stato addestrato, di volta in volta, con tweet classificati manualmente differenziandoli e contraddistinguendoli in tweet positivi e tweet negativi. Dato il numero consistente di tweet raccolti, il classificatore, è stato allenato di volta in volta con tweet diversi effettuando i test su quelli rimanenti. Lo schema utilizzato per la classificazione e i test è una matrice 4x4 come quella mostrata di seguito.

TRAIN / TEST	AMAZON [50:50]	TRIPADIVISOR [50:50]	MYMOVIES [50:50]	VARIE [25:25]
AMAZON [50:50]				
TRIPADIVISOR [50:50]				
MY MOVIES [50:50]				
VARIE [25:25]				

Figura 6: Matrice 4x4: schema classificazione e test

Le righe e le colonne della matrice sono state etichettate con nomi e valori: i nomi indicano i siti web da dove sono stati raccolti i dati, mentre i valori indicano rispettivamente il numero di tweet positivi e negativi raccolti. Per ogni sito web, sono state raccolte 50 recensioni positive e 50 recensioni negative, per un totale di cento recensioni per ogni sito web, ad esclusione di "Varie" che ne contiene 50 in

tutto. Le colonne della matrice indicano il train set utilizzato per l'allenamento del classificatore; mentre le righe i test.

Il classificatore bayesiano funziona estraendo dalle frasi del train le parole che compaiono con maggiore frequenza. Tale classificatore, utilizza la probabilità a priori di ogni etichetta, data dalla frequenza di ciascuna etichetta nel training set, e questo contributo viene dato in modo indipendente. Nel caso, oggetto di studio, la frequenza di ciascuna etichetta è la stessa per positiva e negativa.

Di seguito, una figura a scopo di esempio, contenente le caratteristiche più informative per il classificatore allenato con le recensioni di Varie.

```
Most Informative Features
       contains(product) = True           positi : negati =      2.3 : 1.0
          contains(nice) = True           positi : negati =      2.3 : 1.0
       contains(service) = True           negati : positi =      2.3 : 1.0
       contains(quality) = True           positi : negati =      1.7 : 1.0
          contains(good) = False          negati : positi =      1.5 : 1.0
     contains(beautiful) = False          negati : positi =      1.2 : 1.0
           contains(bad) = False          positi : negati =      1.1 : 1.0
           contains(she) = False          negati : positi =      1.1 : 1.0
         contains(works) = False          negati : positi =      1.1 : 1.0
          contains(nice) = False          negati : positi =      1.1 : 1.0
```

Figura 7: Most informative features (trainVarie)

La figura 7 mostra le Most Informative Features. Ad esempio: se il tweet in ingresso contiene la parola "product" , la razione positivi è 2.3; mentre se il tweet in ingresso non contiene la parola "good" la razione negativa è di 1.5.

3.3. Dictionary Based-Approach

Un'applicazione basata sul Dictionary Based-Approach necessita dell'ausilio di un dizionario da cui ricavare il peso associato alle parole. Nel caso di studio sono stati presentati due database lessicali: WordNet e SentiWordNet, due risorse lessicali utilizzati con scopi diversi. Nello specifico il primo è stato utilizzato per estrapolare la posizione della parola contenuta all'interno della frase, mentre il secondo per assegnare il peso alle parole, e quindi all'intero tweet.

WordNet e SentiWordNet identificano quattro categorie di termini da cui ricavare le informazioni: nomi, verbi, aggettivi ed avverbi; tutto ciò che non rientra in una di queste categorie sono considerati inutili al fine di ricavare la sentenza sentimentale. Grazie alle risorse messe a disposizione della libreria NLTK è possibile procedere con un'analisi grammaticale del testo in modo da etichettare ogni termine in base alla propria funzione e decidere quindi quali parole sono significative e quali invece sono di intralcio. La figura 8 riportata di seguito riassume i concetti chiave che partendo dal tweet di input permette di giungere alla sentenza positiva, negativa o oggettiva.

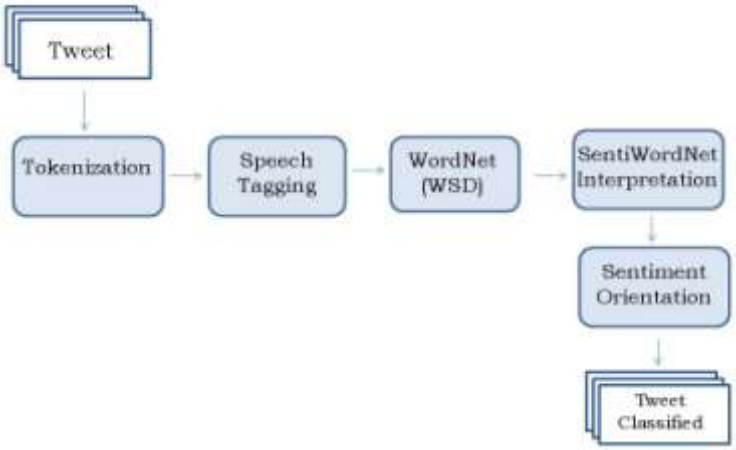

Figura 8: Sentiment Classification Phases

La parte centrale della figura descrive la fase di preprocessing del testo. Ciò consiste nel ripulire il testo da tutti quegli elementi considerati di intoppo al fine di analisi. La fase di preprocessing, dato un tweet di input, prevede:

- la fase di **tokenization,** che consiste nel dividere la sequenza di caratteri in unità minime di analisi dette "token";
- la fase **part-of-speech tagging o POS tagging** che aggiunge, alla fase di tokenization l'etichetta corrispondente al termine; ad esempio: [NN] per i sostantivi, [VB] per i verbi, [JJ] per gli aggettivi, [RB] per gli avverbi;
- la fase **Word Sense Disambiguation**, discusso nel paragrafo 2.4, è la fase che ricerca nel testo parole considerate ambigue

Una volta ultimata questa fase di analisi testuali si giunge alla classificazione del tweet con sentenza positiva, negativa o oggettiva.

Capitolo 4

Valutazione efficacia dei due metodi

In questo capitolo verranno mostrati i risultati dei test effettuati sui dati raccolti sottoposti ai due metodi algoritmici discussi nel Capitolo 3. Inoltre, servendomi dei risultati ottenuti, sarà possibile dare una motivazione su quale algoritmo ha funzionato meglio e in quale circostanza.

4.1. Risultati dei test su Machine Learning

Come già discusso nel Capitolo 3, la fase di addestramento del classificatore è avvenuto seguendo uno schema matriciale 4x4 ed effettuando i test sui rimanenti dati. Trattandosi di una matrice, si è prestata molta attenzione nell'effettuare i test su quelli che sono i dati della diagonale evitando che i dati di allenamento e di test coincidessero.

TRAIN / TEST	AMAZON [50:50]	TRIPADIVISOR [50:50]	MYMOVIES [50:50]	VARIE [25:25]
AMAZON [50:50]				
TRIPADIVISOR [50:50]				
MY MOVIES [50:50]				
VARIE [25:25]				

Figura 9: Matrice 4x4 (Test)

Per svolgere questi test, ho utilizzato la metà dei dati per allenare il classificatore e la restante per l'analisi. Poiché, sulla diagonale l'addestramento e l'analisi dei dati avvenivano sulle medesime informazioni con le medesime caratteristiche ho deciso di svolgere svariati test cercando di variare le sottocategorie a cui appartenevano i dati. Nello specifico:

- addestramento sui primi N/2 dati, e test sui restanti;
- addestramento sugli ultimi N/2 dati, e test sui restanti;
- addestramento sugli N/2 dati centrali, e test sui restanti;
- addestramento sugli N/2 dati scelti casualmente tra ogni N/2 sottocategoria, e test sui restanti.

Ciò ha portato ai seguenti risultati:

TRAIN / TEST	AMAZON [50:50]		TRIPADIVISOR [50:50]		MY MOVIES [50:50]		VARIE [25:25]	
AMAZON [50:50]	Pos	Neg						
	56%	72%						
	Tot: 64%							
TRIPADIVISOR [50:50]			Pos	Neg				
			92%	64%				
			Tot: 78%					
MY MOVIES [50:50]					Pos	Neg		
					64%	68%		
					Tot: 66%			
VARIE [25:25]							Pos	Neg
							38,46%	92.30%
							Tot: 65,38%	

Figura 10: Matrice 4x4 valori diagonale addestramento primi N/2 dati

TRAIN / TEST	AMAZON [50:50]		TRIPADIVISOR [50:50]		MY MOVIES [50:50]		VARIE [25:25]	
AMAZON [50:50]	Pos	Neg						
	96%	52%						
	Tot: 74%							
TRIPADIVISOR [50:50]			Pos	Neg				
			52%	100%				
			Tot: 76%					
MY MOVIES [50:50]					Pos	Neg		
					56%	80%		
					Tot: 68%			
VARIE [25:25]							Pos	Neg
							41,67%	83,33%
							Tot: 62,5%	

Figura 11: Matrice 4x4 valori diagonale addestramento ultimi N/2 dati

TRAIN / TEST	AMAZON [50:50]		TRIPADIVISOR [50:50]		MY MOVIES [50:50]		VARIE [25:25]	
AMAZON [50:50]	Pos	Neg						
	84%	48%						
	Tot: 66%							
TRIPADIVISOR [50:50]			Pos	Neg				
			40%	100%				
			Tot: 70%					
MY MOVIES [50:50]					Pos	Neg		
					52%	88%		
					Tot: 70%			
VARIE [25:25]							Pos	Neg
							48%	70%
							Tot: 59%	

Figura 12: Matrice 4x4 valori diagonale addestramento sugli N/2 dati centrali

TRAIN / TEST	AMAZON [50:50]		TRIPADIVISOR [50:50]		MY MOVIES [50:50]		VARIE [25:25]	
AMAZON [50:50]	Pos	Neg						
	56%	100%						
	Tot: 78%							
TRIPADIVISOR [50:50]			Pos	Neg				
			88%	92%				
			Tot: 90%					
MY MOVIES [50:50]					Pos	Neg		
					92%	88%		
					Tot: 90%			
VARIE [25:25]							Pos	Neg
							75%	83,33%
							Tot: 79,17%	

Figura 13: Matrice 4x4 valori diagonali addestramento sugli N/2 dati scelti a casualmente tra ogni N/2 sottocategoria

Come è possibile osservare, le quattro tabelle, riportano valori diversi. Questa diversità è dovuta alla frequenza dei termini che compongono le frasi utilizzate per allenare il classificatore. Ogni cella della matrice mostra la percentuale di correttezza dei tweet positivi, dei tweet negativi e la percentuale totale di correttezza di entrambi i tweet. Analizzando tali valori, è possibile osservare, che i risultati dei test hanno subito delle variazioni significative.

Riassumo in un istogramma i valori delle quattro tabelle per avere una visione d'insieme dell'andamento dei dati.

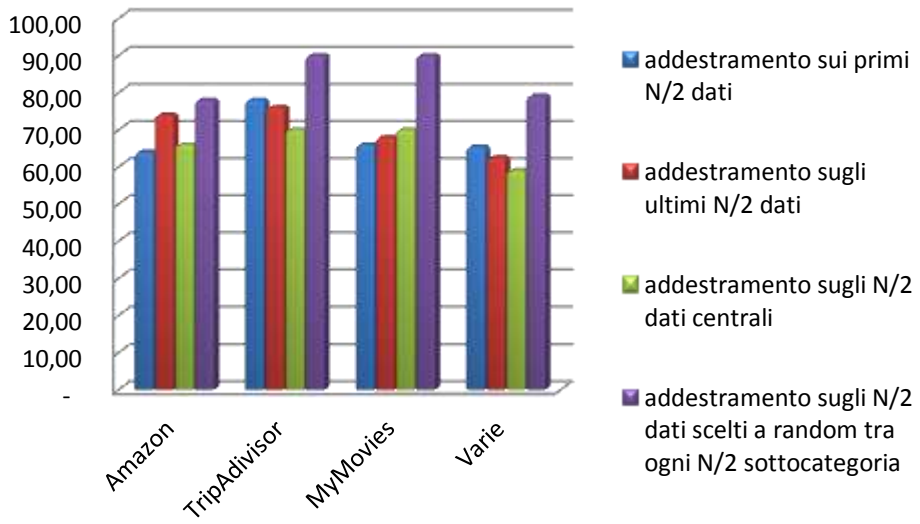

Figura 14: Istogramma dei quattro casi

L'istogramma mostra sull'asse delle ordinate la percentuale di correttezza, mentre sull'asse delle ascisse viene riportato il nome del sito che raggruppa l'andamento dei risultati al variare dell'addestramento del classificatore identificato dal colore nella legenda. L'andamento di questi valori è dovuto al fatto che, i termini che compongono le rispettive frasi si combinano in modo diverso in base al contesto di riferimento. Stessi termini infatti, possono essere utilizzati indistintamente sia per le frasi positive sia per le frasi negative. Ad esempio, date queste due recensioni, è possibile osservare che i termini *staff, very, return, hotel* sono utilizzati sia per la recensione positiva che per quella negativa.

I am very disappointed in the hotel. The staff was unfriendly and arrogant and the service was slow as much as the Wi-Fi connection. Next time in London I would never return there. [negative]

Fantastic Hotel. We enjoyed the large rooms. The staff was very, very helpful. Love this Hotel and I'm looking forward to return. [positive]

L'allenamento sui restanti casi ha condotto ai seguenti risultati:

TRAIN / TEST	AMAZON [50:50]		TRIPADIVISOR [50:50]	MY MOVIES [50:50]	VARIE [25:25]
AMAZON [50:50]					
TRIPADIVISOR [50:50]	Pos	Neg			
	96%	46%			
	Tot: 71%				
MY MOVIES [50:50]	Pos	Neg			
	86%	32%			
	Tot: 59%				
VARIE [25:25]	Pos	Neg			
	96%	16%			
	Tot: 56%				

Figura 15: Matrice 4x4, train Amazon

TRAIN / TEST	AMAZON [50:50]	TRIPADIVISOR [50:50]		MY MOVIES [50:50]	VARIE [25:25]
AMAZON [50:50]		Pos	Neg		
		58%	94%		
		Tot: 76%			
TRIPADIVISOR [50:50]					
MY MOVIES [50:50]		Pos	Neg		
		52%	96%		
		Tot: 74%			
VARIE [25:25]		Pos	Neg		
		20%	88%		
		Tot: 54%			

Figura 16: Matrice 4x, train TripAdivisor

TRAIN / TEST	AMAZON [50:50]	TRIPADIVISOR [50:50]	MY MOVIES [50:50]	VARIE [25:25]
AMAZON [50:50]			Pos: 80%, Neg: 84%, Tot: 82%	
TRIPADIVISOR [50:50]			Pos: 78%, Neg: 86%, Tot: 82%	
MY MOVIES [50:50]				
VARIE [25:25]			Pos: 24%, Neg: 88%, Tot: 56%	

Figura 17: Matrice 4x, train MyMovies

TRAIN / TEST	AMAZON [50:50]	TRIPADIVISOR [50:50]	MY MOVIES [50:50]	VARIE [25:25]
AMAZON [50:50]				Pos: 48%, Neg: 78%, Tot: 63%
TRIPADIVISOR [50:50]				Pos: 38%, Neg: 90%, Tot: 64%
MY MOVIES [50:50]				Pos: 28%, Neg: 84%, Tot: 56%
VARIE [25:25]				

Figura 18: Matrice 4x4, train Varie

Anche per questi test verrà riportato un istogramma allo scopo di riassumere i risultati dell'analisi svolti allenando di volta in volta il classificatore con l'intero set di dati raccolti dallo stesso sito e svolgendo i test sui dati rimanenti.

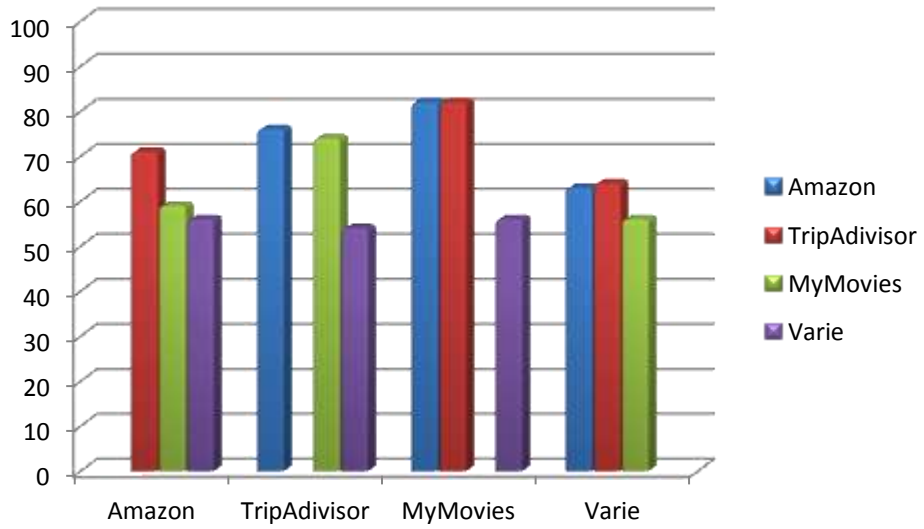

Figura 19: Istogramma riassuntivo allenando il classificatore con un solo set di dati per volta

L'istogramma mostra sull'asse delle ordinate la percentuale di correttezza, mentre sull'asse delle ascisse viene riportato il nome del sito in cui sono state raccolte le recensioni utilizzate per allenare il classificatore. Il colore delle rispettive barre è indicato nella legenda affianco. In questo diagramma, così come nelle figure precedente, non viene riportato l'analisi dei test sulla diagonale poiché già discussi.

Osservando il grafico è possibile notare che non si è giunti ad una correttezza massima ma comunque maggiore del 50%.

4.2. Risultati dei test su Dictionary Based-Approach

La tecnica Dictionary Based-Approach basata sui database lessicali WordNet e SentiWordNet con l'ausilio della libreria NLTK, ha permesso di effettuare due diversi tipi di test: con e senza disambiguation (discusso al Capitolo 2). Poiché, il database SentiWordNet restituisce, per ogni tweet di input, una sentenza positiva, negativa ed oggettiva; e poiché i dati in ingresso sono solo positivi e negativi, eventuali sentenze oggettive verranno considerate erronee. Il risultato dei test è stato raccolto in una tabella, che come per il paragrafo precedente, riporta per ogni cella la percentuale di correttezza dei tweet positivi, dei tweet negativi e la percentuale totale di correttezza di entrambi i tweet. Qui di seguito le due tabelle e il relativo istogramma riassuntivo dei valori:

	AMAZON [50:50]		TRIPADIVISOR [50:50]		MY MOVIES [50:50]		VARIE [25:25]	
Senza Disambiguation	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg
	86%	58%	84%	56%	90%	38%	76%	48%
	Tot: 72%		**Tot: 70%**		**Tot: 64%**		**Tot: 62%**	
Con Disambiguation								

Figura 20: Tabella risultati Senza Disambiguation

	AMAZON [50:50]		TRIPADIVISOR [50:50]		MY MOVIES [50:50]		VARIE [25:25]	
Senza Disambiguation								
Con Disambiguation	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg
	86%	66%	72%	68%	76%	54%	73%	53%
	Tot: 76%		**Tot: 70%**		**Tot: 65%**		**Tot: 63%**	

Figura 21: Tabella risultati Con Disambiguation

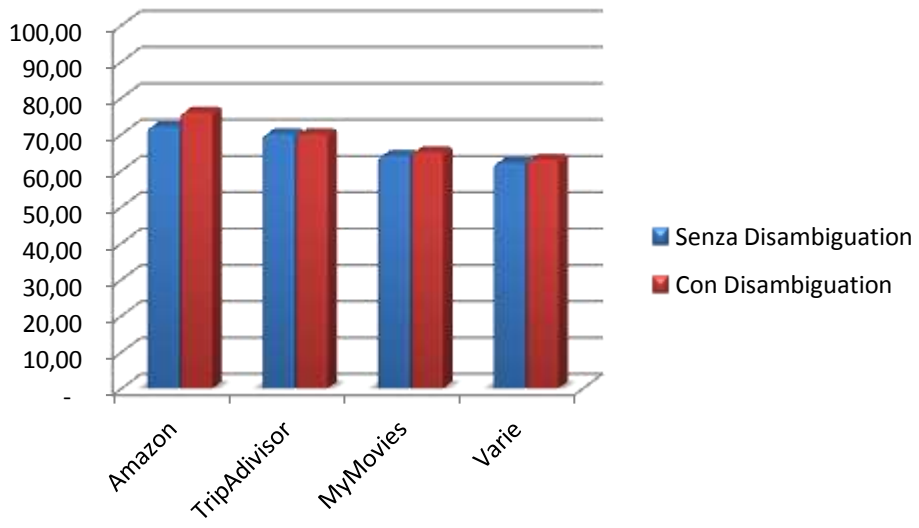

Figura 22: Istogramma riassuntivo Dictionary Based-Approach

Come è possibile notare dall'istogramma, l'andamento dei dati con e senza disambiguation ha condotto a dei cambiamenti lievi. Questo è dovuto fondamentalmente ad una struttura algoritmica molto semplice e/o dalla poca presenza di termini ambigui.

4.3. Valutazione finale test

Nei precedenti paragrafi sono state riportate le percentuali di correttezza dei test effettuati con i metodi Naïve Bayes e Dictionary Based-Approach.

Secondo il mio punto di vista migliori risultati sono stati ottenuti con la tecnica del Machine Learning poiché, avendo comunque a disposizione un set di dati molto ristretto per il train e per i test si è giunti ad una percentuale di correttezza del 70% contro il 67% raggiunto dal Dictionary Based-Approach il quale comunque ha a disposizione (secondo le ultime stime del 2012) 155,287 parole organizzate in 117,659 synset per un totale di 206,941 coppie di parole di senso compiuto.

Riporto le tabelle complete, analizzate in questo capitolo per una visione d'insieme dei dati ad esclusione dei valori contenuti sulla diagonale, per i quali si rimanda alle figure successive.

TRAIN / TEST	AMAZON [50:50]		TRIPADIVISOR [50:50]		MY MOVIES [50:50]		VARIE [25:25]	
AMAZON [50:50]			Pos	Neg	Pos	Neg	Pos	Neg
			58%	94%	80%	84%	48%	78%
			Tot: 76%		Tot: 82%		Tot: 63%	
TRIPADIVISOR [50:50]	Pos	Neg			Pos	Neg	Pos	Neg
	96%	46%			78%	86%	38%	90%
	Tot: 71%				Tot: 82%		Tot: 64%	
MY MOVIES [50:50]	Pos	Neg	Pos	Neg			Pos	Neg
	86%	32%	52%	96%			28%	84%
	Tot: 59%		Tot: 74%				Tot: 56%	
VARIE [25:25]	Pos	Neg	Pos	Neg	Pos	Neg		
	96%	16%	20%	88%	24%	88%		
	Tot: 56%		Tot: 54%		Tot: 56%			

Figura 23: Matrice Naïve Bayes

TRAIN / TEST	AMAZON [50:50]		TRIPADIVISOR [50:50]		MY MOVIES [50:50]		VARIE [25:25]	
AMAZON [50:50]	Pos	Neg						
	56%	72%						
	Tot: 64%							
TRIPADIVISOR [50:50]			Pos	Neg				
			92%	64%				
			Tot: 78%					
MY MOVIES [50:50]					Pos	Neg		
					64%	68%		
					Tot: 66%			
VARIE [25:25]							Pos	Neg
							38,46%	92.30%
							Tot: 65,38%	

Figura 24: Matrice 4x4 valori diagonale addestramento primi N/2 dati

TRAIN / TEST	AMAZON [50:50]		TRIPADIVISOR [50:50]		MY MOVIES [50:50]		VARIE [25:25]	
AMAZON [50:50]	Pos	Neg						
	96%	52%						
	Tot: 74%							
TRIPADIVISOR [50:50]			Pos	Neg				
			52%	100%				
			Tot: 76%					
MY MOVIES [50:50]					Pos	Neg		
					56%	80%		
					Tot: 68%			
VARIE [25:25]							Pos	Neg
							41,67%	83,33%
							Tot: 62,5%	

Figura 25: Matrice 4x4 valori diagonale addestramento ultimi N/2 dati

TRAIN / TEST	AMAZON [50:50]		TRIPADIVISOR [50:50]		MY MOVIES [50:50]		VARIE [25:25]	
AMAZON [50:50]	Pos	Neg						
	84%	48%						
	Tot: 66%							
TRIPADIVISOR [50:50]			Pos	Neg				
			40%	100%				
			Tot: 70%					
MY MOVIES [50:50]					Pos	Neg		
					52%	88%		
					Tot: 70%			
VARIE [25:25]							Pos	Neg
							48%	70%
							Tot: 59%	

Figura 26: Matrice 4x4 valori diagonale addestramento sugli N/2 dati centrali

TRAIN / TEST	AMAZON [50:50]		TRIPADIVISOR [50:50]		MY MOVIES [50:50]		VARIE [25:25]	
AMAZON [50:50]	Pos	Neg						
	56%	100%						
	Tot: 78%							
TRIPADIVISOR [50:50]			Pos	Neg				
			88%	92%				
			Tot: 90%					
MY MOVIES [50:50]					Pos	Neg		
					92%	88%		
					Tot: 90%			
VARIE [25:25]							Pos	Neg
							75%	83,33%
							Tot: 79,17%	

Figura 27: Matrice 4x4 valori diagonali addestramento sugli N/2 dati scelti a casualmente tra ogni N/2 sottocategoria

	AMAZON [50:50]		TRIPADIVISOR [50:50]		MY MOVIES [50:50]		VARIE [25:25]	
Senza Disambiguation	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg
	86%	58%	84%	56%	90%	38%	76%	48%
	Tot: 72%		Tot: 70%		Tot: 64%		Tot: 62%	
Con Disambiguation	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg
	86%	66%	72%	68%	76%	54%	73%	53%
	Tot: 76%		Tot: 70%		Tot: 65%		Tot: 63%	

Figura 28: Matrice Dictionary Based-Approach

Capitolo 5

Considerazioni finali

Gli obiettivi di questa tesi erano quelli di studiare ed approfondire le conoscenze sulle tecniche di Sentiment Analysis e valutare, quale fosse il miglior metodo algoritmico in base ai risultati ottenuti.

L'analisi condotta sulle recensioni prese dai vari siti web ha permesso di giungere alla conclusione, nonostante le varie difficoltà dovute alla vastità del lessico, che a livello aziendale per sondare il gradimento dei prodotti immessi sul mercato sarebbe più opportuno fare uso di un analizzatore lessicale specifico per un dato argomento in modo da omettere i termini che potrebbero essere specifiche di quel settore.

Basandomi sugli studi condotti ho potuto notare che nomi di alcune locande o ristoranti, quali ad esempio "*Sweet Sara*"; "*Larry delicious food*" ; contribuivano comunque alla sentenza finale della recensione. Rimanendo nell'ambito delle recensioni del sito TripAdivisor riporto alcune recensioni che avevano un esplicito riferimento alle Stelle Michelin:

"My partner and I stayed here for dinner in June and I was really shocked at how bad it was. If you consider that this is supposed to be Michelin starred restaurant the food was terrible. We had the tasting menu. They had an overly salty and sometimes overly acidic sauce. I've never had such a bad meal for a long time, and I've never had such a bad meal at a Michelin-starred restaurant."

"Restaurant menu could be very close to be a Michelin star or even better is some areas. Wine selection is well balanced with occasional surprises of regional interest. I like the butler style room attendances, where magical things are done."

L'analisi svolta dal punto di vista umano fa emergere chiaramente un riferimento esplicito a quello che è definito il maggior riferimento mondiale per la valutazione della qualità dei ristoranti e alberghi a livello nazionale e internazionale.
Altri comportamenti strani del sistema si notavano ogniqualvolta, il contenuto di un tweet riguardava un paragone, una similitudine, frasi fatte o metaforiche. Riporto una recensione dalla collezione MyMovie:

"I'm sorry to say this but I didn't enjoy this movie at all. It was just too boring. So boring that I couldn't watch the rest of it. It's not as interesting as the Lion King or Aladdin. That's why I hate this stupid days because these days are just not what they used to be. The movies that Disney have released lately really suck! I miss the old days when they used to produce good movies like the Lion King, Aladdin, Pocahontas, Cinderella, The Aristocats, and Robin Hood. This movie is really awful. I really tried my best to be interested in this movie but I just couldn't. I don't think I could recommend this film to anyone. I would recommend the ones that I just mentioned above"

Altra tipica fonte di errore riguarda l'accostamento di termini con significato opposto:

"I'm always nervous when I order online because I haven't the assurance of the quality and the product itself. As expected the product is of poor quality and is flammable. I do not think I will order anymore online also because the return is at my expense."

Alcuni di questi “inconvenienti” potrebbero essere in parte risolti inserendo ad esempio:

- dei tag relativi al nome del locale o del prodotto, con il beneficio di avere già delle recensioni raggruppate e quindi più semplici da reperire per quel locale;
- utilizzare una valutazione diversa dei dati ogniqualvolta viene citato un riferimento esplicito per la valutazione della qualità, come nel caso delle Stelle Michelin.

Acronimi

SA	*Sentiment Analysis*
OM	*Opinion Mining*
ML	*Machine Learning*
SVM	*Support Vector Machine*
NN	*Neural Network*
NB	*Naïve Bayes Classifier*
BN	*Bayesian Network*
ME	*Maximum Entropy Classifier*
WSD	*Word Sense Disambiguation*
NLTK	*Natural Language Toolkit*

Bibliografia

[1] "*SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining*"; di Baccianella Stefano; Esuli Andrea e Sebastiani Fabrizio. http://nmis.isti.cnr.it/sebastiani/Publications/LREC10.pdf

[2] "*Sentiment analysis algorithms and applications: A survey.";* di Medhat Walaa; Hassan Ahmed e Korashy, Hoda Korashy; 27 Maggio 2014. http://www.sciencedirect.com/science/article/pii/S2090447914000550#b0410

[3] *"Introduction to WordNet: An On-line Lexical Database";* di Miller George A; Beckwith Richard; Fellbaum Christiane; Gross Derek; Miller Katherine; Agosto 1993.
http://wordnetcode.princeton.edu/5papers.pdf

[4] *"Python 3 Text Processing with NLTK 3 Cookbook";* di Perkins Jacob; Agosto 2014.

[5] "*Word Sense Disambiguation: A Survey*"; di Navigli Roberto; Università di Roma La Sapienza; Febbraio 2009.
http://wwwusers.di.uniroma1.it/~navigli/pubs/ACM_Survey_2009_Navigli.pdf

[6] *"Sentiment Analysis of Figurative Language using a Word Sense Disambiguation Approach"*; di Rentoumi Vassiliki e Giannakopoulos George ; International Conference RANLP 2009 - Borovets, Bulgaria. http://www.aclweb.org/old_anthology/R/R09/R09-1.pdf#page=394.

[7] "*Introduction to Sentiment Analysis."* http://www.lct-master.org/files/MullenSentimentCourseSlides.pdf

[8] *"Python 3 Text Processing with NLTK 3 Cookbook";* di Perkins Jacob; Agosto 2014.

[9] http://www.di.uniba.it/~semeraro/LT/NLP_intro.pdf

[10] http://www.nltk.org/