ALMA MATER STUDIORUM - UNIVERSITÀ DI BOLOGNA

CAMPUS DI CESENA
SCUOLA DI SCIENZE
Corso di laurea in Ingegneria e Scienze Informatiche

Strumenti di monitoraggio di dati web non strutturati

Relatore:
Prof. Antonella Carbonaro

Presentata da:
Andrea Pruccoli

Sessione III
Anno Accademico 2014/2015

PAROLE CHIAVE

Sentiment Analysis

Opinion Mining

Text Mining

Twitter

Indice

Introduzione

Una parte fondamentale dell'era dell'informazione è sempre stata quella di scoprire quali fossero le opinioni della gente, e prima della diffusione della rete internet e del Web era di uso comune chiedere a parenti e ad amici il loro pensiero in merito ad un determinato argomento prima di prendere decisioni.

Negli ultimi anni i documenti web hanno attratto molta attenzione, poiché vengono visti come un nuovo mezzo che porta quello che sono le esperienze ed opinioni di un individuo da una parte all'altra del mondo, raggiungendo quindi persone che mai si incontreranno. Ed è proprio con la proliferazione del Web 2.0 che l'attenzione è stata incentrata sul contenuto generato dagli utenti della rete, i quali hanno a disposizione diverse piattaforme sulle quali condividere i loro pensieri, opinioni o andare a cercarne di altrui, magari per valutare l'acquisto di uno smartphone piuttosto che un altro o se valutare l'opzione di cambiare operatore telefonico, ponderando quali potrebbero essere gli svantaggi o i vantaggi che otterrebbe modificando la sia situazione attuale.

Per anni le aziende di prodotti o di servizi hanno condotto e somministrato sondaggi o questionari per capire il sentimento e l'opinione che i consumatori provano nei confronti di quello che offrono. Ora queste aziende ed organizzazioni si sono evolute, non somministrano più, non sottopongono più direttamente i consumatori a questionari o sondaggi, ma hanno a disposizione siti web di recensioni dove i loro clienti esternano le loro impressioni in merito ai prodotti e servizi offerti, elogiandone i pregi o accentuandone i difetti,

raccomandando un determinato servizio piuttosto che un altro per via di un determinato fatto.

Questa grande disponibilità di informazioni è molto preziosa per i singoli individui e le organizzazioni, che devono però scontrarsi con la grande difficoltà di trovare le fonti di tali opinioni, estrapolarle ed esprimerle in un formato standard. Queste operazioni risulterebbero quasi impossibili da eseguire a mano, per questo è nato il bisogno di automatizzare tali procedimenti, e la Sentiment Analysis è la risposta a questi bisogni.

Sentiment analysis (o Opinion Mining, come è chiamata a volte) è uno dei tanti campi di studio computazionali che affronta il tema dell'elaborazione del linguaggio naturale orientato all'estrapolazione delle opinioni. Negli ultimi anni si è rilevato essere uno dei nuovi campi di tendenza nel settore dei social media, con una serie di applicazioni nel campo economico, politico e sociale. Possiede un grande potenziale per essere sfruttato in strategie di impresa ed ha già aiutato aziende ed organizzazioni ad ottenere un riscontro in tempo reale delle reazioni dei consumatori ai nuovi prodotti lanciati o del pubblico alle nuove pubblicità, grazie alla condivisione di questi ultimi di stati e post sui social network più famosi.

Questa tesi ha come obiettivo quello di fornire uno sguardo su quello che è lo stato di questo campo di studio, con presentazione di metodi e tecniche e di applicazioni di esse in alcuni studi eseguiti in questi anni.

Più precisamente, nel primo capitolo verrà fornita una rapida panoramica sull'evoluzione che ha subito il web dalla nascita e perché sia considerato l'elemento fondamentale di questo campo di studio, seguita dall'introduzione dei temi del Data Mining e del Text Mining.

Nel secondo capitolo approfondiremo le tecniche e metodologie più utilizzate in letteratura per effettuare sentiment analysis.

Nel terzo capitolo vedremo un po' più da vicino le applicazioni di sentiment analysis sui social network, andando a vedere quali sono gli studi effettuati di recente sulla piattaforma di microblogging Twitter.

Concluderemo poi la trattazione presentando quelli che sono alcuni dei campi di studio strettamente legati alla sentiment analysis.

1 Panoramica

1.1 Breve storia del Web

Il *Web* [82] (abbreviazione di *World Wide Web – WWW*) nacque nel 6 agosto 1991, quando Berners-Lee, a quel tempo ricercatore presso i laboratori del CERN (l'organizzazione europea per la ricerca nucleare), rese disponibile su Internet il primo sito web, grazie al contributo del suo collega Cailliau. La necessità dell'implementazione di una tale struttura era data dall'esigenza di poter condividere documenti tra diverse postazioni.

I principali standard con cui il Web è implementato sono:

- *HTML (Hyper Text Markup Language):* linguaggio di formattazione che descriver le modalità di impaginazione o visualizzazione grafica del contenuto, testuale e non, di una pagina web attraverso tag di formattazione.

- *HTTP (HyperText Transfer Protocol):* protocollo di rete che permette la trasmissione delle informazioni.

- *URL (Uniform Resource Locator):* lo schema di identificazione e rintracciabilità dei contenuti e dei servizi del Web

In questa prima implementazione gli standard e i protocolli utilizzati permettevano la sola gestione di pagine HTML statiche, dove i file ipertestuali

presenti erano visualizzabili e consultabili come un libro da ogni utente che potesse accedervi, tramite l'utilizzo di apposite applicazioni, i browser. Per questo motivo oggi pensiamo a quello spazio virtuale come al Web 1.0, ovvero come ad una biblioteca dove poter consultare le informazioni che ci interessano.

La grande diffusione dell'utilizzo del Web inizio circa a metà del 1993, quando il CERN decise di rendere pubblica tale tecnologia, che fino a quel momento era rimasta ad uso esclusivo della comunità scientifica. Molte fra le grandi e piccole aziende videro una grande opportunità nell'utilizzo del Web [83] e decisero di investire in questa tecnologia, consentendo ad Internet di crescere commercialmente ed attirando sempre più nuovi utenti. Il grande sviluppo dello spazio virtuale si deve anche all'introduzione di JavaScript e CSS, il primo un linguaggio di scripting utilizzato per integrare, all'interno delle pagine web statiche, aspetti dinamici, il secondo un linguaggio di formattazione, che rende più "attraente" la visualizzazione di una pagina. I continui investimenti dalle aziende all'interno di questo nuovo settore in sviluppo vengono effettuati senza tenere conto dei modelli di business fondamentali, pensando solamente a recuperare i patrimoni investiti senza prendere in considerazione gli andamenti dei mercati sui quali le start-up finanziate erano impegnate. Questa disattenzione ha portato, nel 2001, a quello che viene ricordato come *crollo della bolla dot-com*, dove numerose furono le aziende che fallirono, alcune invece hanno resistito ed oggi sono ancora qua.

Dopo questo evento si inizia a pensare al Web come ad una tecnologia a disposizione degli utenti non più solamente per fattori economici. Da questo momento l'utente non resta semplicemente un lettore di informazioni, ma si trasforma anche in produttore di informazioni, diventa un *prosumer* (unione delle parole tra *producer* e *consumer*). I modi con cui l'utente produce informazioni e fornisce quindi conoscenza sul web sono i blog, i social network, le piattaforme di caricamento di contenuti multimediali, piattaforme di raccomandation e altre. Questi sistemi, dove l'utente svolge anche un ruolo attivo e non più solamente un ruolo passivo diventando così una figura fondamentale per il processo di sviluppo, vengono definiti come appartenenti al

Web 2.0 [84], termine utilizzato per la prima volta da O'Riley Media come titolo per una serie di conferenze aventi per oggetto una nuova generazione di servizi internet che enfatizzano la collaborazione online e la condivisione fra utenti.

La grande diffusione che ili Web ha ottenuto nel corso degli anni è dovuto sicuramente anche alla possibilità di accesso dai dispositivi mobili. Sin dagli anni '90 era possibile connettersi alla rete tramite i dispositivi mobili, cambiando in tal modo il concetto di fruizione dei servizi, ma l'esperienza che si viveva non era per niente comparabile con la comodità che aveva tramite navigazione da computer desktop o laptop, per questo motivo la diffusione era contenuta. Una maggiore ondata invece si ha dal 2007, quando con la commercializzazione del primo iPhone si introdusse il concetto di applicativi mobili e consentendo tutte quelle attività che oggi ormai sembrano normali come il geotagging o lo sharing di immagini immediato.

1.2 Data Mining

Il *Data Mining* può essere definito come il processo di estrazione delle informazioni implicite dai dati, precedentemente sconosciute e potenzialmente utili, o come il procedimento di esplorazione ed analisi di grandi quantità di dati al fine di scoprire pattern significativi, per mezzo di sistemi automatici o semi-automatici.

Le attività di un sistema di data mining sono tipicamente di due tipologie:

1. *Predizione:* si utilizzano alcune variabili per predire il valore di altre variabili, incognito o futuro.

2. *Descrizione:* inteso come il procedimento di identificazione di modelli ricorrenti nei dati, interpretabili dall'uomo in grado di descrivere i dati.

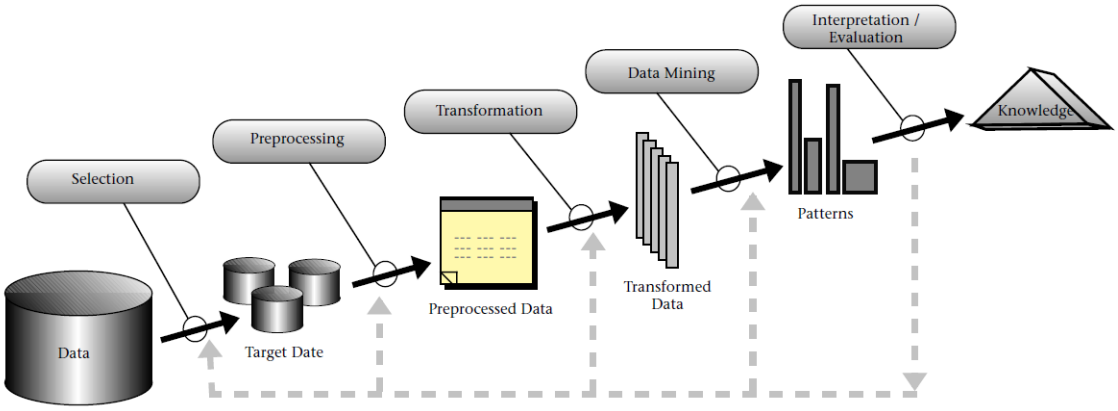

Figura 1: Estrazione della conoscenza

La figura soprastante [85] riporta quello che è un comune procedimento per l'estrazione della conoscenza, da un insieme di dati a disposizione, utilizzando algoritmi e tecniche di data mining.

La prima fase consiste nella raccolta di un'ingente quantità di informazioni del dominio di interesse dai quali si vuole estrapolare la conoscenza. Su questo grande insieme di dati raccolti si effettua una selezione, prendendo in considerazione quindi un sottoinsieme significativo di dati che si ritengono utili per il problema in esame. Una volta ottenuto un insieme limitato di dati, questi presentano informazioni in eccesso che possono, nella fase di mining, "sporcare" i risultati ottenuti. Per questo motivo il data set ottenuto dalla selezione viene sottoposto ad operazioni di pulitura dei dati e preprocessamento, grazie ai quali si riduce il rumore all'interno dei dati. Superata la fase di pulitura. Prima di raggiungere la fase di mining vera e propria, i dati, ora preprocessati, vengono sottoposti ad una fase di trasformazione, con la quale si possono ridurre il numero di variabili da tenere in considerazione o trovare rappresentazioni invarianti dei dati. Eseguite queste operazioni di preparazione, l'applicazione di metodi e algoritmi di data mining porta all'estrazione dei modelli ricercati, i quali vengono interpretati e valutati, ottenendo i risultati finali dell'estrazione della conoscenza dall'insieme iniziale di dati.

Il processo di mining dei dati può essere notevolmente agevolata, eseguendo correttamente le fasi che la precedono o addirittura ripetendole più volte, ovviamente se questo porta ad evidenti miglioramenti.

Gli obiettivi di predizione e descrizione possono essere raggiunti utilizzando una certa varietà di metodi:

- *Classificazione:* la classificazione è una funzione di apprendimento che mappa (classifica) l'oggetto di un dato all'interno di una tra diverse classe predefinite.

- *Regressione:* la regressione è una funzione di apprendimento che mappa l'oggetto di un dato in una variabile di predizione a valori reali.

- *Clustering:* il clustering (o analisi dei gruppi) è un compito descrittivo che si prepone di identificare un insieme infinito di categorie o gruppi per descrivere i dati.

- *Summarization:* prevede l'utilizzo di metodi per la ricerca di una descrizione compatta per un sottoinsieme di dati.

- *Modellazione delle dipendenze:* modellare le dipendenze significa cercare un modello che descriva in modo appropriato le dipendenze significative che intercorrono fra le diverse variabili.

- *Identificazione del cambiamento e della deviazione:* si incentra nel trovare i cambiamenti più significativi nei dati confrontati con valori precedentemente misurati.

1.3 Text Mining

La diffusione del web e la facilità di accesso ad esso, dovuto grazie all'evoluzione delle tecnologie hardware e software, combinato con il grande sviluppo dei dispositivi mobili, hanno reso possibile la raccolta di grandi quantità di contenuti testuali sulla rete. Blog, social networks, forum, sono tutte piattaforme su cui gli utenti della rete hanno pieno accesso e libertà di condivisione di informazione. La grande disponibilità testuale ha generato l'esigenza di evolvere gli attuali metodi ed algoritmi di analisi dei dati per estrarre pattern interessanti in maniera dinamica e scalabile.

Il *Text Mining* è una branca del data mining che si concentra sull'analisi dei dati testuali, poiché questi differiscono da molti altri tipi di dati. Infatti i dati testuali sono di tipo non strutturato, ossia privi di uno schema o di un modello che permette di attribuire ad essi una semantica ben definita. In questo si differenziano dai dati di tipo strutturato, i quali presentano invece uno schema ben preciso, come ad esempio i record salvati all'interno di un database.

Data la presenza predominante dei dati testuali all'interno della rete (circa l'80% dell'informazione globale [86]), al text mining è attribuito un grande valore potenziale commerciale.

Nel contesto del text mining Aggarwal e Zhai [32] espongono diverse applicazioni possibili:

- *Estrazione di informazione da dati testuali:* applicazione chiave nel text mining, l'estrazione delle entità e delle relazioni che intercorrono tra esse è in grado di estrapolare informazioni dall'alto contenuto semantico.

- *Sintetizzazione dei dati:* i testi vengono processati con lo scopo di ottenere riassunti di documenti molto ricchi o facenti riferimento allo stesso argomento.

- *Text mining multi lingua:* la facilità di reperimento di documenti testuali in diverse lingue ha interessato lo studio dell'estrapolazione di informazioni da documenti provenienti dalle diverse zone del globo e scritti in lingua madre che riguardano lo stesso argomento.

- *Text mining nei social media:* l'estrapolazione di informazioni da dati testuali ottenuti dai social media è una delle sfide più sostenute, poiché deve affrontare la dinamicità dei contenuti, espressi molto spesso in modo non comune.

- *Opinion Mining (o Sentiment Analysis):* il campo di applicazione di interesse di questa tesi, ha come obiettivo l'elaborazione di dati testuali provenienti da varie fonti, al fine di ottenere l'opinione generale che la gente ha nei confronti di un certo argomento.

2 Sentiment Analysis

2.1 Introduzione

Quello che gli altri pensano, in merito ad un determinato argomento, è sempre stato un'informazione di importante valore durante il nostro processo di decisione. Quale prodotto di quale linea comprare, i pro e i contro delle singole proprietà di un prodotto, chi votare alle prossime elezioni, da chi andare per fare un controllo sulla propria auto di cui si teme essere giunto il momento della sua dipartita? Queste, e tante altre ancora, sono le domande che spesso e volentieri si chiedevano al proprio vicino, al collega o all'amico del bar, per avere consigli, per aiutarci nella nostra scelta.

Grazie alla diffusione che Internet ha avuto sin dall'inizio del terzo millennio, e grazie anche alla grande reperibilità di computer, tablet e smartphone coi quali navigare nel Web, scoprire quale opinione la gente (diverso dal conoscente o da un critico di grande fama) ha relativamente alla reputazione di un candidato politico o di una più comune lavatrice è sempre di più alla portata di tutti.

Perché in questi anni l'utente medio non è più solo spettatore all'interno del World Wide Web (WWW), ma ricopre anche il ruolo di attore, esprimendo le sue considerazioni su blog, social network e altri social media.

Lasciare che questa immensa mole di informazioni, formata da opinioni e pareri personali (in gran parte o il più delle volte), rimanga sepolta all'interno della piattaforma in cui è riversata non è l'obiettivo della sentiment analysis, la

quale si fa forza di questi elementi, andando ad esaminarle e ad estrare il contenuto che più è prezioso: il sentimento che l'autore dell'opinione vuole comunicare.

Con sentiment analysis [27] ci si riferisce a quel campo di studio che mira ad analizzare, all'interno di uno scritto, le opinioni, i sentimenti, le valutazioni, le emozioni, la mentalità e i giudizi di una persona rispetto a prodotti, servizi, organizzazioni, personaggi, eventi, problematiche, argomenti e loro attributi.

Tale procedura è molto utilizzata, ed è applicabile ad innumerevoli campi, come l'identificazione di parole chiave per migliorare le ricerche di mercato o semplicemente per comprendere meglio i gusti degli utenti. Sempre più imprese di medie e grandi dimensioni sfruttano questo procedimento per comprendere meglio i pareri degli utenti e riuscire a prevedere il mercato al meglio.

Le preferenze delle persone sono difficili da comprendere e fino a poco tempo fa erano anche difficili da ottenere, bisognava far compilare moduli specifici in occasioni determinate, come per esempio accade nei call center che richiedono un parere sulla gestione della chiamata. Gli svantaggi del vecchio metodo sono tutti relativi alla limitatezza delle informazioni e alla difficoltà nell'ottenerle, i detentori di tali informazioni sono solo alcune aziende le quali giocano un ruolo di mediazione, inoltre, gli utenti non sono sempre disposti a compilare documenti poiché richiede tempo.

Un'altra difficoltà è legata al documento stesso da compilare, l'utente risponde solo alle domande poste, pertanto risulta importante e difficile anche definire quali sono le domande migliori da fare per ottenere le risposte desiderate, ma è pur sempre vero che ciò può comportare un errore, ad esempio una domanda che può apparire come ottima, potrebbe risultare inutile nei confronti delle informazioni che si vogliono ottenere.

Il campo della sentiment analysis [1], e della opinion mining in generale, ha subito una forte crescita nella ricerca grazie a:

- Sviluppo di metodi di machine learning (ML) nell'ambito della natural language processing (NLP) e dell'information retrieval (IR);

- Disponibilità di data-set con cui "allenare" gli algoritmi di ML grazie alla nascita di siti di recensioni;

- Alla presa di coscienza delle sfide e delle applicazioni che questa area di studio offre.

2.1.1 Il sentimento e l'opinione

Il vocabolario della nostra lingua definisce in tale modo le parole sentimento ed opinione:

- *Sentimento [28]:* 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. […] Sensibilità, sensitività, capacità di sentire con l'animo, finezza di sentire.

- *Opinione [29]:* 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, si propone un'interpretazione personale che si ritiene esatta e a cui si dà perciò il proprio assenso, ammettendo tuttavia la possibilità di ingannarsi nel giudicarla tale.

Per una trattazione computazionale del problema queste due definizioni sono ad un livello troppo astratto, inoltre nel campo della sentiment analysis si tende ad usare in modo intercambiabile i termini sentimento ed opinione.

Una definizione più formale di questi due termini viene data da Liu e Zhang [30].

Si indichi con il termine ***entità*** una persona, un servizio, un prodotto, un evento o un'organizzazione. L'entità *e* è associata ad una coppia (*T, W*) dove *T* rappresenta le componenti di *e*, mentre *W* rappresenta l'insieme degli attributi di *e*. Un esempio di entità può essere un telefono cellulare, formato da un certo numero di componenti (quali schermo, batteria etc.) e possiede un insieme di attributi (quali resistenza agli urti, durata della batteria, qualità audio/video etc.).

Un'opinione o un sentimento sono definite da Liu e Zhang come una quintupla (e_i, a_{ij}, oo_{ijkl}, h_k, t_l), dove:

- e_i rappresenta l'entità,
- a_{ij} è un aspetto di e_i,
- oo_{ijkl} rappresenta l'orientamento dell'opinione riguardante l'aspetto a_{ij} dell'entità e_i,
- h_k ci dice chi esprime l'opinione (*opinion holder*),
- t_l Ci dice quando l'autore ha espresso l'opinione (il tempo).

L'opinione (o il sentimento) espressa dall'autore può essere positiva, negativa, neutrale o essere espressa con diversi livelli di intensità (utilizzando un sistema a stelle, come per esempio una valutazione da 1 a 5 stelle).

2.2 Fonti dei dati

Il campo della ricerca nella sentiment analysis si è evoluto rapidamente grazie alle piattaforme rese disponibili dall'avvento del Web 2.0, che hanno

permesso ai ricercatori di ottenere in questo modo una fonte quasi infinita di dati e informazioni su cui lavorare:

- *Siti di recensioni:* un sito di recensioni permette agli utenti di scrivere e condividere recensioni contenenti pareri personali in merito a persone, prodotti, servizi e aziende. Poiché lo scopo di una recensione è quella di valutare un argomento in particolare, l'analisi di questi dati si concentra in un singolo dominio, che è quello di discussione. Riuscire a classificare le recensioni in positive e negative è un grande vantaggio per le aziende e i consumatori, in quanto vengono messi in luce i pregi e i difetti dei vari prodotti e/o servizi, concedendo la possibilità al produttore di poter indirizzare la sua attività verso una certa direzione e al consumatore un mezzo di comparazione.

- *Blog:* con il termine blog ci si riferisce ad una pagina web dove vengono condivisi brevi paragrafi contenenti informazioni, opinioni o parti di un diario personale (dell'autore), ordinati cronologicamente e con una frequenza di aggiornamento non sempre regolare, che può variare da intervalli di qualche ora a intervalli di diversi giorni, concedendo una visione quasi in tempo reale degli avvenimenti.

- *Forum:* i forum sono spazi web all'interno dei quali gli utenti possono discutere liberamente (rispettando comunque certe regole di comportamento) di vari argomenti. Di solito ogni forum tratta un argomento in particolare, per cui anche qui, come nel caso dei siti di recensione, l'analisi dei dati si concentra in un solo dominio.

- *Social Network:* i social network sono siti o servizi online che tentano di emulare le relazioni sociali tra persone che si conoscono

o che condividono uno stesso interesse. Permettono agli utenti di condividere opinioni, idee, pensieri, attività, eventi e interessi che riguardano qualsiasi argomento, i quali creano in questo modo un enorme base di dati di informazioni.

Tutte queste risorse online, racchiuse nel termine *social media*, sono la nuova fonte di informazioni del web, poiché offre la possibilità all'intero mondo di connettersi, consentendo a persone che vivono a migliaia di distanza l'una dall'altra di influenzare reciprocamente il proprio pensiero.

2.3 Struttura

La sentiment analysis è un compito impegnativo che comprende l'elaborazione del linguaggio naturale, recupero di dati su cui basarsi per costruire modelli e/o per effettuare test, utilizzo di diversi approcci combinati o utilizzati separatamente.

Data la complessità che presenta, la sentiment analysis viene generalmente decomposta in diversi sotto compiti, trattando aspetti più particolari del problema:

- *Subjectivity Classification:* è il compito che ha come obiettivo la classificazione delle frasi come soggettive o oggettive, ovvero riuscire a distinguere quelle frasi che sono portatrici di opinioni personali dell'autore del testo da quelle che descrivono in modo oggettivo l'argomento trattato.

- *Sentiment Classification:* una volta trovate all'interno del testo frasi soggettive, l'obiettivo della sentiment classification è quello di riuscire ad assegnare un valore positivo o negativo alla polarità della frase che si sta analizzando.

- *Object/Feature Extraction:* considerato un compito opzionale della sentiment analysis come l'opinion holder detection, la object/feature detection ha come scopo l'identificazione dell'entità o delle diverse componenti dell'entità di cui si sta discutendo all'interno di un discorso.

- *Opinion Holder Extraction:* ha come obiettivo l'identificazione delle fonti, dirette o indirette, autrici dell'opinione in esame.

Ognuna delle parti sopra descritte può essere portata a termine andando ad analizzare il testo in esame a diversi livelli di dettaglio:

- *Document level:* la sentiment analysis effettuata a livello del documento tratta l'intero documento come l'unità base del quale si vuole determinare l'orientamento di opinione. Per semplificare questo compito si assume che tutte le opinioni che si trovano all'interno del testo sono appartenenti ad un singolo autore e che si riferiscono ad una singola entità o caratteristica. Qui la difficoltà risiede nel fatto che all'interno di un documento possono essere presenti (con grande probabilità) opinioni diverse e contrastanti tra loro, oppure che l'autore utilizzi modi indiretti per esprimere la sua opinione. Come inoltre accennato poco sopra, all'interno di un documento, che contiene un vasto numero di frasi, sono presenti sia frasi soggettive che frasi oggettive.

- *Sentence level:* il livello successivo a quello del documento è l'analisi delle singole frasi di un documento, trattandole come l'unità base da cui partire per calcolare la polarità globale. Un problema che si può incontrare a questo livello, sebbene tratti una porzione di testo molto minore rispetto a quella di livello

precedente, è quello di trovare frasi oggettive che però al loro interno presentano parole che hanno valenza sentimentale.

- *Word level:* il livello più profondo di analisi di un testo è quello a livello di parola o espressione. Si analizzano le singole parole o espressioni verbali, si calcolano le polarità di esse e poi si calcola la polarità globale della frase/documento.

2.4 Approcci e tecniche

In questo paragrafo presenteremo le principali metodologie usate in letteratura per affrontare il tema della sentiment analysis [31].

2.4.1 Feature Selection

Questo sotto paragrafo tratta la *Feature Selection*, l'estrazione e la selezione delle principali qualità del testo, intendendo con questa definizione il ruolo che una parola o un insieme di parole possono giocare all'interno di un documento.

2.4.1.1 Principali feature

Le principali caratteristiche [32] del testo che vengono prese in esame per operare feature selection sono:

- *Terms presence and frequency:* vengono considerate il numero di volte che una parola o un insieme di parole compaiono all'interno del testo. Si può utilizzare dando alle parole un peso binario in base alla presenza o meno della parola/e, oppure si utilizza il numero di conteggi per assegnare il peso in base all'importanza che la parola o le parole hanno all'interno del documento o frase.

- *Parts of speech (POS):* vengono presi in esame le unità grammaticali del testo, quali aggettivi, avverbi, verbi etc.

- *Opinion words and phrases:* esistono parole che vengono comunemente utilizzate per esprimere opinioni, come ad esempio *buono* o *cattivo*, *amore* o *odio*. Sono presenti anche espressioni verbali che, anche senza contenere parole che esprimono direttamente un'opinione o un sentimento, ne sono comunque portatrici, come ad esempio "*Questo libro mi è costato un occhio della testa!*".

- *Negations:* la presenza di negazioni invertono completamente la polarità di una parola, per esempio *non buono* equivale a *cattivo*.

2.4.1.2 Metodi per eseguire Feature Selection

I metodi per eseguire feature selection possono essere rozzamente suddivisi in metodi lexicon-based e metodi statistici. Mentre i primi necessitano dell'intervento dell'uomo in alcune fasi della loro realizzazione, i metodi statistici sono completamente automatici e per questo maggiormente utilizzati in questo ambito.

Le tecniche di feature selection trattano il documento in esame come un gruppo di parole (*Bag of Words, BOW [33])*, dove la rappresentazione del documento ignora l'ordine delle parole, oppure come una stringa, che conserva la sequenza di parole nel testo. Uno dei più comuni passi che si effettua quando si esegue l'estrazione delle caratteristiche da un documento è la rimozione delle *stop-words* (parole di uso molto comune all'interno di un testo) e lo *stemming [34]* (riduzione delle forma flessa di una parola alla sua forma radice).

2.4.1.2.1 Point-wise Mutual Information (PMI)

Il valore di mutua informazione fornisce la possibilità di modellare la mutua informazione che intercorre tra le feature e le classi (positiva, negativa, neutrale) ed è derivato dalla teoria dell'informazione [35].

La mutua informazione puntuale (*point-wise mutual information – PMI*) $M_i(w)$ tra la parola *w* e la classe *i* è definita sulla base del livello di coincidenza

fra la classe e la parola. La coincidenza attesa della classe *i* e della parola *w*, sulle basi della mutua indipendenza, è data da $P_i \bullet F(w)$, mentre la coincidenza reale è data da $F(w) \bullet p_i(w)$, dove P_i è la frazione globale dei documenti che contengono la classe *i*, $F(w)$ è la frazione globale dei documenti che contengono la parola *w* e $p_i(w)$ è la probabilità condizionale della classe *i* per i documenti che contengono *w*.

La mutua informazione è definita in termini di rapporto tra i due valori di coincidenza sopracitati, data dalla seguente equazione:

$$M_i(w) = \log\left(\frac{F(w)\cdot p_i(w)}{F(w)\cdot P_i}\right) = \log\left(\frac{p_i(w)}{P_i}\right)$$

Formula 1: calcolo del valore di mutua informazione

La parola *w* correlata in modo positivo con la classe *i* quando $M_i(w)$ è maggiore di 0, mentre è correlata in modo negativo alla classe *i* quando il valore è minore di 0. Il valore puntuale di mutua informazione prende in considerazione solo la forza di coincidenza tra la classe *i* e la parola *w*.

2.4.1.2.2 Chi-square (χ^2)

Considerato *n* il numero totale di documenti in esame, la distribuzione statistica di chi quadrato della parola *w* e la classe *i* è definita come:

$$\chi_i^2 = \frac{n\cdot F(w)^2\cdot (p_i(w) - P_i)^2}{F(w)\cdot(1-F(w))\cdot P_i\cdot(1-P_i)}$$

Formula 2: distribuzione statistica di chi quadrato

PMI e χ^2 sono due modi diversi di misurare la correlazione che esiste tra termini e categorie. Inoltre, il valore di χ^2 viene considerato migliore rispetto a

quello restituito dalla mutua informazione puntuale, perché è un valore normalizzato, pertanto i valori ottenuti dalla distribuzione di chi quadro sono più facilmente comparabili tra termini della stessa categoria [32].

2.4.1.2.3 Latent Semantic Indexing (LSI)

I metodi di feature selection tentano di ridurre la grandezza dello spazio dei dati andando ad estrarre gli attributi dall'insieme di origine, andando ad utilizzare metodi di trasformazione delle feature i quali creano un insieme degli attributi più piccolo tramite funzioni sull'insieme originale.

LSI [36] è uno di questi metodi, il quale trasforma lo spazio del testo in un nuovo sistema di assi che è una combinazione lineare delle feature originali. LSI usa tecniche di *Principal Component Analysis (PCA)* [37] per raggiungere questo scopo. Determina il sistema di assi che racchiude il più grande livello di informazioni riguardanti le variazioni dei valori degli attributi sottostanti.

Il principale svantaggio dell'LSI è che è una tecnica di tipo *unsupervised*, il che la rende cieca alle distribuzioni di classi sottostanti, perciò le feature individuate da tale metodo non sono necessariamente le vie migliori per cui la distribuzione delle classi dei documenti trattati possono essere separate.

Altri approcci statistici che possono essere usati per operare feature selection sono Hidden Markov Model (HMM) e Latent Dirichlet Allocation (LDA) [38].

2.4.2 Tecniche di Sentiment Classification

Le tecniche di sentiment classification possono essere rozzamente suddivise in approcci di tipo *machine learning (ML)* e approcci di tipo *lexicon-based*. Gli approcci di tipo machine learning utilizzano algoritmi di machine learning famosi in letteratura e usano feature linguistiche, mentre gli approcci di tipo lexicon-based si basano su dizionari del sentimento, ovvero una collezione di termini polarizzati conosciuti e precompilati.

I metodi di classificazione del testo di tipo machine learning possono essere generalmente suddivisi in metodi *supervised* e *unsupervised*. I primi fanno utilizzo di un vasto numero di documenti etichettati durante la fase di training (allenamento del classificatore), mentre i secondi vengono utilizzati quando non è possibile ottenere i documenti etichettati, rendendo difficile e/o svantaggioso il training.

I metodi lexicon-based sono vincolati alla scelta del dizionario di opinioni utilizzato per analizzare il testo, e sono l'approccio dictionary-based e l'approccio corpus-based. L'approccio dictionary-based (basato sul dizionario) cerca le parole che esprimono un'opinione all'interno del testo, dopodiché ne assegna il valore andando a cercare nel dizionario i vari sinonimi o contrari. L'approccio di tipo corpus-based parte da una lista base di parole portatrici di un'opinione, per poi trovarne altre all'interno di un vasto corpus, in modo da facilitare la ricerca di parole che esprimono un'opinione per un contesto specifico.

2.4.2.1 Approcci Machine learning

Gli approcci di tipo machine learning si basano sui famosi algoritmi di machine learning per risolvere l'analisi del sentimento come un regolare problema di classificazione del testo che fa uso di feature sintattiche e/o linguistiche.

Definiamo innanzitutto il problema: abbiamo un insieme $D = \{X_1, X_2, X_3, \ldots, X_n\}$ di elementi utilizzati per il training, ognuno dei quali è etichettato ad una delle classi possibili. Il modello di classificazione è relativo alle feature dell'elemento sottostante ad una delle etichette delle classi. Poi, per una data istanza di una classe sconosciuta, si utilizza il modello per predire a quale etichetta quella istanza appartiene.

2.4.2.1.1 Metodi supervised

Come già precedentemente detto, i metodi di machine learning di tipo supervised sono dipendenti dall'esistenza di documenti già etichettati utilizzati

per il training del classificatore. In letteratura sono presenti diversi generi di classificatori di tipo supervised.

Classificatori Probabilistici

I classificatori probabilistici fanno uso di modelli mixture [39] per la classificazione, i quali assumono che ogni classe è un componente della popolazione. Ogni componente è un modello generativo [40] che fornisce la probabilità di campionare un particolare termine per quel componente. Per questo motivo questo genere di classificatori sono anche detti *generativi*. I più famosi classificatori probabilistici sono il *Naive Bayes Classifier*, il *Bayesian Network* e il *Maximum Entropy Classifier*.

Il classificatore Naive Bayes(NB) è il più semplice e comune classificatore probabilistico utilizzato in letteratura. Questo modello calcola la probabilità a posteriori di una classe basata sulla distribuzione delle parole nel documento. Fa utilizzo della feature BOW (bag-of-words vista preceentemente), la quale ignora la posizione delle parole nel documento, e del teorema di Bayes per predire la probabilità con la quale una data feature appartenga ad una particolare etichetta.

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

Formula 3: teorema di Bayes

P(label) è la probabilità a priori di un'etichetta o la probabilità con cui un insieme di feature casuale ricada in quell'etichetta, *P(features|label)* è la probabilità a priori che l'insieme di feature dato sia classificato con quell'etichetta, mentre *P(features)* è la probabilità a priori di un insieme di feature di apparire.

L'assunzione principale che il classificatore di Naive Bayes opera è che le feature siano fra loro indipendenti. L'altra assunzione, quella opposta, è che tutte

le feature sono fortemente dipendenti tra loro. Questo ci conduce ai classificatori Bayesian Netword (BN), il quale è un grafo aciclico direzionato, i quali nodi rappresentano variabili casuali e gli archi rappresentano dipendenze condizionali. Nel text mining, la complessità di computazione della rete bayesiana è molto alto, questo fatto denota il suo scarso utilizzo.

Il classificatore Maximun Entropy (chiamato anche MaxEnt – ME), conosciuto per essere un classificatore esponenziale condizionale, converte tramite codifica gli insiemi di feature etichettate in vettori. I vettori codificati così ottenuti sono utilizzati per calcolare i pesi di ogni feature, i quali possono essere combinati per determinare l'etichetta più adatta per un dato insieme di feature. Questo classificatore è parametrizzato da un insieme di pesi *X*, utilizzato per combinare le feature unite che sono state generate dall'insieme di feature tramite un insieme di codifiche *Z*. Le codifiche mappano ogni coppia di insieme di feature ed etichetta *C{(feature-set, label)}* all'interno di un vettore. La probabilità di ogni etichetta è successivamente calcolata utilizzando la formula seguente:

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

Formula 4: formula calcolo probabilità in ME

Dove *dotprod* indica un prodotto scalare tra i pesi e il vettore ottenuto tramite codifica e *sum(dotprod(..)..)* rappresenta la somma di tutti i prodotti scalari, per ogni etichetta.

Classificatori Lineari

Definito come $\bar{U} = \{u_1,\ldots,u_n\}$ il vettore normalizzato delle frequenza di una parola all'interno del documento, $\bar{A} = \{a_1,\ldots,a_n\}$ un vettore di coefficienti lineari che ha la stessa dimensione dello spazio delle feature, e *b* uno scalare, il risultato di un classificatore lineare è dato da $p = \bar{A} \bullet \bar{U} + b$, dove *p* è un

iperpiano che divide, separa le diverse classi. I due più famosi classificatori di questa categoria sono il *Support Vector Machines* e il *Neural Network.*

I classificatori di tipo Support Vector Machines (SVM) hanno come principio quello di determinare i separatori lineari che meglio sono in grado di dividere tra loro le diverse classi.

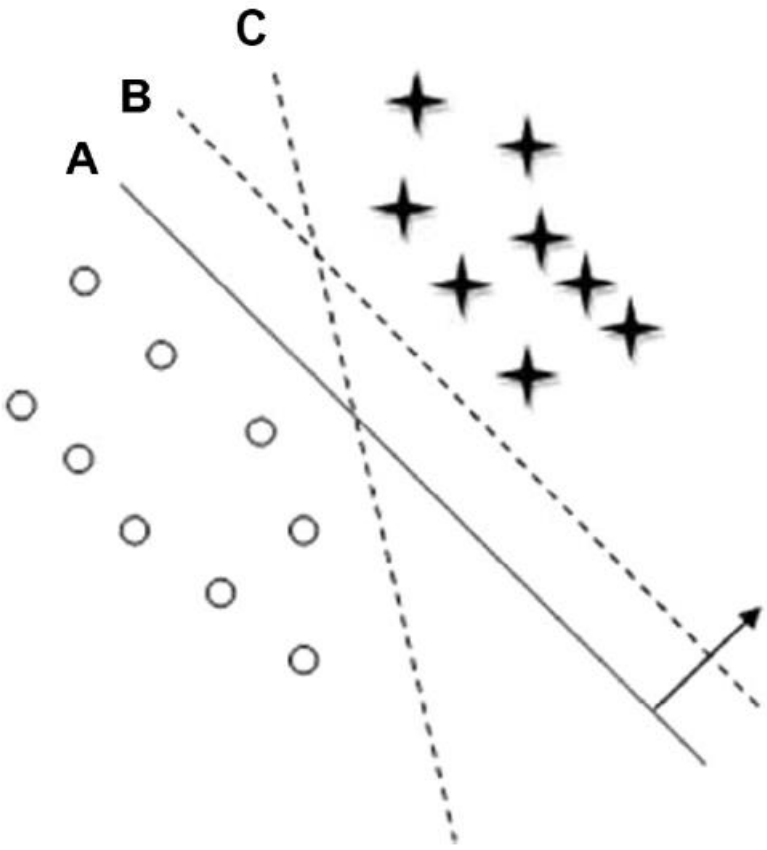

Figura 2: SVM in un problema di classificazione delle classi

Come mostrato in figura, le due classi *x* e *o* sono separate da tre iperpiani A, B e C. L'iperpiano A è quello che offre la miglior separazione tra le due classi, poiché la distanza normale di ogni punto è la maggiore, offrendo così la rappresentazione del massimo margine di separazione.

Dati ottenuti dai testi sono i più adatti per le classificazioni tramite SVM, per via della natura sparsa del testo nella quale poche delle feature presenti sono irrilevanti, ma che comunque tendono ad essere correlate le une con le altre e

generalmente organizzate per essere all'interno di categorie linearmente separabili. Classificatori SVM possono costruire piani di decisione non lineari nello spazio originale delle feature, mappando in modo non lineare le istanze di dati con uno spazio vettoriale nel quale è definito un prodotto interno, dove le classi possono essere separate linearmente tramite iperpiani [41].

I Neural Network (NN) sono formati da tanti neuroni, dove il neurone è la sua unità di base. Gli input dei neuroni sono denotati dal vettore $\bar{U}$, che sono le frequenze delle parole all'interno dell'*i*-esimo documento. Ad ogni neurone è associato un insieme di pesi A, usati all'interno della funzione $p_i = A \bullet \bar{U}_i$. In un problema di classificazione binaria (positivo, negativo) si assume che l'etichetta della classe di $\bar{U}_i$ è rappresentata da y_i e il segno della funzione di predizione p_i svela l'etichetta della classe.

Alberi di decisione

I classificatori che utilizzano gli alberi di decisione eseguono una decomposizione gerarchica sullo spazio dei dati utilizzati per il training, all'interno della quale una condizione sul valore di un attributo è utilizzata per dividere i dati [42]. La condizione, o il predicato, è la presenza o meno di una o più parole. La suddivisione dello spazio dei dati è eseguita in modo ricorsivo, fino a quando i nodi foglia non contengono un numero sufficiente di elementi che vengono utilizzati proprio per la classificazione.

Esistono altri tipi di predicati che dipendono sulla similarità dei documenti per correlare gli insiemi di elementi che potrebbero essere utilizzati per partizionare ulteriormente i documenti. Tra questi ci sono il *Single attribute split*, che opera la divisione in base alla presenza o meno di particolari parole o espressioni all'interno di uno dei nodi dell'albero, il *Similarity-based multi-attribute split*, che fa uso di gruppi di documenti o di parole frequenti e la similarità tra questi per eseguire la suddivisione, il *Discriminant-based multi-attribute split*, che fa uso di discriminanti per la suddivisione [43, 44].

Classificatori rule-based

Nei classificatori rule-based, lo spazio dei dati è modellato in base ad un insieme di regole. La parte sinistra rappresenta una condizione sull'insieme di feature espresso in forma normale disgiunta, mentre la parte destra rappresenta l'etichetta della classe in cui ricadono. L'assenza del termine è raramente utilizzata come regola perché non è molto informativa per uno spazio di dati sparso.

Esistono numeri criteri che vengono utilizzati per generare l'insieme di regole e la fase di training fa uso di questi criteri per costruire le regole. I due criteri più comuni sono il *supporto* e la *confidenza* [45]. Il supporto è il numero assoluto di istanze nel data set utilizzato per il training che sono attinenti alla regola, mentre la confidenza fa riferimento alla probabilità condizionale che la parte destra della regola è soddisfatta se la parte sinistra è soddisfatta.

Sia gli alberi di decisione che le regole di decisione tendono a codificare delle regole sullo spazio delle feature. Quinlan [42] ha studiato i problemi relativi agli alberi di decisione e alle regole di decisione all'interno di una singola struttura. La principale differenza tra alberi di decisione e regole di decisione è che i primi effettuano un rigido partizionamento gerarchico dello spazio dei dati, mentre i classificatori di tipo rule-based permettono le sovrapposizioni all'interno dello spazio delle decisioni.

2.4.2.1.2 Metodi unsupervised

Il principale obiettivo della text classification è quella di classificare i documenti entro un certo numero di categorie. Per raggiungere tale scopo si fa largo utilizzo di documenti etichettati per eseguire il training nei metodi di supervised learning. A volte però risulta difficile creare questi documenti etichettati, mentre è molto più facile collezionarne di non etichettati.

Ko e Seo [46] hanno proposto un metodo che divide il documento in frasi, che vengono categorizzate utilizzando liste di parole chiave per ogni categoria e valori di similarità per le frasi.

Xianghua e Guo [47] hanno utilizzato un approccio di tipo unsupervised per scoprire in modo automatico gli aspetti discussi nelle recensioni sociali in lingua cinese e il sentimento che veniva espresso in ogni diverso aspetto. Hanno fatto uso di un modello LDA per trovare i diversi argomenti globali delle recensioni, dopodiché hanno assegnato la polarità del sentimento da ogni argomento estratto in modo locale. Hanno dimostrato che il loro approccio ottiene buoni risultati per quanto riguarda la partizione degli argomenti, favorendo anche il miglioramento della precisione nella sentiment analysis.

Approcci di tipo supervised e unsupervised posso essere combinati per ottenere risultati migliori. Su questa convinzione si basa il lavoro svolto da Martín-Valdivia et al [48], i quali propongono l'uso di meta classificatori con lo scopo di sviluppare un sistema di classificazione delle polarità.

Il loro lavoro si è basato sull'utilizzo di un corpus di recensioni di film in lingua spagnola, con un corpus parallelo delle stesse recensioni, ma tradotte in lingua inglese. Inizialmente hanno creato due modelli individuali tramite l'uso dei due corpora, successivamente hanno applicato diversi algoritmi di machine learning (SVM, NB e altri). Successivamente hanno integrato il corpus di SentiWordNet all'interno di quello in lingua inglese, generando un nuovo modello unsupervised utilizzando approcci di orientazione semantica. Infine, tramite l'uso di un meta classificatore per combinare i tre sistemi ottenuti, Martín-Valdivia et al hanno ottenuto dei risultati che superavano quelli ottenuti usando i corpora a disposizione in maniera individuale, dimostrando così che il loro approccio potrebbe essere considerata una buona strategia da prendere in considerazione quando si hanno a disposizione corpora paralleli.

2.4.2.2 Approcci Lexicon-based

Le parole che trasmettono comunemente un'opinione (opinion word) vengono tenute in considerazione in molti dei compiti della classificazione del sentimento. Quelle positive sono utilizzate per esprimere uno stato in cui ci si trova piacevolmente, mentre quelle negative indicano uno stato in cui non ci si

trova piacevolmente. Congiuntamente con le singole parole, in molti ambiti vengono prese in considerazione anche le espressioni verbali che esprimono un'opinione e le espressioni idiomatiche, le quali formano l'*opinion lexicon*, il vocabolario specializzato dell'opinione.

Esistono principalmente tre approcci che permettono di collezionare liste di tali parole. L'approccio manuale è quello che richiede direttamente l'intervento dell'uomo per categorizzare le parole e le espressioni nelle varie classi. Questo approccio richiede molto tempo per ottenere un risultato e quindi viene generalmente utilizzato come fase di controllo in uno degli altri due approcci automatizzati, nel caso servisse correggere a posteriori le etichette di alcune parole o espressioni.

2.4.2.2.1 Approccio dictionary-based

Negli approcci di tipo dictionary-based, ovvero basati sui dizionari, un insieme ridotto di opinion word di cui si conosce la polarità viene manualmente raccolto e collezionato. Successivamente questo insieme viene espanso, grazie all'integrazione nell'insieme di sinonimi e/o contrari delle parole presenti all'interno di corpora già assestati (come WordNet) o tramite l'utilizzo di dizionari di sinonimi [49]. I nuovi termini trovati vengono aggiunti alla lista iniziale, iterando tale procedimento in modo ricorsivo, fino a quando non si individua nessuna nuova parola. Al termine del procedimento di ricerca si può eseguire un controllo manuale sull'insieme ottenuto, effettuando eventuali correzioni.

Il grande svantaggio che questo tipo di approccio possiede è l'impossibilità di trovare termini che possiedono un'orientazione specifica del dominio a cui si riferiscono.

2.4.2.2.2 Approccio corpus-based

I metodi basati sul corpus aiutano a risolvere il problema definito poco sopra. I metodi di questo approccio dipendono da schemi sintattici o da schemi

che si presentano, usati in concomitanza con una lista di partenza di opinion word con l'obiettivo di trovarne altre, all'interno di un corpus più grande.

Uno di questi metodi è stato presentato da Hatzivassiloglou e McKeown [50], i quali hanno utilizzato una lista di partenza di aggettivi che esprimono un opinione insieme ad un insieme di vincoli linguistici per identificare ulteriori aggettivi e/o parole e la loro orientazione. I vincoli sono rappresentati da connettivi come *e*, *o*, *ma*, *l'uno o l'altro*, etc. Per esempio, la congiunzione *e* indica solitamente che gli aggettivi collegati hanno la stessa orientazione, mentre *ma* indica un cambio di opinione.

Per determinare se due aggettivi collegati esprimono un sentimento della stessa polarità o meno, tecniche di apprendimento vengono utilizzate all'interno di un grande corpus. I collegamenti tra gli aggettivi formano un grafo su cui viene eseguito *clustering* (analisi di gruppi), in modo da produrre due insiemi di parole distinte, quelle con valenza positiva e quelle con valenza negativa.

Questo approccio applicato da solo non è molto efficace come quello basato sul dizionario, perché è difficile riuscire a creare una grande raccolta in grado di mappare tutte le parole in lingua, ma grazie alla sua peculiarità, è in grado di trovare quelle parole e quelle espressioni che sono tipiche del dominio che si sta analizzando. Gli approcci corpus-based sono effettuati utilizzando approcci statistici o semantici, di cui daremo una breve presentazione di seguito.

Approccio statistico

Tramite approcci statistici è possibile eseguire una ricerca per trovare schemi di coincidenza o parole base che esprimono opinioni, derivando le polarità a posteriori tramite le co-occorrenze di aggettivi all'interno di una raccolta, o utilizzando l'intero insieme di documenti indicizzati sul web come corpus per la costruzione del dizionario, superando il problema della non disponibilità di alcune parole in raccolte non molto grandi [51,52].

La polarità di una parola può essere identificata studiando la frequenza con cui la parola compare in una raccolta. Se compare più frequentemente fra testi positivi, allora la sua polarità sarà positiva, mentre se compare più

frequentemente in testi negativi, la sua polarità sarà negativa. Nel caso in cui le frequenze con cui compare nei testi delle due polarità sono pressoché uguali, la parola avrà valenza neutrale.

Parole che hanno valenza similare compaiono frequentemente all'interno di una raccolta, questa la principale osservazione su cui si basano i metodi all'attuale stato dell'arte. Quindi due parole che compaiono spesso all'interno dello stesso contesto tenderanno ad avere la stessa polarità, questo ci permette di determinare la polarità di parole di cui non si sa la valenza calcolando la frequenza relativa con la quale le parole compaiono, all'interno di un testo, con un'altra parola di cui si conosce la polarità. Per questo scopo si può utilizzare la PMI [52].

Tramite *Latent Semantic Analysis* (*LSA*) [36] si possono analizzare le relazioni che intercorrono tra un insieme di documenti e i termini che vengono utilizzati in questi documenti, con lo scopo di produrre un insieme di schemi significativi collegati ai documenti e ai termini.

Approccio semantico

L'approccio semantico assegna direttamente i valori del sentimento e fa affidamento su diversi principi per calcolare la similarità fra le parole. Questo principio assegna valori di sentimento simili a parole che sono vicine semanticamente. Per esempio WordNet mette a disposizione diverse tipologie di relazione semantica tra le parole, usate per calcolare le polarità. Può essere utilizzato anche per ottenere una lista di parole polarizzate espandendo in modo iterativo l'insieme di partenza, cercando sinonimi e contrari delle parole ed espressioni presenti, per riuscire a determinare la polarità di una parola sconosciuta tramite il conteggio relativo dei suoi sinonimi positivi e negativi [53].

Metodi statistici e metodi semantici possono essere combinati per eseguire analisi del sentimento, come proposto dal lavoro di Zhang e Xu [54], dove l'utilizzo di entrambi i metodi è servito per la ricerca delle debolezze di un

prodotto ottenute tramite recensioni online. Il loro strumento per la ricerca di debolezze estrae le feature e gruppi espliciti di feature utilizzando un metodo basato sul morfema [55] per identificare le feature delle parole dalle recensioni. La ricerca di feature frequenti e non frequenti che descrivono uno stesso aspetto è stata svolta utilizzando la misura di similarità basata su HowNet, l'identificazione di feature implicite è stata eseguita tramite PMI. Tramite utilizzo di metodi semantici hanno raggruppato le feature relative ai prodotti con i corrispondenti aspetti, infine tramite metodi di analisi del sentimento basati sulle frasi hanno determinato la polarità di ogni aspetto, tenendo in considerazione l'impatto degli avverbi all'interno del contesto.

Come risultato, Zhang e Xu sono riusciti a determinare le debolezze dei prodotti, rappresentate dagli aspetti meno soddisfacenti riscontrati dai consumatori ed esternate nelle loro recensioni, o quegli aspetti che sono più insoddisfacenti quando comparati con i prodotti di aziende competitrici.

2.4.2.2.3 Tecniche lexicon-based e NLP

Tecniche di *natural language processing* (*NLP*) vengono a volte utilizzate assieme a tecniche basate sul lessico per trovare la struttura sintattica dei documenti, in modo da aiutare la ricerca di relazioni semantiche.

Tecniche NLP sono state utilizzate da Moreo et al [56] come fase di precompilazione prima di utilizzare algoritmi basati sul lessico. Il sistema che propongono si compone di un modulo di rilevazione automatica del focus e un modulo per l'analisi del sentimento capaci di valutare le opinioni di un utente in merito agli argomenti di alcune notizie che sfrutta un lessico di tassonomie progettato specificatamente per l'analisi delle notizie. Si sono riscontrati buoni risultati quando a predominare è il linguaggio colloquiale.

Caro e Grella [57] propongono un approccio basato su un'analisi profonda delle frasi tramite NLP, usando in una fase di precompilazione un decodificatore di dipendenze. Il loro algoritmo di analisi del sentimento si basa sul concetto di *Sentimento Propagation* (propagazione del sentimento), con il

quale si assume che ogni elemento linguistico come un sostantivo, un verbo, un avverbio etc. possono possedere un valore del sentimento intrinseco, che si propaga all'interno della struttura sintattica della frase analizzata. Caro e Grella presentano un insieme di regole basate sulla sintassi che mirano a coprire una parte significante del rilievo sentimentale espresso da un testo e propongono un sistema di visualizzazione dei dati che mostra all'utente solo le informazioni rilevanti ai fini della sua ricerca, filtrando o contestualizzando i dati. Per la realizzazione di questo sistema hanno utilizzato un metodo basato sul contesto che visualizza le opinioni in basse alla misura della distanza tra le chiavi di ricerca e le polarità delle parole contenute nei testi.

2.4.2.3 Altri approcci

Esistono tecniche che non possono essere definite in modo preciso di machine learning o lexicon-based, come ad esempio la *Formal Concept Analysis* (*FCA*) proposta da Wille, un approccio matematico utilizzato per la strutturazione, analisi e visualizzazione dei dati basato su una nozione di dualità chiamata connessione di Galois [58,59]. I dati sono costituiti da insiemi di entità e le loro feature sono strutturate in astrazioni formali chiamate *concetti formali*, che formano un reticolo di concetti ordinato secondo una relazione d'ordine parziale. I reticoli sono costruiti tramite l'identificazione degli oggetti e dei loro attributi corrispondenti per un dominio specifico, denotati come *strutture concettuali*, e infine vengono mostrate le relazioni che intercorrono tra essi.

La tecnica *Fuzzy Forma Concept Analysis* (*FFCA*) [60] è stata sviluppata per meglio gestire le informazioni incerte e poco chiare.

2.5 Valutazione di un classificatore

Le prestazioni di un classificatore, nell'analisi del sentimento, possono essere valutate tramite il calcolo e lo studio di quattro indici:

- *Accuracy*: misura la capacità predittiva del classificatore.

$$Accuracy = \frac{VP + VN}{VP + VN + FP + FN}$$

Formula 5: calcolo dell'accuratezza di un classificatore

- *Precision*: denota la proporzione dei casi positivi predetti che sono realmente positivi.

$$Precision = \frac{VP}{VP + FP}$$

Formula 6: calcolo della precisione di un classificatore

- *Recall*: misura la proporzione dei casi positivi reali che sono stati correttamente predetti.

$$Recall = \frac{VP}{VP + FN}$$

Formula 7: calcolo del richiamo di un classificatore

- *F-measure:* o F_1 score è un valore che indica l'affidabilità del classificatore relativamente ai test eseguiti, potendolo considerare come una media pesata della precision e della recall.

$$F - measure = 2 * \frac{precision * recall}{precision + recall}$$

Formula 8: calcolo della F-measure di un classificatore

3 Sentiment Analysis nei social: Twitter

3.1 La piattaforma

Twitter [2] è un servizio gratuito di social networking e microblogging, creato nel marzo del 2006 dalla Obvious Corporation di San Francisco, il quale offre ai suoi utenti la possibilità di condividere messaggi di breve lunghezza sulla propria personale.

Questi messaggi, chiamati *tweet* (dal verbo inglese *to tweet*, ovvero "cinguettare"), sono diventati oggetto di interesse per la sentiment analysis negli ultimi anni, proprio grazie al vincolo imposto sulla lunghezza di questi messaggi (140 caratteri) nei quali l'autore deve condensare il suo pensiero da condividere con la comunità.

Utilizzando tecniche di text mining e NLP, una grande mole di informazioni utili può essere recuperata dai tweet sui quali si può effettuare opinion mining, per estrarre quello che è il pensiero della massa relativamente ad un argomento.

3.2 Caratteristiche del linguaggio

Negli anni della sua presenza nel web, Twitter ha sviluppato una particolare struttura nella composizione del messaggio, rendendo l'elaborazione del testo un passo fondamentale per il recupero di informazioni "pulite".

Alcune caratteristiche [6] chiave dei messaggi sono:

- *Lunghezza del messaggio:* Come detto anche sopra, la lunghezza massima di un tweet è di 140 caratteri. Questo è molto diverso rispetto a ricerche precedenti di classificazione del sentimento che si concentravano sulla classificazione di porzioni di testo più lunghe, come le recensioni di un prodotto o di un film.

- *Tecniche di scrittura:* Rispetto ad altri domini, la presenza di errori di battitura e di utilizzo di *slang* è maggiore. Data la loro ristretta lunghezza, gli autori dei messaggi usano acronimi, spesso sbagliano a scrivere le parole, utilizzano le emoticon e altri caratteri speciali ("#" = *hash tag*, *"@"* = *target*) che, per convenzione, assumono significati propri.

- *Disponibilità:* La quantità di informazioni disponibile è veramente enorme. Data la diversità nei termini di privacy rispetto ad altre piattaforme, i messaggi condivisi su Twitter sono visualizzabili da chiunque [3]. Twitter stesso mette a disposizione degli sviluppatori API specifiche atte al recupero dei tweet [4].

- *Topic* (Argomenti): Gli utenti pubblicano messaggi che possono riguardare argomenti diversi da loro, diversamente da altri siti che invece sono progettati e specifici per la trattazione di un dato argomento.

- *Tempo reale:* Mentre i blog vengono aggiornati in intervalli di tempo lunghi (ad esempio una volta al giorno, alla settimana, al mese etc.), i messaggi pubblicati su Twitter vengono aggiornati molto più velocemente, data la loro lunghezza, fornendo una visione in tempo reale delle reazioni agli eventi.

La piattaforma presenta anche quella terminologia propria del dominio, alcune già incontrate nell'elenco appena esposto:

- *Emoticon:* Sono rappresentazioni illustrate di espressioni facciali realizzate tramite la punteggiature e le lettere. Il loro scopo è quello di fornire una rappresentazione diretta del sentimento dell'autore.

- *Target:* Gli utenti fanno largo uso del carattere "@", chiamato target (o più comunemente tag), per fare riferimento ad altri utenti. Grazie a questa meccanica, l'utente riferito viene notificato automaticamente qualora vengano menzionati da qualcuno.

- *Hash tag:* Identificati dal simbolo "#", gli hash tag vengono utilizzati per evidenziare i topic di cui l'autore del messaggio sta scrivendo, in modo da renderli maggiormente visibili al pubblico. Infatti dalla home page di Twitter è possibile effettuare ricerche anche per hash tag.

- *Simboli speciali:* "RT" è una sequenza particolare che sta ad indicare un *re-tweet*, ovvero l'azione di condividere un tweet pubblicato precedentemente da un altro autore.

3.3 Recupero dei tweet

In questa sezione presentiamo quali sono i metodi più comuni e utilizzati per recuperare i messaggi dalla piattaforma di microblogging.

3.3.1 Streaming API

Poiché accennati precedentemente, introduciamo brevemente le API fornite da Twitter per il recupero dei messaggi condivisi ogni giorno dai suoi utenti, riferendoci in particolare a quelle dirette a recuperare il flusso di tweet relativo a determinati argomenti o utenti.

Le *Streaming API* forniscono tre livelli di recupero dei tweet: *POST statuses / filter*, *GET statuses / sample* e *GET / firehose.* Quella che più interessa a noi è la prima [5], poiché è in grado di restituire un insieme di tweet filtrati per determinati parametri:

- *follow:* questo predicato fornisce la possibilità di indicare di quali utenti recuperare i relativi messaggi. I messaggi recuperati saranno quelli creati dall'utente, quelli "retweettati" dall'utente, quelli che sono il retweet di un suo messaggio e quelli in risposta ad un tweet dell'utente. All'interno dell'archivio recuperato non saranno presenti i messaggi in cui l'utente è menzionato (tramite il carattere @), i retweet creati manualmente senza l'utilizzo dello specifico bottone e quelli degli utenti che hanno impostato a "protetto" il proprio profilo;

- *track:* forse il parametro più utilizzato, consente di recuperare i tweet che contengono le parole chiave descritte all'interno di questo predicato;

- *locations:* è possibile specificare una serie di coppie di coordinate geografiche (longitudine, latitudine) in modo da recuperare i messaggi geo localizzati all'interno di una delle zone definite;

- *delimeted:* specifica se i tweet recuperati debbano essere o meno delimitati dalla lunghezza in byte che occupano all'interno dello stream dati, in modo da rendere noto a priori quanto contenuto leggere prima di arrivare a processare il tweet successivo (utile per il parse automatico dei tweet tramite programmi ad hoc);

- *stall_warnings:* se impostato a "true", questo parametro invia periodicamente messaggi di avvertimento all'utilizzatore delle API nel momento in cui è a rischio disconnessione (utile per un monitoraggio real time del flusso di dati recuperato).

Quando si vuole eseguire un recupero di messaggi tramite questa tipologia di API, almeno uno dei primi tre parametri deve essere utilizzato nella richiesta.

I tweet vengono recuperati sotto forma di file *JSON* (*JavaScript Object Notation*), contenente, per ogni tweet, molte informazioni, tra cui le più importanti sono:

- *id:* valore che identifica in modo univoco il tweet;

- *text:* testo completo del messaggio, di massimo 140 caratteri;

- *user:* valore che identifica in modo univoco l'utente autore del messaggio;

- *lang:* lingua in cui il tweet è stato scritto;

- *created_at:* data del momento di creazione del messaggio, espresso nel formato UTC;

- *coordinates:* coordinate geografiche del luogo da cui è stato postato il messaggio;

- *favourite_count:* numero di volte in cui il messaggio è stato scelto come "preferito" (equivalente alla funzionalità "Mi piace" di Facebook);

- *retweet_count:* numero di volte in cui il messaggio è stato retweettato;

- *in_reply_to_status_id:* se il tweet corrente è una risposta ad un altro tweet, questo campo conterrà l'id di tale messaggio.

3.3.2 Extraction Tools

Diversamente dalle API, esistono anche piattaforme web che svolgono operazioni di recupero dei tweet:

- *Sentiment140 [24]:* restituisce i tweet più recenti relativi alla parola chiave inserita nella barra di ricerca, evidenziando in verde quelli calcolati avere una polarità positiva, rossi se negativi e bianchi se neutrali. Mostra anche un grafico riassuntivo del numero di tweet positivi e di quelli negativi.

- *StreamCrab [25]:* simile a Sentiment140, StreamCrab restituisce i tweet più recenti relativi alla parola chiave di interesse, mostrando grafici riguardanti alla percentuale di tweet positivi e negativi, la

somma delle polarità, la tendenza che ha avuto la polarità nel corso del tempo e la distribuzione della polarità nel tempo.

- *Sentiment viz [26]:* suddiviso in sezioni, Sentiment viz offre la possibilità all'utente non solo di leggere i tweet e vederne le polarità calcolate, ma permette anche di vedere classificazioni più specifiche dei sentimenti presenti all'interno dei tweet, vederne il raggruppamento per argomento e un grafico dove vengono mostrate le affinità con altri argomenti/utenti.

3.4 Sfide nell'estrapolazione del sentimento in Twitter

Di seguito verranno elencate alcune delle sfide [7] tutt'ora attuali quando si vuole effettuare sentiment analysis su dati raccolti da Twitter.

NB: le frasi utilizzate come esempio nell'elenco che segue sono in lingua inglese, poiché è in tale lingua che si sono svolte e sviluppate la maggior parte delle ricerche.

1. *Identificare le porzioni del testo che sono soggettive*: Le porzioni di testo soggettive sono quelle che portano con loro informazioni sul sentimento proprio dell'autore. Il problema di identificare una porzione del testo come soggettiva, portatrice del sentimento che si vuole trovare, risiede nel fatto che una parola può essere considerata soggettiva se usata all'interno di un contesto, ma può essere anche considerata oggettiva all'interno di un altro. Per esempio, "The language of the author was very crude" (Il linguaggio dell'autore era molto crudo) esprime un'opinione nei confronti del linguaggio dell'autore, mentre "Crude oil is extracted from the sea beds" (Il petrolio greggio è estratto dai fondali marini)

non esprime nessuna opinione personale, poiché il termine "crude" (crudo, grezzo, greggio) assume una valenza diversa.

2. *Associare il sentimento con parole chiave*: A volte risulta difficile localizzare la fonte di questi sentimenti in frasi che esprimono una forte opinione personale, questo fa sì che creare un'associazione con una parola chiave o una frase risulti essere un compito arduo. Preso l'esempio "Every time I read 'Pride and Prejudice' I want to dig her up and beat her over the skull with her own shin-bone" (Ogni volta che leggo 'Orgoglio e Pregiudizio' mi viene voglia di riesumarla e colpirla in testa col suo stesso stinco), con "her" ci si vuole riferire all'autrice del libro, che però non è esplicitamente menzionata. Non si riesce quindi a legare il sentimento negativo con il personaggio.

3. *Dipendenza dal dominio*: Simile al caso della soggettività, una parola o una frase può assumere un valore positivo all'interno di un contesto come la recensione di un film, ma può anche assumere un valore negativo se riferito allo sterzo di un veicolo.

4. *Rilevazione del sarcasmo*: Le frasi sarcastiche (utilizzate molto spesso sul web) esprimono opinioni negative nei confronti di un argomento o di una persona, ma lo fanno utilizzando parole prevalentemente positive. "Nice perfume. You should marinate in it." (Buon profumo. Dovresti starci a mollo.) non presenta nessuna parola o sequenza di parole che esprimano direttamente un parere o opinione negativa in riferimento all'odore del profumo, ma nel loro complesso sono dirette a criticarne la caratteristica.

5. *Espressioni contrastanti*: Ci sono alcune frasi nelle quali solo una parte di esse sono utili per determinare l'opinione globale del

documento. “This movie should be amazing. It sounds like a great plot, the poupolar actors, and the supporting cast is talented as well. However, it can’t hold up.” (Questo film dovrebbe essere fantastico. Sembra avere una buona trama, gli attori famosi e il restante cast è sono molto talentuosi. Comunque sia, non regge.). Utilizzando un semplice approccio di “bag-of-words”, la frase nel complesso verrebbe classificata come prevalentemente positiva, mentre in realtà è prevalentemente negativa.

6. *Negazione indiretta del sentimento*: Il sentimento può essere negato in altri vari modi oltre ai soliti no, mai, etc. Questo tipo di negazioni sono difficili da individuare. Nella frase “It avoids all clichés and predictability found in Hollywood movies.” (Evita tutti quei cliché e quella prevedibilità che si può trovare nel film di Hollywood.), i termini “clichés” e “predictability” portano con loro una valenza negativa, mentre l’uso del verbo “avoids” nega la valenza negativa per conferirne una positiva al soggetto della frase.

7. *Ordine di apparizione*: L’analisi della struttura del discorso in esame è fondamentale per riuscire a estrarre correttamente il sentimento tramite sentimen analysis o opinion mining. “A is better than B / B is better than A” (A è meglio di B / B è meglio di A) è un semplice esempio di come l’ordine con cui i soggetti compaiono nella frase possa capovolgere il significato (e quindi la valenza).

8. *Riconoscimento dell’entità*: Se una frase presenta più entità verso i quali si esprimono opinioni diverse, è necessario prima di tutto separare le singole parti che trattano una specifica entità e poi analizzarle, per riuscire ad estrarre l’opinione a cui questa è collegata. “I hate Microsoft, but I love Linux” (Odio Microsoft, ma

amo Linux) verrebbe categorizzata come frase neutrale da un semplice approccio di tipo "bag-of-words", mentre sono presenti due opinioni con valenze opposte ma per due entità separate tra loro.

9. *Costruire un classificatore "soggettivo vs oggettivo" per i tweet*: La costruzione di un classificatore per la classificazione di un tweet oggettivo/soggettivo è un tema ancora in forte sviluppo.

10. *Gestire la comparazione*: Modelli di tipo "bag-of-words" non gestiscono molto bene la comparazione fra entità: "IIT's are better than most of the private colleges" (Gli IIT sono migliori rispetto a molti altri college privati), il messaggio in questo caso verrebbe considerato prevalentemente positivo sia per gli IIT che per i college privati, poiché i "bag-of-words" non tengono in considerazione la relazione che intercorre con "better".

11. *Internazionalizzazione*: Lo studio della sentiment analysis è svolto prevalentemente in lingua inglese, poiché quella più comune e vicina al modo dell'informatica. Non mancano comunque studi e sviluppi per effettuare sentiment analysis su testi in lingue diverse.

3.5 Alcuni studi

Ad oggi [8] Twitter conta più di 600 milioni di utenti iscritti, di cui 289 milioni condividono ogni giorno contenuti, con una frequenza di più di 9000 tweet al secondo.

Negli anni sono stati eseguiti vari studi sulle diverse modalità di elaborazione dei tweet, utilizzando modelli formali conosciuti nel campo della NLP in generale, ma tentando anche di proporne di nuovi, combinando modelli esistenti, considerando diversi aspetti del problema in esame.

Proponiamo di seguito una breve serie di studi che sono stati svolti più o meno recentemente sulla piattaforma di microblogging di San Francisco.

3.5.1 Metodi

Nel loro studio, Agarwal et al [9] confrontano i risultati ottenuti nell'analisi del sentimento, su un set di tweet recuperati manualmente, usando modelli unigram, tree kernel e feature-based. Successivamente effettuano test anche con una combinazione del modello unigram con quello feature-based e con una combinazione del modello tree kernel con quello feature-based.

Nella fase di pre-processing dei messaggi all'interno del data set, Agarwal et al. presentano anche due nuove risorse che utilizzano per meglio categorizzare e valorizzare la positività o la negatività del sentimento complessivo del tweet: un dizionario di emoticon raccolte da Wikipedia, alle quali viene assegnata una fra cinque etichette (neutrale, positiva, negativa, estremamente positiva e estremamente negativa); un dizionario degli acronimi, ottenuto inserendo manualmente le traduzioni di vari acronimi ottenuti da un sito web [10] spesso utilizzati all'interno dei messaggi. Aspetto interessante da sottolineare, nella fase di pre-processing del testo, è la scelta di sostituire le sequenze di caratteri ripetuti con tre caratteri al posto di due, in modo da tenere in considerazione l'aspetto enfatizzante che quella parola porta con sé.

Il tree kernel è una rappresentazione ad albero dei tweet che rende più semplice confrontare tra loro le componenti caratterizzanti del messaggio e assegnare il valore di polarità complessiva. Per assegnare i valori di polarità ad ogni componente, Agarwal et al hanno fatto utilizzo di una rivisitazione del *Dictionary of Affect Language* (DAL) [12] combinato con WordNet [11].

Il modello feature-based proposto, chiamato da loro *Senti-features*, è un modello di 100 caratteristiche che possono essere raggruppate in tre categorie:

1. Caratteristiche che assumono valori dell'insieme dei numeri naturali, come ad esempio il numero di avverbi che assumono valori positivi o il numero di avverbi che assumono valori negativi.

2. Caratteristiche che assumono valori dell'insieme dei numeri reali, come il singolo valore di polarità di una parola o la somma delle polarità di parole che compongono un'espressione presente nel tweet.

3. Caratteristiche che assumono valori booleani, come la presenza di punti di esclamazione o di testo scritto completamente in maiuscolo.

Queste categorie sono ulteriormente suddivise in "Polar" e "Non-polar", dove per "Polar" si intendono quelle caratteristiche che possiedono un valore di polarità, ottenuto da uno dei dizionari prima elencati.

Un'ulteriore suddivisione è quella che divide le caratteristiche in "POS" o "Others", andando quindi a distinguere ulteriormente quelle caratteristiche che catturano le statistiche delle parole della categoria "parts-of-speech" (POS) e quelle che non fanno parte di quest'ultima.

I confronti tra i tre modelli sono stati ottenuti eseguendo l'analisi dei tweet con due diverse finalità: classificare i tweet in positivi e negativi e classificare i tweet in positivi, negativi e neutrali.

In tutti e due i test, Argwal et al hanno riscontrato che il tree kernel analizza in maniera più precisa i tweet, soprattutto nel secondo caso, dove si considera anche la polarità neutrale. Per quanto riguarda *Senti-features*, i ricercatori hanno concluso che la caratteristica più importante e che ha portato maggior contributo ai risultati è quella che tiene in considerazione la polarità a priori delle POS.

Nel loro studio, Akshi e Teeja [6] utilizzano un approccio ibrido, utilizzando un metodo di tipo corpus-based per estrapolare l'orientamento semantico degli aggettivi e un metodo di tipo dictionary-based per estrapolare l'orientamento semantico di verbi e avverbi. Ottenuti in questo modo i valori di polarità degli elementi caratterizzanti il messaggio, la polarità complessiva del

tweet è stato calcolato utilizzando un'equazione lineare che fa uso degli intensificatori emozionali.

Anche qui, durante la fase di pre-processing vengono tenuti in considerazione particolarità del messaggio, come la percentuale di testo in maiuscolo, la presenza di caratteri ripetuti e punti esclamativi.

Nella fase di calcolo della polarità complessiva del tweet, Akshi e Teeja hanno deciso di raggruppare gli avverbi, gli aggettivi e i verbi in due gruppi, uno composto dagli aggettivi e gli avverbi che qualificano questi ultimi (chiamato gruppo aggettivo), un altro composto dai verbi e dagli avverbi che qualificano questi ultimi (chiamato gruppo verbo). Il valore di polarità di ciascun gruppo è calcolato tramite il prodotto del valore dell'avverbio per il valore dell'altro elemento del gruppo, questo calcolo eseguito per ogni aggettivo del gruppo. Nel caso non sia presente un avverbio nel gruppo, viene assegnato un valore di base di 0.5.

$$S(T) = \frac{(1 + (P_c + \log(Ns) + \log(N_x))/3)}{|OI(R)|} * \sum_{1=1}^{|OI(R)|} S(AG_i) + S(VG_i) + N_{ei} * S(E_i)$$

Formula 9: Calcolo dell'orientamento globale del tweet

La figura sopra riportata mostra la formula utilizzata da Akshi e Teeja. |OI(R)| indica la dimensione del set dei gruppi (aggettivo e verbo) e delle emoticon estratte dal messaggio; P_c indica la percentuale in cui il testo del messaggio è in maiuscolo; N_s indica il numero di lettere ripetute; N_x rappresenta il numero di punti esclamativi; N_{ei} è relativo al numero di volte che la *i*-esima emoticon compare; $S(AG_i)$ indica il valore di polarità dell'*i*-esimo gruppo aggettivo; $S(VG_i)$ indica il valore di polarità dell'*i*-esimo gruppo verbo; $S(E_i)$ indica il valore di polarità dell'*i*-esima emoticon.

P_c, N_s e N_x fanno parte degli intensificatori emozionali, perché rappresentano l'enfasi che si vuole dare al sentimento a cui si riferiscono.

S(T) è un valore che deve essere compreso nell'intervallo [-1,1], se supera questo intervallo viene approssimato al limite più vicino.

3.5.2 Applicazioni

Fino ad ora ci siamo concentrati sulla parte che riguarda l'elaborazione e i metodi per l'estrapolazione e il calcolo della polarità di un tweet, sembra dunque doveroso ora parlare anche di qualche applicazione pratica relativamente all'analisi del sentimento sulla piattaforma di microblogging.

Pecionchin e Usman [13] hanno tentato di trovare una correlazione tra l'umore di una persona e le decisioni sugli investimenti degli investitori, osservando come caso di studio gli indici NASDAQ e il suo volume dell'intermediazione finanziaria.

Il loro data set è composto da tweet recuperati nell'arco di tempo da luglio a dicembre 2013, servendosi dell'aiuto di Archive Team [14] per il recupero delle informazioni in tempo reale. Il confronto è stato poi eseguito sui dati finanziari del NASDAQ aggiornati ogni ora per il tempo di apertura del mercato azionario, sempre il quel periodo di tempo.

Per il calcolo della polarità di ogni tweet si sono basati sul dizionario ANEW [15], andando ad analizzare sei dimensioni della scala emozionale che sono, rispettivamente, "happy", "sad", "excited", "calm", "dominant" e "submissive", andando poi ad aggiustare il valore in base al "peso" [16] che l'autore del messaggio possiede. Con "peso" si intende l'influenza che il proprietario del tweet possiede all'interno della comunità, qualità che deriva dal numero di follower che seguono quell'utente.

Per ottenere i risultati dello studio hanno poi applicato il test della causalità di Granger [17], dove si ipotizza che una serie storica ne anticipi un'altra in un qualche modo e che eventualmente la causi. Grazie ai risultati ottenuti, Pecionchin e Usman hanno dimostrato che i messaggi che ricadono nella dimensione "happy" o "calm" causano, in un arco di tempo che varia tra le 48 e le 55 ore, un alzamento del prezzo di chiusura del NASDAQ, mentre i

messaggi che ricadono nella dimensione “sad” causano un abbassamento del prezzo di chiusura dopo un arco di tempo di approssimativamente 72 ore.

Per quanto riguarda il volume dell’intermediazione finanziaria, l’alzamento è causato da messaggi della dimensione “excited” e l’abbassamento da quelli della dimensione “calm”, andando ad influire dopo un periodo di approssimativamente 62 ore.

Anche Skuza e Romanowski [18] hanno basato il loro studio sulla ricerca di un legame tra quelli che sono gli stati d’animo che la gente esprime tramite messaggi istantanei su Twitter e i valori di mercato delle grandi aziende. Il loro lavoro prende in esame la società Apple Inc., uno dei più grandi colossi.

Per i loro studi, i tweet sono stati recuperati tramite le Streaming API di Twitter, in un periodo che copre i primi tre mesi dell’anno 2013. Le parole chiave utilizzate sono state “Apple”, nome ufficiale con la quale ci si riferisce all’azienda in generale, e “AAPL”, che invece è la sigla utilizzata nei mercati azionari.

I data set utilizzati per il training e il test della classificazione sono stati ottenuti in due modi diversi. Il primo consiste nell’utilizzare SentiWordNet [19] per rilevare in modo automatico il sentimento dei messaggi, il secondo consiste nel definire manualmente la polarità dei messaggi, marcando con l’etichetta positivo, negativo o neutrale.

Grazie alla creazione di questi data set, due classificatori sono stati creati, uno per la rilevazione della soggettività del tweet, l’altro per la classificazione della polarità dei messaggi ottenuti dal primo. Per meglio analizzare i risultati della classificazione, Skuza e Romanowski hanno introdotto la concezione di “sentiment value” come funzione logaritmica in base 10 del rapporto tra tweet positivi su tweet negativi, mentre per stimare i prezzi delle azioni hanno fatto uso di una funzione di regressione lineare dei prezzi passati, utilizzando come peso il “sentiment value” sopra definito.

Per ogni data set sono stati svolti test di predizione del prezzo con intervalli di tempo differenti, cercando di prevedere l’andamento del titolo ogni ora, ogni mezz’ora, ogni 15 minuti e ogni 5 minuti.

Dai risultati ottenuti da questi test, Skuza e Romanowski hanno osservato che il data set contente i tweet filtrati usando il nome del titolo azionario (AAPL) è stato quello più utile nel predire l'andamento azionario, poiché il data set contenente i messaggi filtrati con il nome della società (APPLE) spesso riguardavano argomenti dell'azienda in generale e non propri del mercato azionario.

Altro fattore determinante per effettuare predizioni sulle fluttuazioni è il numero di tweet che fanno riferimento al determinato intervallo di tempo, questo fattore ha impedito di eseguire test utilizzando il data set costruito sul filtro del titolo azionario poiché i messaggi recuperati per questo erano in numero insufficiente.

Skuza e Romanowski, nonostante abbiano evidenziato un legame tra i messaggi condivisi dagli utenti di Twitter e le variazioni dei prezzi dei titoli azionari, sottolineano che la struttura da loro presentata non è in grado di predire le variazioni dovute dai cosiddetti "black swans" [20], metafora utilizzata per indicare eventi che causano un forte impatto ma che non sono prevedibili.

Nel loro lavoro, Chen et al. [21] utilizzano i messaggi recuperati da Twitter all'interno del loro modello per la predizione dei crimini di furto all'interno di una certa area, combinando questi dati con dati storici su crimini passati e le segnalazioni delle temperature nel momento dei crimini.

Anche in questo caso i tweet sono stati recuperati usando le Streaming API di Twitter, filtrando i messaggi geo localizzati nella zona nord e sud della città di Chicago, i quali sono stati analizzati tramite tecniche lexicon-based [23] e aggiustando il valore della polarità di una parola polarizzata, andando ad analizzare le parole che la circondano. Queste parole sono state categorizzate in neutrali, negatori, amplificatori e de-amplificatori che, a seconda della categoria in cui ricadono, modificano il valore della polarità della parola a cui fanno riferimento.

Con il loro studio, Chen et al hanno dimostrato che è possibile riuscire a prevedere in quali zone della città potrebbe verificarsi un crimine, fornendo così un potente mezzo alle forze dell'ordine locali per meglio gestire le risorse a

proprie disposizione, andando ad aumentare il numero di pattuglie presso le zone ritenute più “calde”.

4 Settori correlati

Chiudiamo questo sguardo generale sulla sentiment analysis presentando alcune tematiche molto legate a questo campo di ricerca ma che non sono state trattate nei capitoli precedenti.

4.1 Costruzione di risorse

La costruzione di risorse (in inglese *Building Resources – BR*) ha come obiettivo quello di creare, e mettere a disposizione della comunità scientifica, vocabolari, dizionari e raccolte di documenti all'interno dei quali vengono annotate espressioni che esternano opinioni a seconda della loro polarità.

Non è considerata un sotto compito della sentiment analysis, ma può aiutare a migliorarne le prestazioni quando vengono applicati metodi lexicon-based. Come descritto da Montoyo et al [61], le principali difficoltà che si incontrano in questo ambito sono:

- *Ambiguità delle parole*

- *Multilinguismo:* la necessità che si ha di avere a disposizione risorse in grado di andare a coprire più lingue (come visto anche nel capitolo precedente, il problema dello studio su lingue diverse da quella di lingua inglese è forte).

- *Granularità:* le opinioni possono celarsi all'interno dei diversi livelli del testo, quindi a livello di parola, di un modo di dire o a livello di un'intera frase.

- *La differenza nell'esternazione dell'opinione per ogni tipologia di testo trattato:* può succedere che le opinioni vengano espresse in maniere differenti spostandosi da una piattaforma all'altra.

La costruzione di vocabolari è una tematica presente nel lavoro di Tan e Wu [62], dove propongono un algoritmo di random walk per riuscire ad ottenere un vocabolario orientato al dominio grazie all'utilizzo simultaneo di parole e documenti portatori del sentimento reperiti da vecchi domini e dal dominio corrente. I loro esperimenti, basati sull'utilizzo di tre diversi data set specifici per il dominio, hanno ottenuto risultati promettenti, validando le teorie proposte di migliorare il processo di costruzione automatica di vocabolari.

Di Caro e Robaldo [63] introducono nel loro studio la costruzione di corpus tramite Opinion Mining-ML, un nuovo formalismo basato su XML per l'etichettatura di espressioni testuali che descrivono aspetti ed entità che sono considerati rilevanti nella situazione in esame. Si propone come un nuovo standard a fianco di Emotion-ML e WordNet.

Il loro lavoro si articola in due fasi. Inizialmente presentano una metodologia standard per l'annotazione di asserzioni affettive che non sono dipendenti dal dominio applicativo corrente, e che quindi hanno valenza in senso comune. Successivamente prendono in considerazione l'adattamento specifico per il dominio, che fa affidamento sull'uso dell'ontologia, che è dipendete dal dominio in esame. Nel loro esperimento prendono come caso di studio un data set di recensioni di ristoranti, al quale applicano un processo di estrazione basato sull'interrogazione, valutando poi l'efficienza del progetto proposto tramite l'analisi dettagliata dei disaccordi tra vari annotatori.

I risultati ottenuti mostrano la capacità del sistema di essere uno schema di annotazioni effettivo, in grado di coprire una vasta area di termini complessi,

mantenendo un buon grado di accordo tra le varie persone coinvolte nell'annotazione durante il test.

Steinberger et al [64] presentano invece un approccio semi-automatio per la creazione di dizionari in diverse lingue. Come primo passo hanno realizzato dizionari per due lingue di alto livello, traducendoli successivamente in modo automatico per ottenere il dizionario nella terza lingua. Quelle parole che possono essere trovate in entrambe le liste di parole delle lingue sono più propense ad essere utili, perché il loro significato tende ad essere simile a quello presente all'interno delle due risorse in lingua.

Durante il loro lavoro, Steinberger et al hanno affrontato due problemi. La flessione morfologica e la soggettività coinvolti nell'annotazione umana e nella prova della valutazione. Il loro studio si è basato sul dominio delle notizie.

4.2 Emotion Detection

L'analisi del sentimento è considerato un compito di natural language process per estrapolare le opinioni in merito ad una o più determinate entità. Esistono ambiguità sulla differenza tra opinione, sentimento ed emozione. L'opinione viene definita come il concetto tradizionale che riflette l'atteggiamento, la mentalità nei confronti di una entità, il sentimento riflette i sentimenti o le emozioni, mentre l'emozione riflette l'atteggiamento [65].

Plutchik afferma che ci sono otto emozioni elementari e originali, che sono *gioia, tristezza, rabbia, paura, fiducia, disgusto, sorpresa e aspettativa* [66]. La *Emotion Detection* (ED) può essere considerato un sotto compito della sentiment analysis, dove l'obiettivo di quest'ultima è principalmente quella di riuscire ad assegnare una valenza positiva o negativa alle opinioni, mentre quello dell'emotion detection è quella di riuscire a trovare e ad estrapolare le varie emozioni che l'autore del testo vuole trasmettere.

Come per l'analisi del sentimento, anche il rilevamento delle emozioni può essere implementato tramite l'utilizzo di approcci di machine learning o lexicon-based, e soprattutto questi ultimi sono i più utilizzati in letteratura.

Lu et al [67] hanno affrontato il problema della rilevazione delle emozioni adottando un approccio di text mining web-based su frasi in inglese relative ad un singolo evento. L'approccio è basato sulla distribuzione di probabilità di legami reciproci comuni tra il soggetto e l'oggetto di un evento. Non hanno fatto uso di vaste risorse lessicali, ma hanno mostrato come il loro approccio sia risultato soddisfacente nella rilevazione di emozioni positive, negative e neutrali, sottolineando anche il fatto che questo tipo di problema è molto sensibile al contesto in esame.

Balahur et al [68] hanno utilizzato un approccio combinato di machine learning e lexicon-based, proponendo un metodo basato sul senso comune archiviato nella base di conoscenza del corpus di emozioni EmotiNet [69]. Balahur et al affermano che le emozioni non vengono sempre espresse in modo diretto tramite l'utilizzo di parole dal valore emotivo come ad esempio *felice* o *triste*, ma possono essere espresse tramite la descrizione di situazioni di vita reale, che poi i lettori assumono essere correlati ad un determinato stato emotivo. Per il raggiungimento del loro obiettivo hanno fatto uso di algoritmi SVM, mostrando che l'approccio basato su EmotiNet è il più adeguato per la rilevazione di emozioni all'interno di contesti dove vi è assenza di parole che esprimono direttamente un'emozione.

Affect Analysis (AA) è quella serie di tecniche che mirano al riconoscimento delle emozioni che sono dedotte tramite ragionamento da certe modalità semiotiche.

Ptaszynski et al [73] hanno lavorato sull'AA basato sul testo della narrativa giapponese reperibile su Aozora Bunko [74]. Nella loro ricerca hanno affrontato il problema del riconoscimento delle emozioni dell'individuo o del personaggio nei racconti. Inizialmente hanno estratto il soggetto dell'emozione da una frase in base all'analisi di espressioni anaforiche, per poi stimare, tramite analisi, in quale tipo di stato emotivo ogni personaggio si trovava all'interno di ogni parte del racconto.

Mohammad [75] ha concentrato il suo lavoro nello studio dell'AA nelle e-mail e nei libri. Ha analizzato il corpus di e-mail Enron [76], con il quale ha

dimostrato che esistono differenze ben marcate nell'uso di parole che esternano emozioni nelle e-mail di lavoro tra i diversi sessi. Mohammad ha creato un vocabolario composto da annotazioni manuali delle associazioni di polarità positiva o negativa di una parola e delle otto emozioni base tramite sviluppo collettivo. Ha utilizzato questo vocabolario per analizzare e tracciare la distribuzione delle parole che esprimono emozioni all'interno di libri e di e-mail. Ha inoltre introdotto il concetto di densità di parole emotive tramite studio si romanzi e favole, dal quale studio ha dimostrato che le favole hanno una distribuzione maggiore rispetto ai romanzi.

L'impiego di parole relative ad emozioni può essere utilizzato assieme a tecniche basate sul corpus, come presentato da Keshtkar e Inkpen [70], i quali hanno introdotto un algoritmo di bootstrap [71] basato su feature lessicali e contestuali, per l'identificazione di parafrasi e le estrazioni da esse di termini emotivi. Sono partiti da una lista ristretta di parole (WordNet Affect [72]), riuscendo a far imparare al loro approccio schemi per sei classi di emozioni. Per l'estrazione di parafrasi hanno fatto utilizzo di blog commentati e altri data sets, mentre il loro dominio di utilizzo si è focalizzato su dati reperiti da blog commentati, blog di notiziari e favole. I loro risultati hanno mostrato buone prestazioni per l'algoritmo da loro implementato.

4.3 Transfer Learning

La tecnica denominata di *Transfer Learning* estrae la conoscenza da domini ausiliari, con lo scopo di migliorare il procedimento di apprendimento all'interno del dominio di interesse. Per esempio si può trasferire la conoscenza dai documenti di Wikipedia per analizzare i tweet, o ottenere conoscenza da una ricerca in inglese per migliorare una ricerca in arabo. La transfer learning è considerata una nuova tecnica di apprendimento intersettoriale, poiché affronta i vari aspetti delle differenze di vari domini. È utilizzata per migliorare alcuni compiti di text mining come la classificazione del testo, l'analisi del sentimento ed altri. [1, 77,78].

Nell'ambito della sentiment analysis, tecniche di transfer learning posso essere utilizzare per trasferire le conoscenze ottenute dalla classificazione del sentimento da un dominio ad un altro, o per costruire un ponte tra i due [79, 80].

Tan e Wang [79] propongono, nel loro lavoro, un algoritmo basato sull'entropia per estrarre feature specifiche del dominio con un'alta frequenza (*high-frequency domain-specific – HFDS*), così come un modello di pesatura che pesa le feature trovate e le singole istanze. Hanno assegnato un peso minore alle feature trovate, mentre hanno assegnato un peso maggiore alle istanze con la stessa etichetta della feature utilizzata come pivot. Il loro studio si è concentrato sulle recensioni di computer, di mercato e sull'educazione provenienti da un data set cinese specifico del dominio. Dai loro risultati hanno provato che il modello da loro proposto può superare l'influenza sfavorevole delle feature HFDS, così come hanno mostrato che il modello è un'ottima scelta per le applicazioni di analisi del sentimento che necessitano di una classificazione ad alta precisione, ma che hanno pochi dati etichettati utilizzabili per la fase di addestramento.

Wu e Tan [80] hanno proposto una struttura per la classificazione del sentimento a due stadi. Nel primo stadio hanno costruito un ponte tra il dominio sorgente e il dominio di interesse, con l'intento di ottenere qualche documento etichettato più affidabile per il dominio di interesse. Nel secondo stadio hanno sfruttato la struttura intrinseca, rivelata dai documenti etichettati ottenuti, per riuscire ad etichettare i dati del dominio su cui lavoravano. Come per Tan e Wang, i dati per i loro esperimenti sono stati recuperati da data set cinesi specifici del dominio, che riguarda le recensioni di libri, hotel e notebook.

La diversità tra le varie fonti da cui si recuperano i dati è un problema che riguarda la modellazione della congiunzione tra diverse fonti di dati. Gupta et al [81] hanno provato a risolvere questo problema, poiché la modellazione di questa congiunzione è uno dei temi più importanti della transfer learning. Nel loro lavoro, Gupta et al propongono una struttura di sottospazio di apprendimento regolarizzata e condivisa, che può sfruttare la mutua forza delle fonti di dati connesse, restando nel contempo inalterata dagli effetti dovuti dalla variabilità di ogni risorsa. Nel loro lavoro si sono concentrati sulle notizie dei

social media reperiti da siti famosi come Blogspot, Flickr, Youtube, CNN e BBC. Dai risultati del loro studio hanno dimostrato che il loro approccio ha portato prestazioni migliori rispetto ad altri approcci del settore.

Conclusioni

Concludiamo la tesi presentata riassumendo quello che è stato presentato nei vari capitoli.

Nel primo capitolo abbiamo dato una breve panoramica di quello che è il Web, agli inizi e nei nostri giorni, ovvero un enorme contenitore di informazioni (che per la maggior parte è in forma testuale) dal quale si può estrarre ed analizzare la conoscenza ivi contenuta tramite tecniche di analisi ed elaborazione dei dati. Abbiamo quindi presentato quello che è il Data Mining e un suo sotto ramo, il Text Mining, come un'insieme di metodologie per affrontare i compiti appena descritti.

Successivamente abbiamo presentato quello che è l'argomento principale della trattazione in corso, la Sentiment Analysis, quella serie di tecniche e metodi utilizzati con lo scopo di analizzare porzioni di testo, di dimensioni più o meno contenute, estraendo da essi le opinioni che l'autore del documento prova nei confronti di uno o più degli argomenti che ha trattato. Il caso più generale di categorizzazione dell'opinione è quello di prendere in considerazione solamente un aspetto binario del sentimento, ovvero se l'autore sta esternando un'opinione positiva o negativa verso l'oggetto del suo discorso. Abbiamo visto a che livelli può essere effettuata l'analisi, gli approcci più comuni di Feature Selection per l'identificazione e lo sfruttamento di particolari qualità del testo e di Sentiment Classification, con l'obiettivo di riuscire a fornire una valutazione complessiva della polarità dei documenti analizzati. È stato infine presentato uno dei metodi più classifici per valutare le prestazioni di un classificatore.

Nel terzo capitolo abbiamo dato uno sguardo più approfondito del tema trattato, andando a prendere in considerazione una delle piattaforme web più sfruttate in questo campo, Twitter. In questo capitolo abbiamo presentato velocemente quelle che sono le proprietà principali di questa piattaforma e i metodi esistenti per il recupero dei messaggi condivisi all'interno di questo social, considerando sia le API da poter utilizzare all'interno di applicazioni che risorse web, siti internet dai quali è possibile effettuare interrogazioni per la ricerca di particolari tweet. Abbiamo poi esposto alcuni studi fatti in merito alle tecniche per classificare le opinioni estrapolate dai tweet, seguiti da alcuni esempi di applicazione di tecniche di opinion mining sulla piattaforma.

Abbiamo infine chiuso la trattazione di questa tesi parlando di alcuni dei campi di ricerca strettamente correlati alla sentiment analysis, quali l'identificazione delle emozioni all'interno dei testi, la costruzione di risorse per un'applicazione più efficiente delle tecniche basate sul lessico e il trasferimento della conoscenza da un dominio all'altro, in modo da aiutare il procedimento di apprendimento all'interno del dominio di interesse.

I classificatori Naive Bayes e le macchine a vettori di supporto (SVM) sono gli algoritmi di machine learning più frequentemente utilizzati per la classificazione del sentimento, considerati come modelli di riferimento per la comparazione con nuovi algoritmi.

L'analisi del sentimento in lingue diverse da quella inglese ha attratto molti ricercatori negli ultimi anni che hanno provato ad estrapolare e a classificare i testi in italiano, francese, tedesco, cinese, giapponese, arabo etc. Nonostante il forte e crescente interesse per idiomi diversi da quello inglese, la disponibilità delle risorse su cui effettuare sentiment analysis ed opinion mining (parliamo sia di testi, che di vocabolari/dizionari) è in numero molto inferiore, motivo per cui la lingua inglese è ancora quella più studiata.

Essendo un campio di studio piuttosto recente, non si è ancora in possesso di tecniche ed algoritmi assolutamente perfetti per la determinazione dell'opinione di uno o più individui in merito ad uno o più argomenti. Numerosi sono gli studi e gli articoli, accademici e non, che hanno trattato questo problema

sotto vari aspetti, e sicuramente tanti altri e molti di più saranno gli studi eseguiti in questo campo per fini che possono essere quello commerciale (si pensi alle recensioni di prodotti), politico (previsione delle elezioni) o sociale (prevenzione di crimini, di cui è solo un esempio).

Bibliografia

[1] Pang B., Lee L., Opinion Mining and Sentiment Analysis, 2008

[2] Wikipedia, https://it.wikipedia.org/wiki/Twitter

[3] Twitter privacy policy, https://twitter.com/privacy?lang=it

[4] Twitter APIs, https://dev.twitter.com/overview/documentation

[5] Twitter API, Streaming API, POST statuses / filter : https://dev.twitter.com/streaming/reference/post/statuses/filter

[6] Akshi K., Teeja M.S., Sentiment Analysis on Twitter, *International Journal of Computer Science Issues*, Vol. 9, Issue 4, No. 3, Luglio 2012

[7] Kharde V.A., Prof Sonawane S.S., Sentiment Analysis of Twitter Data : A Survey of Techniques, 26 Gennaio 2016. Articolo disponibile a : http://arxiv.org/abs/1601.06971

[8] Twitter statistics : http://www.statisticbrain.com/twitter-statistics/

[9] Agarwal A., Xie B., Vovsha I., Ranbow O., Rebecca Passonneau, Sentiment Analysis of Twitter Data, *Proceedings of the Workshop on Language in Social Media* (LSM 2011), pages 30–38, Portland, Oregon, 23 June 2011

[10] Internet Slang Dictionary & Translation : http://www.noslang.com/

[11] WordNet : https://wordnet.princeton.edu/

[12] Dictionary of Affect Language, Whissell C. : https://www.shaktitechnology.com/whissel-dictionary-of-affect/index.htm

[13] Pecionchin M, Usman M., Data Mining Twitter To Predict Stock Market Movements, 23 febbraio 2015, disponibile a : https://ideas.repec.org/a/nos/ycriat/192.html

[14] Archive Team : https://archive.org/details/twitterstream

[15] Bradley M.M., Lang P.J., Affective Norms for English Words (ANEW) : Instruction Manual and Affective Ratings, 1999, disponibile a : http://goo.gl/TzT2nu

[16] Brin. S, Page L., The PageRank Citation Ranking : Bringing Order to the Web, 1999, disponibile a : http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf

[17] Causalità di Granger : https://it.wikipedia.org/wiki/Causalità_di_Granger

[18] Skuza M., Romanowski A., Sentiment Analysis of Twitter Data within Big Data Distributed Environment for Stock Prediction, *Proceedings of the Federated Conference on Computer Science and Information System*, 13-16 settembre 2015, pp. 1349-1354.

[19] Esuli A., Sebastiani F., SentiWordNet: A publicly available lexical resource for opinion mining, *Proceedings of the 5th Conference on Language Resources and Evaluation* (LREC'06, pages 417–422, 2006.

[20] Teoria del Cigno Nero: https://it.wikipedia.org/wiki/Teoria_del_Cigno_Nero

[21] Litton I., #TwitterCritic: Sentiment Analysis of Tweets to Predict TV Ratings, giugno 2015, disponibile a: http://digitalcommons.calpoly.edu/statsp/52/

[22] Chen X., Cho Y., Jang S., Gerber M.S., Crime Prediction Using Twitter Sentiment and Weather, *Systems and Information Engineering Design Symposium (SIEDS)*, pp. 63-68, 24 aprile 2015.

[23] Hu M., Bing L., Mining Opinion Features in Customer Reviews, 2004, disponibile a: http://www.aaai.org/Papers/AAAI/2004/AAAI04-119.pdf

[24] Sentiment140: http://www.sentiment140.com/

[25] StreamCrab: http://www.streamcrab.com/

[26] Sentiment viz: https://www.csc.ncsu.edu/faculty/healey/tweet_viz/tweet_app/

[27] Cosenza V., Cos'è la Sentiment Analysis? 2012, disponibile a: http://vincos.it/2012/08/30/cose-la-sentiment-analysis

[28] Treccani, sentimento: http://www.treccani.it/vocabolario/sentimento/

[29] Treccani, opinione: http://www.treccani.it/vocabolario/opinione/

[30] Liu B., Zhang L., A survey of opinion mining and sentiment analysis, in *Mining Text Data*, pp. 415-463, Springer, 2012

[31] Medhat W., Hassan A., Korashy H, Sentiment analysis algorithms and applications: A survey, *Ain Shams Engineering Journal (2014) 5*, pp 1093-1113, 2014

[32] Aggarwal C.C., Zhai C.X., Mining Text Data, Springer Science + Business Media, LLC'12; 2012.

[33] Bag of Words: https://it.wikipedia.org/wiki/Modello_della_borsa_di_parole

[34] Stemming: https://it.wikipedia.org/wiki/Stemming

[35] Wikipedia, Teoria dell'informazione: https://it.wikipedia.org/wiki/Teoria_dell%27informazione

[36] Deerwester S., Dumais S., Landauer T.K., Furnas G.W., Harshman R., Indexing by Latent Semantic Analysis, *Journal of the American Society Information Science,* 41.6:pp. 391-407, 1 settembre 1990

[37] Jolliffe I.T., Principal Component Analysis, Springer, 2002

[38] Duric A., Song F., Feature selection for sentiment analysis based on content and syntax models, *Decision Support System*, 53:pp. 704-711, 2012

[39] Wikipedia, Mixture model: https://en.wikipedia.org/wiki/Mixture_model

[40] Wikipedia, Generative model: https://en.wikipedia.org/wiki/Generative_model

[41] Aizerman M., Braverman E., Rozoner L., Theoretical foundations of the potential function method in patter recognition learning, *Automation and Remote Control,* pp. 821-837, 1964

[42] Quinlan J.R., Induction of Decision Trees, *Machine Learning*, Volume 1, Issue 1, pp. 81-106

[43] Chakrabarti S., Roy S., Soundalgekar M.V., Fast and accurate text classification via multiple linear discriminant projections, *The VLDB Journal*, Volume 12, Issue 2, pp. 170-185, agosto 2003

[44] Lewis D.D., Ringuette M., A comparison of two learning algorithms for text categorization, *Symposium on Document Analysis and IR*, 1994

[45] Liu B., Hsu W., Ma Y., Integrating classification and association rule mining, *ACM KDD conference*, 1998

[46] Ko Y., Seo J., Automatic text categorization by unsupervised learning, *Proceedings of COLING-00, 18th international conference on computational linguistics*, 2000

[47] Xianghua F., Guo L., Yanyan G., Zhiqiang W., Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon, *Knowledge-Based Systems*, Volume 37, pp. 186-195, gennaio 2013

[48] Martín-Valdivia M.T., Martínez-Cámara E., Perea-Ortega J.M., Ureña-López L.A., Sentiment polarity detection in Spanish reviews combining supervised and unsupervised approaches, *Expert Systems with Applications*, Volume 40, Issue 10, pp. 3934-3942, agosto 2013

[49] Miller G.A., Beckwith R., Fellbaum C., Gross D., Miller K.J., Introduction to WordNet: an On-Line Lexical Database, *Oxford International Journal of Lexicography*, Volume 3, Issue 4, pp. 235-244, 1990

[50] Hatzivasiloglou V., McKeown K., Predicting the semantic orientation of adjectives, *Proceedings of the 35th annual meeting of the Association for Computational Linguistics*, pp.174-181, 1997

[51] Turney P.D., Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews, *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 417-424, 2002

[52] Fharni A., Klenenr M., Old wine or warm beer: target-specific sentiment analysis of adjectives, *Proceedings of the Symposium on Affective Language in human and machine*, Volume 2, pp. 60-63, 2008

[53] Kim S.M., Hovy E., Determining the sentiment of opinions, *Proceedings of the 20th international conference on Computational Linguistics*, Articolo n° 1367, 2004

[54] Zhang W., Xu H., Weakness finder: find product weakness from Chinese reviews by using aspects based sentiment analysis, *Expert Systems with Applications*, Volume 39, Issue 11, pp. 10283-10291, 1 settembre 2012

[55] Wikipedia, Morpheme-based morphology: https://en.wikipedia.org/wiki/Morphology_(linguistics)#Morpheme-based_morphology

[56] Moreo A., Romero A., Casto J.L., Zurita J.M., Lexicon-based Comments-oriented News Sentiment Analyzer sistem, *Experts Systems with Applications*, Volume 39, Issue 10, pp. 9166-9180, agosto 2012

[57] Di Caro L., Grella M., Sentiment analysis via dependency parsing, *Computer Standards & Interfaces*, Volume 35, Issue 5, pp. 442-453, settembre 2013

[58] Wille R., Restructuring lattice theory: an approach based on hierarchies of concepts, 1982

[59] Priss U., Formal concept analysis in information science, *Annual Review of Information Science and Technology*, 2006

[60] Li S.T., Tsai F.C., Noise control in document classification based on fuzzy forma concept analysis, *IEEE International Conference on Fuzzy SystemI*, pp. 2583-2588, giugno 2011

[61] Montoyo A., Martínez-Barco P., Balahur A., Subjectivity and sentiment analysis: an overview of the current state of the area and envisaged developements, *Decision Support System*, pp. 675-679, 2012

[62] Tan S., Wu Q., A random walk algorithm for automatic construction of domain-oriented sentiment lexicon, *Expert System with Applications*, pp. 12094-12100, 2011

[63] Robaldo L., Di Caro L., OpinionMining-ML, *Computer Standard & Interfaces*, pp. 454-469, 2013

[64] Streinberger J., Ebrahim M., Ehrmann M., Hurriyetoglu A., Kabadjov M., Lenkova P., Steinberger R., Tanev H., Vázquez S., Vanni Z., Creating sentiment dictionaries via triangulation, *Decision Support Systems*, pp. 689-694, 2012

[65] Tsytsarau M., Palpanas T., Survey on mining subjective data on the web, *Data mining and Knowledge discovery*, Volume 24, Issue 3, pp. 478-514, maggio 2012

[66] Plutchik R., A general psychoevolutionary theory of emotion, *Emtion Theory, Research and Experience, Volume 1, Theories of Emotions*, capitolo 1, 1980

[67] Lu C.Y., Lin S.H., Liu J-C., Cruz-Lara S., Hong J-H., Automatic event-level textual emotion sensing using mutual action histogram between entities, *Expert Systems with Applications*, Volume 37, Issue 2, pp. 1643-1653, marzo 2010

[68] Balahur A., Hermida J.M., Montoyo A., Detecting implicit expressions of emotion in a text: A comparative analysis, *Decision Support System*, Volume 53, Issue 4, pp. 742-753, novembre 2012

[69] Balahur A., Hermida J.M., Montoyo A., Muñoz R., EmotiNet: a knowledge base for emotion detection in text built on the appraisal theories, *Proceedings of the 16th international conference on Natural language processing and information systems*, pp. 27-39, 2011

[70] Keshtkar F., Inkpen D., A bootstraping method for extracting paraphrases of emotion expressions from texts, *Computational Intelligence*, Volume 29, Issue 3, pp. 417-435, agosto 2013

[71] Wikipedia, Metodo Bootstrap: https://it.wikipedia.org/wiki/Metodo_bootstrap

[72] WordNet Affect: http://wndomains.fbk.eu/wnaffect.html

[73] Ptaszynski M., Dokoshi H., Oyama S., Rzepka R., Kurihara M., Araki K., Momouchi Y., Affect analysis in context of characters in narratives, *Expert Systems with Applications*, Volume 40, Issue 1, pp. 168-176, gennaio 2013

[74] Wikipedia, Aozora Bunko: https://en.wikipedia.org/wiki/Aozora_Bunko

[75] Mohammad S.M., From once upon a time to happily ever after: tracking emotions in mail and books, *Decision Support Systems*, Volume 53, Issue 4, pp. 730-741, novembre 2012

[76] Enron email dataset: https://www.cs.cmu.edu/~./enron/

[77] Thorsten J., Learning to classify text using support vector machines: methods, theory and algorithms, Kluwer Academic Publishers Norwell, MA, USA 2002

[78] Zhang T., Johnson D., A robust risk minimization based named entity recognition system, *Presented at the seventh conference on Natural language learning at HLT-NAACL*; 2003

[79] Tan S., Wang Y., Weighted SCL model for adaptation of sentiment classification, *Expert Systems with Applications*, Volume 38, Issue 8, pp. 10524-10531, agosto 2011

[80] Qiong W., Tan S., A two-stage framework for cross-domain sentiment classification, *Expert Systems with Applications*, Volume 38, Issue 11, pp. 14269-14275, ottobre 2011

[81] Gupta S.K., Phung D., Adams B., Venkatesh S., Regularized nonnegative shared subspace learning, *Data Mining and Knowledge Discovery*, Volume 26, Issue 1, pp. 57-97, gennaio 2013

[82] Wikipedia, World Wide Web: https://it.wikipedia.org/wiki/World_Wide_Web

[83] Tagliavacche M., La storia del World Wide Web in breve, disponibile a: http://www.webhouseit.com/la-storia-del-world-wide-web-in-breve/

[84] Wikipedia, Web 2.0: https://it.wikipedia.org/wiki/Web_2.0

[85] Fayyad U., Piatetsky-Shapiro G., Smyth P., From Data Mining to Knowledge Discovery in Databases, *Al Magazine*, Volume 17, Numero 3, novembre 1996

[86] Unstructured data and the 80 percent rule, disponibile a: http://breakthroughanalysis.com/2008/08/01/unstructured-data-and-the-80-percent-rule/