
**UNIVERSITA' DEGLI STUDI DI PADOVA**
**DIPARTIMENTO DI SCIENZE ECONOMICHE E AZIENDALI**
**"MARCO FANNO"**

**CORSO DI LAUREA IN ECONOMIA INTERNAZIONALE**
***L-33 Classe delle lauree in SCIENZE ECONOMICHE***


Tesi di laurea

# SENTIMENT ANALYSIS: CONCETTI ED APPLICAZIONI

## *SENTIMENT ANALYSIS: CONCEPTS AND APPLICATIONS*


Relatore:
Prof. TUSSET GIANFRANCO

Laureando:
CAMPANER ANDREA



Anno Accademico 2018-2019

# INDICE

# INTRODUZIONE

Nel seguente elaborato viene trattato il tema della *sentiment analysis*, un approccio allo studio delle opinioni di consumatori ed utenti nato nell'era della digitalizzazione e dei social media. Questo approccio nasce non da un'unica disciplina, ma da una pluralità di aree disciplinari, aspetto che lo rende altresì interessante.

Questa esposizione è organizzata come segue:

- Nel primo capitolo si iniziano ad acquisire i concetti base che caratterizzano e definiscono la *sentiment analysis*, mettendone in luce non soltanto gli aspetti definitori ma anche le applicazioni.
- Nel secondo capitolo vengono esposte ed analizzate alcune problematiche principali riguardanti la *sentiment analysis* di cui bisogna essere consapevoli qualora si decida di adottare questo strumento di ricerca. Fra gli aspetti analizzati le comparazioni di opinioni, il limite dell'*opinion spamming* e relative tipologie di spam.
- Nel terzo capitolo viene presentata un'applicazione della *sentiment analysis* al mercato delle automobili olandesi.

# I

# LUNGO IL SENTIERO DELLA SENTIMENT ANALYSIS

La caratteristica principale nelle scelte che quotidianamente un agente economico si trova ad affrontare è la condizione di incertezza. Un metodo indispensabile che da sempre viene utilizzato per ridurre i costi associati a tale incertezza è l'acquisizione di informazioni e soprattutto ricavare opinioni altrui riguardo, ad esempio, su determinate tipologie di prodotti da acquistare o su un candidato alle elezioni da votare. Allo stesso modo un'azienda potrebbe rispondere ai feedback dei clienti o aspiranti tali modificando i messaggi di marketing, il *brand positioning* e lo sviluppo dei prodotti [Zabin e Jeffries, 2008] così come un candidato politico modellare la propria proposta puntando sui temi più cari agli elettori ed evitare i più insidiosi. A ottenere tutte le opinioni, processarle e sintetizzarle è quel attuale campo di studi chiamato *sentiment analysis* o *opinion mining* (lett. Estrazione di opinioni), una sfida della ricerca scientifica tra diverse discipline. Le opinioni e concetti associati come emozioni, attitudini, valutazioni e sentimenti sono l'oggetto principale degli studi della *sentiment analysis* ma il loro studio è anche legato alla crescita rapida e repentina dei social media (forum, discussioni, blog, recensioni e soprattutto social network). I social media sono "piattaforme virtuali che permettono di creare, pubblicare e condividere contenuti, i quali, a loro volta, sono generati dagli utenti" [Yu e Kak, 2012] ed è proprio l'assenza di barriere nella pubblicazione di contenuti a distinguerla dai media tradizionali. Nel gruppo dei social media fanno parte anche i social network, un luogo virtuale in cui utenti specifici sono collegati tra loro e possono interagire tra loro tramite messaggi privati, commenti pubblici e like.

La ricerca sul tema della *sentiment analysis* è di rilevante importanza se volta a risolvere i problemi tecnici vincolati dall'estrazione, sintesi ed analisi dei dati, ma allo stesso tempo ha, in casi specifici, persino l'obiettivo di prevedere fenomeni sociali. Nonostante questa analisi sia, da un punto di vista strettamente tecnico, una ricerca più vicina al campo dell'ingegneria dell'informazione (*natural processing learning* abbreviato in NLP), le sue applicazioni trovano maggior riscontro in campi come il marketing, *il costumer relations management* e le scienze sociali in generale per poter assumere decisioni in maniera intelligente. La ricerca in ogni caso, persino in ambito ingegneristico, ha bisogno di essere supportata da materie più umanistiche quali la linguistica e la psicologia, che servono a

integrare e coadiuvare la stessa. È per questo che la *sentiment analysis* può essere considerata uno dei più importanti approcci moderni per realizzare analitica sociale [Pedrycz e Chen, 2016].

Obiettivo di questo elaborato, siccome l'argomento pullula di discussioni a riguardo, è chiarire il perché è dedicata tanta attenzione sull'argomento da dipartimenti scientifici differenti ed evidenziare i limiti e i vantaggi che si possono trarre per il futuro.

## ASPETTI DEFINITORI E CENNI STORICI

La *sentiment analysis* è quel campo di studi che analizzale opinioni, i sentimenti, le valutazioni, gli apprezzamenti, le attitudini e le emozioni delle persone in merito a entità come prodotti, servizi, organizzazioni, individui, problemi, eventi e i loro attributi [Liu, 2012]. Nel settore industriale il termine *sentiment analysis* è usato molto più comunemente, ma nel mondo universitario sono impiegati frequentemente sia il termine *sentiment analysis* che *opinion mining*. Il termine *sentiment analysis* forse apparve per la prima volta in [Nasasuka e Yi, 2003], e il termine *opinion mining* apparve per la prima volta in [Dave, Lawrence e Pennock, 2003]. Comunque, la ricerca su *sentiment* e *opinions* apparve prima [Das e Chen, 2001; Morinaga et al., 2002; Pang, Lee e Vaithyanathan, 2002; Tong, 2001; Turney, 2002; Wiebe, 2000]. In questo libro infatti i due termini vennero usati intercambiabilmente, anche se i concetti non sono equivalenti, li distingueremo solo ove necessario. Va in ogni caso sottolineato che estrarre un'opinione da un qualsiasi contesto significa, letteralmente parlando, rilevare una polarità positiva, neutra o negativa mentre se si cerca di analizzare un sentimento è necessario capire a quale emozione esso può essere ricondotto.

Anche se la linguistica e il *natural language processing* (NLP) hanno una lunghissima storia alle spalle, piccole ricerche vennero fatte riguardo l'opinione delle persone e i loro sentimenti prima del nuovo millennio. Da allora, il campo divenne un'area di ricerca molto attiva. Ci sono molte ragioni per questo:

- Prima di tutto esso gode di un vasto ordine applicativo, ciò almeno in ogni dominio. L'industria si circonda della *sentiment analysis* affinché proliferi il commercio delle applicazioni, spingendo in tal modo per una forte motivazione alla ricerca.

- Al secondo punto, essa offre molte sfide che tendano a documentare i problemi, le quali non sono mai state approfondite da un punto di vista di studio prima d'ora. Questo libro vuole sistematicamente definire e discutere tali argomentazioni nonché lo stato corrente, da una situazione prettamente tecnica, per risolverli.
- In terza istanza si tratta di un punto di svolta nella storia umanitaria, noi ora abbiamo un'ingente quantità di opinioni e informazioni dai social media e dal Web. Senza di questi, molta della ricerca, non sarebbe stata possibile.
  Non sorprende la concezione e la rapida crescita della *sentiment analysis* che conclude coincidendo con le statistiche previste.
  Hence ricerca all'interno dell'argomento non solo un importante impatto sulla NLP, ma approfondisce tali studi concentrandosi sulle scienze manageriali, sulle scienze politiche, economiche e sociali che sono affette dall'opinione delle persone.
  Nonostante la maggior parte di ricerche riguardanti la materia in discussione cominci agli inizi del 2000, esistono alcuni lavori di interpretariato di metafore, punti di vista, soggettività e affetti più recenti [Hatzivassiloglou e McKeown, 1997; Hearst, 1992; Wiebe, 1990; Wiebe 1994; Wiebe, Bruce e O'Hara, 1999].

  Le applicazioni della *sentiment analysis* si sono espanse a qualsiasi possibile settore, dai prodotti per i consumatori, servizi, sanità e servizi finanziari a eventi sociali ed elezioni politiche. Distaccata dalla vita reale delle applicazioni, molte di queste sono orientate nella ricerca giornalistica già pubblicata. Per esempio, in [Liu et al., 2007] è stato proposto un modello di predizione della performance delle vendite. In [McGlohon, Glance e Reiter, 2010] si ha una revisione del grado dei prodotti nei mercati. In [Hong e Skiena, 2010] la relazione tra la NFL si indirizza sulla linea delle pubbliche opinioni nei blog e su Twitter. In [Tumasjan et al., 2010], invece, nel medesimo anno si esprime che la *sentiment analysis* venne adottata in Twitter anche per predire i risultati delle elezioni. In [Chen et al., 2010] venne adottato nuovamente per studiare il punto di vista dell'ambito politico. In [Yano e Smith, 2010] venne riportato un metodo adatto a predire il volume di pubblicazione a livello prettamente politico sui siti internet. In [Asur e Huberman, 2010; Joshi et al., 2010, Sadikov, Parameswaran e Venetis, 2009] gli approfondimenti sui dati Twitter, le revisioni dei film e dei blog vennero utilizzati per avere

predizioni in merito ai possibili guadagni al box-office. In [Miller et al., 2011] venne adottato il flusso di sentimenti anche a fini investigativi nei social. In [Mohammad e Yang, 2011] vennero usati i sentimenti per trovare nelle mail la differenza di sesso a livello emozionale scindendo le categorie. In [Mohammad, 2011] le emozioni nelle novelle e nelle storie per bambini furono rintracciate ed analizzate. In [Bollen, Mao e Zeng, 2011] viene tenuto in considerazione il comportamento adottato da Twitter per prevedere le vendite di mercato. In [Bar-Haim et al., 2011; Feldman et al.,2011] esperti investigatori in microblogs furono identificati e la *sentiment analysis* garantì un miglioramento. In [Zhang e Skiena, 2010] i sentimenti tra blog e notizie furono adoperati per studiare strategie di trading. In [Sakunkoo e Sakunkoo, 2009] si studiò l'influenza sociale nei libri online e nelle loro revisioni. In [Groh e Hauffa, 2011] si usò la *sentiment analysis* per approfondirne il contenuto nelle relazioni sociali. Un comprensivo sistema di *sentiment analysis* è stato riportato anche in [Castellanos et al., 2011].

LIVELLI DI ANALISI

Ora faremo una breve introduzione sui principali problemi di ricerca basati su diversi livelli di granularità delle ricerche esistenti. In generale, la *sentiment analysis* è stata studiata principalmente a tre livelli:

- **Documento:** l'incarico in questo livello è quello di classificare se l'intero documento esprime un'opinione positiva o negativa [Pang, Lee e Vaithyanathan, 2002; Turney, 2002]. Per esempio, data la recensione di un prodotto, il sistema determina se la recensione esprime complessivamente un giudizio positivo o negativo riguardante il prodotto in discussione. L'incarico è comunemente conosciuto come *document-level sentiment classification*. Questo livello di analisi assume che ogni documento esprime opinioni su singole entità (ad esempio un singolo prodotto). Perciò, non è applicabile a documenti che comparano entità diverse tra loro.
- **Frase:** l'incarico a questo livello si focalizza sulle frasi e determina se ogni frase esprime un'opinione positiva, negativa o neutra. Neutrale solitamente significa che non viene espressa alcun giudizio. Questo livello di analisi è strettamente collegato alla classificazione soggettiva [Wiebe, Bruce e O'Hara,

1999], che distingue le frasi oggettive, che esprimono informazioni che si attengono ai fatti, dalle frasi soggettive, che esprimono visioni e opinioni soggettive. Comunque, possiamo notare che la soggettività non è equivalente al sentimento così come una frase oggettiva può implicare opinioni, ad esempio "Abbiamo comprato la macchina il mese scorso e il tergicristallo si è staccato". I ricercatori hanno anche analizzato le proposizioni, ma il livello della proposizione non è ancora abbastanza, ad esempio "Apple sta facendo molto bene in questa economia disgustosa".

- **Entità e aspetti:** sia l'analisi a livello del documento che quella a livello della frase non scoprono esattamente cosa piace o non piace alle persone. Il livello dell'aspetto è l'analisi più accuratamente approfondita. Questo livello prima era chiamato livello della caratteristica [Hu e Liu, 2004]. Invece di guardare ai linguistici (documenti, paragrafi, frasi), questo livello guarda direttamente all'opinione stessa, poiché è basato sull'idea che un'idea consiste in un *sentiment* (positivo o negativo) e un target d'opinione. Un'opinione senza il suo target identificato è di uso limitato. Capendo l'importanza dell'uso del target si può anche capire meglio il problema della *sentiment analysis* in maniera più approfondita. Per esempio, nonostante la frase "anche se il servizio non è stato ottimo, mi piace ancora quel ristorante" chiaramente ha un tono positivo, non possiamo affermare che la frase sia interamente positiva, infatti, la frase ha enfasi positiva per quanto riguarda il ristorante, ma negativa riguardo al servizio. In molte applicazioni, i target delle opinioni sono descritti da entità e/o differenti aspetti. Pertanto, l'obiettivo di questo livello di analisi è scoprire sentimenti sulle entità e/o i loro aspetti.

## IL LESSICO DEL SENTIMENT E PROBLEMI INERENTI

I più importanti indicatori di emozioni sono le *sentiment words*, anche chiamate *opinion words*. Queste sono parole che sono comunemente usate per esprimere giudizi, opinioni ed emozioni positivi o negativi. Ad esempio, bello, meraviglioso e fantastico sono parole che esprimono un sentimento positivo, mentre cattivo, povero e terribile ne esprimono uno negativo. Al di là delle parole, ci sono anche frasi ed idiomi che rappresentano ovviamente uno strumento per la *sentiment analysis*, anche se il semplice uso di queste non è affatto sufficiente in quanto il problema stesso è molto più complesso, in altre parole

possiamo dire che questo lessico è necessario ma non sufficiente per questo tipo di analisi. Di seguito sono sottolineati alcuni dei problemi inerenti al lessico:

- Una parola che esprime un sentimento positivo o negativo può avere connotazioni opposte in differenti ambiti di applicazione. Ad esempio, il termine "pauroso" può avere significato sia positivo che negativo: nella frase "quel pallavolista ha un fisico pauroso" si denota una connotazione che fa apprezzare il fisico del pallavolista, mentre nella frase "quel film ha un finale pauroso" l'accezione che prende il termine fa trapelare negatività.
- Una frase contenente delle *sentiment words* può non esprimere alcuna emozione, infatti questo fenomeno accade frequentemente in moltissimi tipi di frase. Nelle frasi interrogative o con il modo condizionale non è raro trovare questa situazione: "Può dirmi quale fotocamera della Sony fa delle belle foto?" oppure "Se riesco a trovare una buona fotocamera della Sony, la compro". Entrambe contengono lo stesso concetto, ma in nessuna delle due viene espresso un commento positivo o negativo su una fotocamera specifica. Come è ovvio che sia, ci sono anche le frasi interrogative o condizionali che esprimono un'idea: "Qualcuno sa dirmi dove posso riparare quella fotocamera maledetta?" oppure "Se stai cercando un'ottima macchina, comprati la Volkswagen T-Roc".
- Il sarcasmo all'interno di frasi con o senza *sentiment words* è molto difficile da decifrare, per esempio "Che macchina fantastica! Ha smesso di funzionare dopo due giorni". Non è frequente il suo uso nelle recensioni dei consumatori riguardo prodotti o servizi, ma è molto frequente in discussioni politiche e ciò rende difficile capire cosa pensano realmente gli elettori.
- Ci sono anche espressioni che non contengono *sentiment words* ma esprimono opinioni. Molte di queste sono fatti e sentimenti oggettivi che sono usati per esplicare informazioni basate su fatti reali.

## NATURAL LANGUAGE PROCESSING

Infine, non dobbiamo dimenticare che la *sentiment analysis* è un problema facente parte dell'ambito del *natural language processing* (NLP). Tocca tutti gli aspetti del NLP, decisioni prese nelle conferenze, trattamento delle negazioni, ambiguità lessicale che aggiungono una maggiore difficoltà in quanto sono problemi tuttora non risolti. Comunque, è utile realizzare che questa analisi è un problema ristretto del NLP perché questo sistema

non ha bisogno di capire interamente la semantica di ogni frase o documento ma necessita di capire solo alcuni aspetti, ad esempio opinioni positive o negative e il target delle loro argomentazioni ed entità. In questo senso la suddetta analisi offre un'ottima piattaforma per i ricercatori del NLP per rendere i progressi tangibili su tutti i fronti con il potenziale di avere un grosso impatto pratico. I ricercatori al giorno d'oggi, rispetto al passato hanno degli strumenti e delle conoscenze più adatte a conoscere l'intero spettro del problema, la sua struttura e i problemi centrali. Numerosi nuovi modelli formali e metodi sono stati proposti e i ricercatori non hanno solo penetrato ma anche aperto significativamente il nocciolo delle varie problematiche con cui si viene a contatto in questa disciplina. In precedenza, essi si sono focalizzati sulla classificazione del *sentiment* o sulle idee soggettive espresse nei documenti e nelle frasi e ciò risulta insufficiente nella maggior parte delle applicazioni riguardanti la vita reale. Applicazioni pratiche spesso richiedono un'analisi molto più profonda, specifica e finemente granulata. Grazie alla maturità raggiunta su questo campo, il problema è ora molto meglio definito rispetto a prima e differenti direzioni di ricerca sono unite in un'unica definizione.

## ESEMPIO DI USO DELL'ANALISI IN DISCUSSIONE

L'8 marzo 2012, in occasione del lancio del nuovo iPad della Apple Computer Inc., sono state analizzati nelle ore seguenti all'uscita 40.000 post su Twitter e piattaforme di altri blog in lingua inglese.[1]

Dai post analizzati, i risultati dell'analisi mostravano che il 76,3% degli utenti avrebbe comprato il nuovo prodotto e il sentiment sull'acquisto del prodotto era pari al 52,6% (Fig. 1).

Questo però è un dato che si può acquisire anche limitandosi semplicemente a guardare la vendita del prodotto a qualche settimana di distanza dal lancio dello stesso. La vera cosa interessante che può attrarre un'impresa nel condurre un'analisi del genere in realtà è che dai *tweet* e dai pensieri degli utenti espressi nei vari blog si possono vedere i punti di forza e di debolezza con gli occhi dell'acquirente. Senza aver definito un set di aspetti positivi o negativi e priori, tramite la codifica dei testi, è possibile captare informazioni direttamente dal consumatore stesso.

[1] Analisi disponibile in forma integrale qui: *http://www.ilsole24ore.com/art/tecnologie/2012-03- 10/nuovo-ipad-cosa-dice-093953.shtml?uuid=AbvOGa5E*.

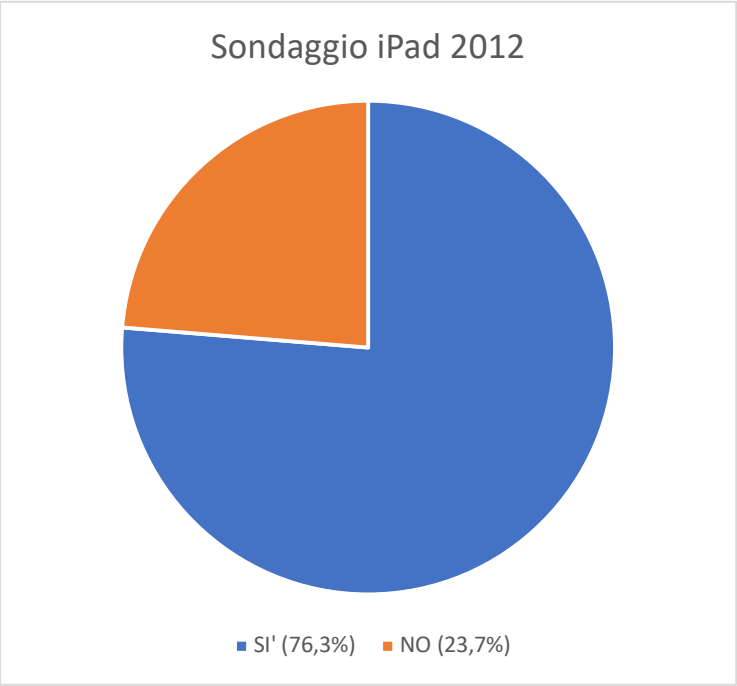
Sondaggio iPad 2012
SI' (76,3%)
NO (23,7%)

**FIG. 1**

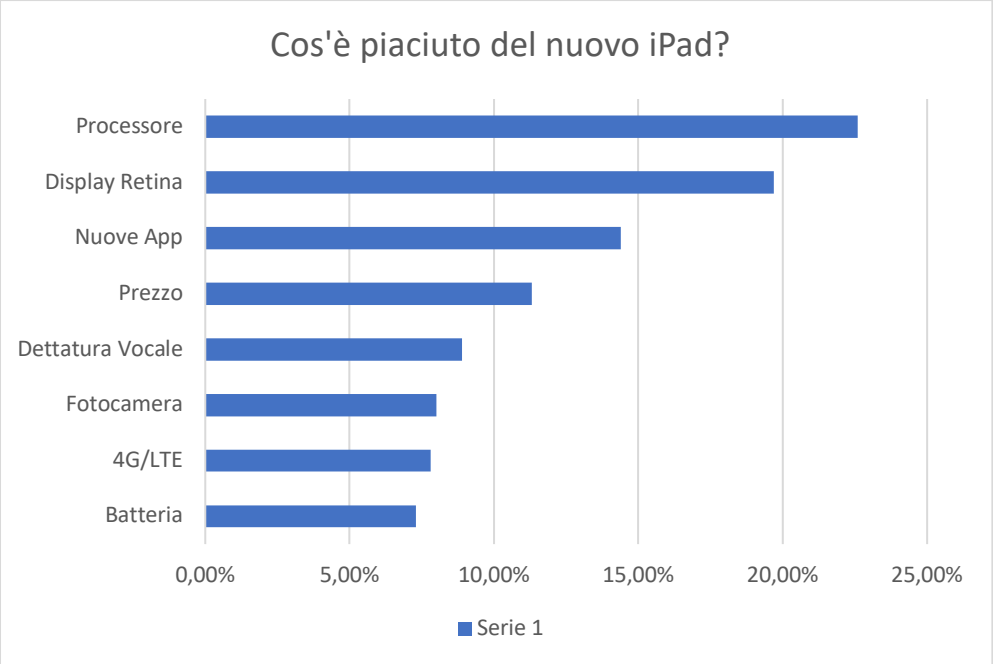
Cos'è piaciuto del nuovo iPad?
Processore
Display Retina
Nuove App
Prezzo
Dettatura Vocale
Fotocamera
4G/LTE
Batteria
0,00%
5,00%
10,00%
15,00%
20,00%
25,00%
Serie 1

**FIG. 2**

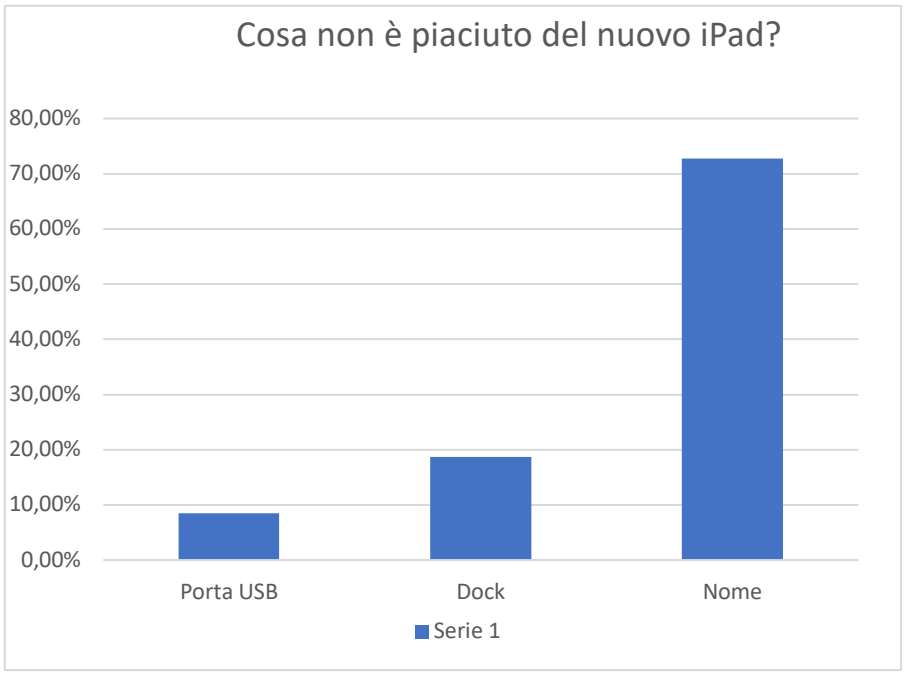


**FIG. 3**

Nella Fig. 2 la cosa lampante che emerge dall'istogramma è che la caratteristica che ha soddisfatto di più i consumatori è stata l'introduzione di un nuovo processore (22,6%) più veloce di quello precedente, il display retina migliorato (19,7%) e la batteria (14,4%) al di là ovviamente del prezzo che è diminuito (11,3%), seguite dagli altri miglioramenti apportati come indicato nel grafico. L'insieme di tutte queste ragioni, sommate una con l'altra, vanno a formare il *sentiment* positivo corrispondente nel grafico precedente al 76,3%.

Tra le motivazioni del *sentiment* negativo invece troviamo come primo aspetto emergente la mancanza di un nome al prodotto, corrispondente esattamente al 72,8%, in quanto gli acquirenti si sarebbero aspettati nei giorni antecedenti al lancio una sovrabbondanza di nomi, ad esempio "iPad3", che non è stata rispettata dalla Apple. La critica mossa dai consumatori è sicuramente un a caratteristica su cui gli esperti di marketing dovranno riflettere per i prodotti futuri. Oltre al nome, ciò che non ha soddisfatto il pubblico è stata la mancanza di un dock (18,7%) e di una porta USB (8,5%).

Questo esempio sottolinea l'efficacia e l'applicazione della *sentiment analysis* per stimare le opinioni che sono state espresse nella rete da un numero di utenti molto ampio, rispettando le varie critiche ed apprezzamenti messi in evidenza da ciascuno di essi. Questo metodo di analisi testuale, che può essere definito quali-quantitativo, presenta un numero

di vantaggi assai proficuo per qualsiasi azienda decida di adottarla in quanto prende in considerazione una ricchezza di informazioni acquisite tramite la rete tale da essere difficilmente eguagliate da altre analisi che agiscono più a strette maglie e funziona in special modo se l'argomento principale trattato sono i vari fenomeni sociali complessi.

# II
# LA SENTIMENT ANALYSIS TRA COMPARAZIONI ED OPINIONI FITTIZIE

## ANALISI DELLE OPINIONI COMPARATIVE

Una persona, come abbiamo visto nel primo capitolo preso in esame, può esprimere direttamente un giudizio positivo o negativo, ma può anche manifestare una propria opinione comparando entità simili. Tali manifestazioni sono chiamate *comparative opinions* [Jindal e Liu, 2006]. Le opinioni comparative sono collegate alle opinioni normali ma sono diverse, in quanto non hanno solo significato semantico differente ma anche altre forme sintattiche: ad esempio nella frase "La qualità della voce nei cellulari Nokia è migliore di quella degli iPhone" non viene mai detto se la qualità della voce del telefono è scarsa o buona, ma semplicemente viene fatto un paragone, quindi si può evincere che bisogna adottare delle tecniche di analisi diverse da quelle viste in precedenza.

Una frase comparativa esprime una relazione basata sulle somiglianze o differenze tra più di un'entità. Ci sono diversi tipi di comparazioni, possono essere raggruppate in due grandi insiemi: comparazioni graduabili e comparazioni non graduabili [Jindal e Liu, 2006; Kennedy, 2005].

Le comparazioni graduabili esprimono una relazione ordinata di entità che sono comparate ed hanno tre sottotipi:

1. **Comparazioni graduabili non eque:** esprimono una relazione del tipo "meglio" o "peggio" che colloca un gruppo di entità al di sopra di un altro gruppo a seconda di alcuni dei loro aspetti condivisi, ad esempio si potrebbe pensare ad un paragone tra Coca-Cola e Pepsi. Questo tipo include le preferenze espresse dal pubblico.
2. **Comparazioni eque:** esprime una relazione del tipo "è uguale a" e stabilisce che due o più entità sono simili a seconda di alcuni dei loro aspetti condivisi.
3. **Comparazioni superlative:** esprimono una relazione del tipo "il migliore" o "il peggiore" rispetto a tutte le altre entità che hanno aspetti condivisi, infatti pone un'entità al di sopra di tutte le altre.

Le comparazioni non graduabili esprimono una relazione tra due o più entità ma non riescono a fornire un grado di paragone e ci sono tre sottotipi:

1. Entità A è simile a o differente dall'entità B secondo alcuni aspetti in comune, ad esempio "la Coca-Cola ha un gusto diverso dalla Pepsi"
2. Entità A ha un aspetto $a_1$, un'entità B ha un aspetto $a_2$ ($a_1$ e $a_2$ sono solitamente sostituibili) ad esempio "I Desktop dei PC usano altoparlanti esterni mentre i portatili usano altoparlanti interni".
3. Entità A ha un aspetto a, invece l'entità B non lo ha, ad esempio "I cellulari Nokia sono forniti di auricolare, mentre quelli Apple non lo sono".

In questo capitolo ci focalizzeremo solo sulle comparazioni graduabili poiché le comparazioni non graduabili possono anch'esse esprimere opinioni, ma sono molto più sottili e difficili da riconoscere.

Nonostante molte frasi comparative contengano parole chiave di maggioranza o superlative, ci sono molte frasi che contengono queste parole e non sono stanno facendo un paragone tra due prodotti né esprimendo una preferenza per uno. In [Jindal e Liu, 2006], venne dimostrato che pressoché in tutte queste tipologie di frasi ci sono delle parole chiave che indicano un paragone, usando una serie di parole chiave, infatti il 98% delle frasi comparative è stato identificato con una precisione del 32%: le parole chiave sono aggettivi comparativi e avverbi comparativi, aggettivi superlativi e avverbi superlativi e altre parole o frasi non indicative. Dal momento in cui le parole chiave da sole hanno un alto richiamo, possono essere usate per non prendere in considerazione espressioni che sono improbabili per essere all'interno di frasi comparative, bisogna solo migliorare la precisione sulle rimanenti frasi.
È stato anche osservato in [Jindal e Liu, 2006] che le frasi comparative hanno un forte coinvolgimento con le parole chiave indicanti una comparazione e per scoprire questi collegamenti venne impiegata la regola sequenziale di classe (CSR, ovvero *class sequential rule*), che è un tipo speciale di collegamento sequenziale in cui ogni formazione di esempio è una coppia ($s_i$, $y_i$), dove $s_i$ è una sequenza e $y_i$ è un'etichetta di classe. La sequenza è generata da una frase e usando la formazione dei dati, può essere generata la CSR.

OPINION SPAM DETECTION

Le opinioni dai social media sono usate sempre di più dai singoli e dalle organizzazioni per fare decisioni d'acquisto, fare scelte in campagna elettorale, marketing e design del prodotto. Avere opinioni positive significa spesso avere dei profitti e fama per attività commerciali e singoli, che, sfortunatamente, danno forti incentivi alla gente a giocare al sistema del postare *fake opinions* e *fake reviews* (opinioni e recensioni fittizie) per promuovere o screditare alcuni tipi di prodotti, servizi, organizzazioni, singoli e persino idee, senza però divulgare le loro reali intenzioni, o le persone o organizzazioni per cui lavorano. Questi individui sono chiamati *opinion spammers* e la loro attività è chiamata *opinion spamming* [Jindal e Liu, 2008]. Fare spam di opinioni riguardo problemi politici e sociali può anche essere terrificante in quanto possono essere stravolte opinioni e mobilizzate masse facendo cambiare loro valori morali o etici. È sicuro dire che, come le opinioni nei social media stanno venendo sempre più usate nella pratica, l'*opinion spamming* sarà un fenomeno sempre più dilagante e sofisticato, che presenta una sfida maggiore per la sua rivelazione; comunque, devono essere rivelati in ordine per assicurarsi che i social media continuino ad essere una ricerca affidabile di opinioni pubbliche, piuttosto che essere pieni di notizie false, bugie ed inganni.

Come rivelare lo spam è una disciplina che è stata studiata in moltissimi campi, ad esempio lo spam tramite e-mail e tramite il web sono due tipi di spam più ampiamente studiati, ma lo spam riguardante le opinioni è molto differente. In [Castillo e Davison, 2010; Liu, 2006] si afferma che ci sono due diversi tipi di e-mail spam: lo spam tramite link e lo spam di contenuti. Il primo avviene tramite link ipertestuale, che a malapena esiste nelle recensioni, anche se i link di pubblicità sono comuni in altre forme dei social media e sono molto facilmente rintracciabili. Il secondo aggiunge parole popolari ma irrilevanti in determinate pagine Web in modo da confondere i motori di ricerca per farli rilevare da molti quesiti di ricerca, ma questo occorre appena nel pubblicare le opinioni. Lo spam tramite e-mail si riferisce a pubblicità non richieste, che sono molto rare nelle opinioni online.

**La sfida:** la chiave della sfida sul riconoscimento dello spam riguardo le opinioni è che diversamente dalle altre forme di spam, è molto difficile, se non impossibile, riconoscere le opinioni false semplicemente leggendole a mano, il che rende difficile trovare database

di spam per aiutare la progettazione della stima di algoritmi volti alla ricerca di notizie false. Per quanto riguarda le altre forme di spam, ciascuno di noi può riconoscerle abbastanza facilmente.

In realtà, nel caso estremo, è logicamente impossibile riconoscere lo spam dalla semplice lettura, ad esempio un individuo può scrivere una vera recensione per un buon ristorante e postarla come recensione fittizia per un ristorante pessimo al fine di promuoverlo. Non c'è modo di svelare queste recensioni non corrispondenti alla realtà senza considerare le informazioni che vanno al di là del semplice testo scritto semplicemente perché la valutazione stessa non può essere veritiera o fittizia allo stesso tempo.

TIPI DI SPAM E MODI DI SPAMMARE

In [Jindal e Liu, 2008] vennero identificati tre tipologie differenti di spam:

- **Tipologia 1 (recensioni false):** queste sono recensioni non corrispondenti al vero e sono scritte non in base all'esperienza genuina del recensore in base all'uso del prodotto o del servizio, ma sono scritte con fini nascosti. Contengono spesso immeritevoli opinioni positive riguardo alcune tipologie di entità con il fine di promuovere le entità e/o smentire o screditare le recensioni negative scritte a riguardo, contemporaneamente hanno anche il fine di danneggiare la reputazione di prodotti o servizi appartenenti alla concorrenza.
- **Tipologia 2 (recensioni riguardanti il solo brand):** queste recensioni non propongono commenti riguardo un prodotto o servizio specifico che si ipotizzano recensire, bensì commentano il brand o la manifattura generale dei prodotti. Anche se possono essere genuine, sono considerate come spam siccome non sono riguardanti una specificità di prodotti e sono spesso di parte. Ad esempio, è come scrivere "Io odio la Samsung, non compro mai nessuno dei loro prodotti".
- **Tipologia 3 (non recensioni):** queste non sono recensioni, ci sono due sottotipi di non recensioni: la pubblicità ed altri testi irrilevanti che non contengono alcuna opinione (domande, risposte e testi random). Strettamente parlando, non sono opinioni in quanto non esprimono pensieri pertinenti degli utenti.

È stato mostrato in [Jindal e Liu, 2008] che lo spam di tipologia numero 2 e 3 nelle recensioni è abbastanza raro e semplice da individuare con l'uso di conoscenze supervisionate. Anche se non rivelato, non è un grosso problema perché i lettori stessi possono facilmente riconoscerlo durante la lettura stessa, per questo in questo paragrafo ci focalizzeremo sulla tipologia 1.

Le recensioni fittizie possono essere viste come una forma speciale di inganno [Hancock et al., 2007; Mihalcea e Strapparava, 2009]. In ogni caso, gli inganni tradizionali solitamente si riferiscono a bugie riguardanti sensazioni vere di persone, fatti o cose. I ricercatori hanno identificato molti segnali ingannevoli nei testi, ad esempio, gli studi hanno mostrato che quando le persone mentono tendono a distaccare essi stessi e usano parole come tu, voi, lui, lei, loro piuttosto che io o me stesso. I bugiardi usano parole correlate all'ambito della certezza molto frequentemente in modo da nascondere la falsità o enfatizzare la verità, però le recensioni false sono differenti dalle bugie in molti aspetti.
Per prima cosa, ai recensori fittizi piace effettivamente usare espressioni quali io, me stesso ecc… per dare l'impressione al lettore che le loro opinioni lasciate nel web siano frutto di esperienze reali. In secondo luogo, le recensioni non sono bugie tradizionali, come ad esempio nel caso in cui uno scrittore di un libro valuti il proprio libro da lettore in modo da promuoverlo alla clientela, questa potrebbe essere la sensazione reale dello scrittore, ma con fini diversi da un semplice lettore esterno che lasci un commento negativo o positivo a riguardo. Oltretutto, c'è da dire che molti improvvisati recensori potrebbero non aver mai usato il prodotto/servizio, si limitano a dare recensioni positive o negative riguardo qualcosa di cui loro non sanno assolutamente niente, ma non stanno mentendo riguardo fatti che conoscono o loro sensazioni ed emozioni vere.

C'è anche da specificare che non tutte le recensioni fittizie sono equamente dannose. Nella tabella qui sotto cerchiamo di dare una visione concettuale di diverse tipologie di *fake reviews*, assumendo di conoscere la vera qualità del prodotto in questione. Lo scopo di queste nella regione 1, 3 e 5 è di promuovere il prodotto, anche se le opinioni nella zona 1 potrebbero essere vere, i recensori non rivelano i loro conflitti di interessi o le loro motivazioni nascoste. Lo scopo nelle regioni 2, 4 e 6 è quello di danneggiare la reputazione del prodotto, anche se le opinioni nella regione 6 potrebbero essere vere, i recensori hanno intenzioni maliziose. Chiaramente, le valutazioni appartenenti alla zona 1 e alla

zona 6 non sono molto dannose, ma quelle nelle zone 2, 3, 4 e 5 sono molto nocive. Perciò, gli algoritmi per la rivelazione delle recensioni fittizie dovrebbero concentrarsi per identificare quali appartengono alle regioni più pericolose, infatti degli algoritmi già esistenti si stanno focalizzando sull'uso di questa idea impiegando tipi diversi di caratteristiche di deviazione della classificazione. Da notare come la buona, scarsa o normale qualità può essere definita basandosi sul giudizio medio dato dalle recensioni riguardanti il prodotto, sebbene questa potrebbe essere non valida nel caso in cui ci fossero pochi recensori o troppi spammer.

| | Positive fake review | Negative fake review |
|---|---|---|
| Good quality product | 1 | 2 |
| Average quality product | 3 | 4 |
| Bad quality product | 5 | 6 |

Le recensioni fittizie possono essere scritte da più tipi di persone, ad esempio familiari, amici, impiegati della compagnia, competitors, aziende che provvedono esse stesse a scriverle, o anche consumatori genuini a cui un'azienda può dare degli sconti o interi rimborsi a patto che scrivano una recensione positiva sul web. In altre forme di social media, agenzie pubbliche e private e organizzazioni politiche possono impiegare persone per pubblicare messaggi in modo da influenzare segretamente le conversazioni dei social per diffondere menzogne e disinformazione.

In generale, uno spammer può lavorare individualmente, oppure essere consciamente o inconsciamente membro di un gruppo di lavoro avente lo stesso fine; lo spam individuale avviene nel caso in cui il recensore lavori singolarmente senza nessun altro, scrive usando un singolo user-id, come ad esempio potrebbe fare un autore di un libro.

## CONNETTERE LA DISTANZA TRA LE VARIE LINGUE

La creazione di un lessico o una collezione di dati etichettati è un compito di tempo/prova intensiva. Dal momento che l'inglese è la lingua dominante nella quale la ricerca riguardo la *sentiment analysis* è stata eseguita, è naturale che molti altri linguaggi abbiano provato ad influenzare le ricerche sviluppate per l'inglese cercando di adattarle o riutilizzarle. L'attraversamento linguistico si riferisce all'uso dello sviluppo di sistemi e ricerche per una lingua per eseguirne un'altra. Il primo linguaggio (quello in cui ricerche/lessico/sistemi sono stati sviluppati) è chiamato linguaggio di ricerca, mentre il secondo linguaggio (dove gli elementi citati precedentemente devono ancora essere sviluppati) è chiamato linguaggio bersaglio. Le basi dell'attraversamento linguistico è la disponibilità del lessico o un'annotata collezione di dati nel linguaggio di ricerca. È doveroso notare che esistono moltissime metodologie sfumate per eseguire l'attraversamento lessicale con la *sentiment analysis*, ma su questo paragrafo ci concentreremo sulle ricerche riguardanti queste differenze lessicali.

Il requisito fondamentale è una mappatura tra i due linguaggi. Spieghiamo ora cosa succederebbe se volessimo mappare il linguaggio X nel linguaggio Y: questa mappa potrebbe essere trasformata sotto forma di dizionario parallelo dove le parole di un dizionario sono tradotte in un altro. ANEW per lo spagnolo [Redondo et al., 2007] descrive la formazione di un nuovo lessico chiamato ANEW. In origine fu creato per le parole inglesi, la sua versione parallela spagnola venne creata per tradurre le parole dall'inglese allo spagnolo, poi convalidandole manualmente. Può essere anche usato sotto forma di collegamento su WordNet (database semantico-lessicale elaborato da George Armitage Miller nell'Università di Princeton), nel caso in cui il lessico si comporti come un insieme di sinonimi (*synset*). Hindi SentiWordNet [Joshi et al., 2010] mappa un insieme di sinonimi in inglese dall'Hindi usando collegamenti con un database semantico, e genera un WordNet dalle sue varianti inglesi. In [Mahyoub et al., 2014] troviamo la descrizione di una tecnica usata per descrivere un lessico riguardante il *sentiment* per l'arabo. Basandosi su un insieme di parole negative e positive, e un database arabo, presentano un'grande algoritmo per creare il linguaggio. L'algoritmo usa le relazioni del database in modo da propagare le etichette del *sentiment* in nuove parole/insiemi di parole. Queste relazioni sono divise in due categorie: la prima che preserva l'orientazione del *sentiment*, la seconda che lo inverte.

Come fa questo processo a mappare le parole di una lingua in altre senza alcuna differenza per le collezioni di dati? In caso sia possibile un sistema di traduzione automatica, questo

passaggio è semplice: una collezione di dati nel linguaggio di ricerca può essere tradotta nel linguaggio bersaglio, questa è una strategia molto comune che è stata impiegata in passato [Mihalcea et al., 2007; Duh et al., 2011]. Ne segue che la traduzione può introdurre errori in più al sistema, causandone un peggioramento in termini di qualità della collezione di dati. Questo è particolarmente applicabile ala traduzione degli idiomi portatori di *sentiment*. [Salameh et al., 2015] eseguirono i loro esperimenti per la lingua araba dove un sistema di traduzione automatica è stato usato per tradurre documenti, dopo di che si è potuta concretizzare l'idea di applicazione della *sentiment analysis*. Un'acuta osservazione fatta dagli autori è che queste traduzioni potrebbero risultare "povere di significato" e quindi sarebbe difficile identificare il *sentiment* espresso, un classificatore interpreta ragionevolmente meglio, ad ogni modo, questo sistema di traduzione non esiste parimente per tutte le lingue. [Balamurali et al., 2012] suggerì un ingenuo rimpiazzo per la traduzione automatica, ovvero per tradurre un corpus dall'Hindi al Marathi e viceversa loro ottengono annotazioni sensate nel dataset, per poi usare un collegamento su WordNet per trasferire le parole dal linguaggio ricerca al linguaggio bersaglio. Un'immediata domanda che si presenta è l'ipotesi alla base di tutti gli approcci tra le varie lingue. Questo significa che se una parola ha un certo *sentiment* nel linguaggio ricerca, la parola tradotta nel linguaggio bersaglio (con un appropriato senso documentato) assume anch'essa lo stesso *sentiment*. Quanto è imparziale l'ipotesi che le parole nelle diverse lingue trasmettano le stesse emozioni? Questo può essere visto dalle correlazioni lineari tra classificazioni per tre dimensioni affettive, come è stato fatto per ANEW per lo spagnolo. ANEW per lo spagnolo [Redondo et el.,2007), come scritto sopra, è un lessico creato usando ANEW in inglese ed i valori correlativi di valenza indicano che una parola che ha un significato positivo in inglese ha molto probabilmente lo stesso significato in spagnolo.

Perciò, abbiamo due opzioni: la prima opzione usa le risorse generate per il linguaggio ricerca e lo mappa nel linguaggio bersaglio; la seconda opzione prevede di creare risorse direttamente per il linguaggio bersaglio senza passare per il linguaggio ricerca. [Balamurali et al., 2013] pesa il metodo nel linguaggio contro il metodo cross-linguale basato sulla traduzione automatica, dimostrando che per inglese, tedesco, francese e russo è molto meglio la seconda opzione.

La prima opzione beneficia anche da raccolte addizionali nel linguaggio bersaglio:

- **Raccolta non etichettata nel linguaggio bersaglio:** questo tipo di raccolta è usata in approcci differenti, il più notevole è quello basato sulla formazione aziendale. [Wan, 2009] assume che è disponibile una raccolta etichettata nel linguaggio ricerca, una raccolta non etichettata nel linguaggio bersaglio e il sistema di traduzione automatica per tradurre reciprocamente i due linguaggi.
- **Raccolta etichettata nel linguaggio bersaglio:** si assume che la dimensione di questa raccolta dati sia molto più piccola che la formazione.
- **Informazioni pseudo-parallele:** [Lu et al., 2011] descrive l'uso di informazioni pseudo-parallele per i suoi esperimenti: questo è l'insieme di frasi nel linguaggio ricerca che sono tradotte nel linguaggio bersaglio e usate come un'informazione aggiuntiva della polarità etichettata e ciò consente al classificatore di essere formato su un più ampio numero di esempi.

# III
# SENTIMENT ANALYSIS: ALCUNE APPLICAZIONI

## SENTIMENT ANALYSIS E IL MERCATO DELLE AUTOMOBILI

L'industria dell'automobile è una tra le più importanti industrie nei Paesi Bassi e impiega più di 50.000 dipendenti. Le previsioni delle vendite nell'industria dell'automobile sono parecchio importanti siccome nel sistema corrente le macchine sono costruite sia per consegna sia per previsione. Comunque, quest'ultimo metodo spesso conduce a un effetto frusta dovuto all'incertezza della domanda e non accuratezza della previsione [Suthikarnnarunai, 2008]. Anche nel caso in cui l'azienda decida di produrre le automobili per consegna, una predizione accurata può tuttavia aiutare i manager a pianificare e allocare meglio le risorse industriali.

I social media si comportano come le parole che, una volta pronunciate, permettono alle compagnie di collezionare opinioni su larga scala e aggiornarsi in base alle opinioni sincere dei consumatori [Ceron et al., 2013; Tuarob et al., 2014]. Molte imprese hanno sviluppato e compreso questa possibilità e pagano per aumentare l'attenzione al contenuto delle proprie pagine social [Karlgren et al., 2012; Liu, 2012; Wijnhoven e Bloemen, 2014]. Come abbiamo visto, un metodo molto diffuso usato da molte aziende per analizzare questi dati è la *sentiment analysis*, che analizza l'opinione delle persone, i sentimenti, le valutazioni, le attitudini e le emozioni dal linguaggio naturale [Pang e Lee, 2008]. Capire questi sentimenti è importante sia per la produzione che per la valutazione dei servizi e, questi sentimenti, potrebbero anche avere un impatto sull'acquisto futuro delle persone. Tra i vari studi che fanno uso della *sentiment analysis* ci sono previsioni di vendite cinematografiche [Asur e Huberman, 2010], movimenti dei prezzi delle azioni [Bing et al., 2014; Nguyen et al., 2015], vendite dei libri [Dijkman et al., 2015] e vendite degli iPhone [Lassen et al., 2014]. Ciò nonostante, la letteratura è ancora carente di informazioni per quanto riguarda la validità e affidabilità della *sentiment analysis* nel contesto di beni più costosi quali ad esempio le automobili.

I consumatori sono soliti dedicare un considerevole lasso di tempo a cercare informazioni riguardo un potenziale veicolo. In uno studio di [Kandaswami e Tiwar, 2014] venne dimostrato che la maggior parte dei consumatori passa più di dieci ore per identificare il

miglior veicolo che soddisfi le loro richieste e soddisfi le loro aspettative. In Cina, questi consumatori corrispondono al 70% della popolazione, mentre nei paesi occidentali quali ad esempio Germania e USA, una percentuale circa del 40-50% ha raggiunto questo lasso di tempo. Conseguentemente, volumi di ricerca, come registrato da motori di ricerca quali Google, possono rappresentare una grandissima parte della ricerca delle persone per informazioni decisionali per l'acquisto della loro macchina. Pertanto, ricerche precedenti hanno studiato un potere di previsione di Google Trends per la vendita di automobili, anche se con un diverso tasso di successo [Barreira et al., 2013; Geva et al., 2017]. Secondo alcuni analisti il miglior modello di predizione della vendita di automobili è una combinazione dei dati di Google Trends, i sentimenti espressi nei Forum e i volumi di menzione dei Forum come visionari. Il nostro scopo in questa contribuzione è un'identificazione più lontana del significato di queste predizioni in modo da riuscire ad utilizzare i dati come strumento di analisi per anticipare il mercato delle automobili. Per questo useremo il modello AIDA del viaggio del consumatore [Gensler et al., 2017]. AIDA sta per i quattro passaggi nel processo di un consumatore quando si trova ad acquistare un prodotto. La A sta per attenzione. L'attenzione è attirata specialmente dai saldi promozionali e dalla pubblicità, che possono essere riscontrate per esempio nei Social Media. La I sta per interesse, che è presente nel comportamento di ricerca dei consumatori, che può essere registrato in Google Trends. La D sta per desiderio, implicito nel sentimento delle persone e che può essere registrato dal rapporto di espressioni positive o negative, rapporto P/N. L'ultima A sta per azione, che è l'atto di acquisto del potenziale compratore.

1) Pertanto, una domanda di ricerca è: "Qual è il potere predittivo dei sentimenti per il modello delle automobili espresso sui social media verso i venditori dell'industria automobilistica nei Paesi Bassi?"
Questa domanda richiede un'esplorazione dell'applicabilità dei risultati della *sentiment analysis* come indicatori del desiderio nei confronti del prodotto in questione, e perciò potrebbe essere utile per costruire una formula predittiva da queste conclusioni. In questa previsione, i lassi di tempo hanno un ruolo chiave, poiché se questi lassi di tempo tra le espressioni del sentimento e il momento della vendita è positivo, ovvero la vendita avviene dopo l'espressione, esso indica un desiderio; se invece sono negativi queste espressioni potrebbero indicare una valutazione. Compareremo anche il potere dei sentimenti

con i poteri di Google Trends perché ricercare in questo motore di ricerca può essere un indicatore di interesse con possibile previsione.

2) La nostra seconda domanda di ricerca è: "Per predire le vendite delle automobili nei Paesi Bassi è più utile usare i dati di Google Trends o i sentimenti espressi nei social media?"
Ci aspettiamo che se le attività AIDA seguono un processo sequenziale, che il volume dei picchi di Google Trends avrà un più ampio rallentamento e una correlazione più debole con gli attuali picchi di vendita rispetto ai picchi dei sentimenti.

In questa sezione verrà offerta una panoramica della letteratura e delle teorie rilevanti utile per presentare un modello di ricerca che intende spiegare le possibili relazioni tra Google Trends, i dati dei social media e le vendite. Successivamente, i dati provenienti da questi canali saranno analizzati considerando le abilità di previsione di vendita e saranno comparati i risultati. I risultati ottenuti saranno allora discussi nel contesto di implicazioni e limitazioni pratiche e accademiche.

ALCUNI MODELLI
La *sentiment analysis* fa riferimento all'estrazione di sentimenti ed opinioni da testi scritti [Liu, 2012]. [Bing et al., 2014] dichiararono che i sentimenti nei social media influenzino le vendite solo con certo ritardo, che chiameremo lasso di tempo. Questo lasso tra una crescita/decrescita di commenti positivi o negativi e una crescita/decrescita nelle vendite è variabile dal momento che i social media potrebbero influenzare il consumatore sia in una situazione precoce che in una situazione avanzata nel processo d'acquisto. Questo tempo diventa chiaro durante l'osservazione del consumatore durante il processo decisionale. Secondo [Kotler, 1994], i consumatori attraversano cinque passaggi quando comprano un prodotto. La decisione iniziale di comprare inizia con il riconoscimento da parte del consumatore di un problema o un'esigenza invece di essere persuaso da un prodotto, in seguito egli percorre i seguenti passaggi: ricerca di informazioni, valutazione delle alternative, decisione d'acquisto; comportamento post acquisto. Il lasso di tempo in questo ciclo si ha tra la fase della ricerca di informazioni e momento d'acquisto mentre sta valutando le alternative.

Come abbiamo visto in precedenza un altro processo d'acquisto che descrive il modello comportamentale di un consumatore in termini di ricerca di informazioni è il modello AIDA. Possiamo aggiungere a quanto detto in precedenza che le attività di ricerca possono essere significative di un'intenzione a comprare e addirittura predire il comportamento del consumatore del suo grado di coinvolgimento nell'acquisto [Choi e Varian, 2012; Goel et al., 2010; Yang et al., 2015]. Ad ogni modo, Google Trends non rappresenta sentimenti positivi o negativi, perciò non è idoneo a indicare il desiderio verso un prodotto, che può invece essere misurato dal rapporto delle opinioni positive e negative espresse nei social media. Seguendo infatti questo modello, l'interesse dà un'indicazione più forte d'intenzionalità all'acquisto rispetto all'attenzione, a sua volta il desiderio è più forte dell'interesse, perciò ci aspetteremo una correlazione molto più forte tra volume delle vendite e desiderio rispetto ad attenzione ed interesse. Allo stesso modo ci aspetteremo un lasso di tempo minore tra i picchi della fase del desiderio e i picchi della fase d'acquisto rispetto al tempo registrato tra i picchi della fase attenzione ed interesse e di acquisto. Google può fortemente influenzare la decisione in base all'ordine dei risultati di ricerca. Ciò è verificato nelle situazioni in cui l'acquirente è molto indeciso nella fase iniziale [Epstein e Robertson, 2015].

Ci sono state ricerche precedenti che hanno esplorato la rilevanza dei social media riguardo la rivelazione delle vendite, come ad esempio il lavoro di [Asur e Huberman, 2010] in cui si riesce a prevedere in maniera notevolmente accurata la vendita dei botteghini includendo diverse variabili quali sentimenti e frequenza dei *tweet* nel loro modello. Molte altre ricerche hanno seguito l'approccio dei due studiosi sopracitati, ad esempio uno studio basato su questo modello svolto da [Lassen et al., 2014] predice la vendita trimestrale di iPhone analizzando il *sentiment* espresso nei *tweet* e usando una ponderazione stagionale degli stessi per calcolare la data proporzione trimestrale dell'ultimo anno di calendario. Ad ogni modo, predire le vendite di automobili dai sentimenti non è un problema irrilevante, molte espressioni di sentimento non sono necessariamente una manifestazione di un desiderio ma possono essere anche valutazioni. Se fosse una valutazione, questo influenzerebbe indirettamente il desiderio. Tuttavia, le valutazioni possono anche essere d'aiuto per conoscere l'utilità di un prodotto [Pang e Lee, 2008]. Se con P chiamiamo i *tweet* positivi, con N chiamiamo i *tweet* negativi e con O i *tweet* neutri, allora la soggettività sarà data da $\frac{p+n}{o}$ e il rapporto positivo o negativo (PNR) sarà dato da $\frac{p}{n}$.

Un approccio similare sarà poi preso dove il PNR sarà una variabile indipendente. In ogni caso, p non sarà definita come il numero di post attivi ma come la percentuale di post positivi rispetto a tutti i post riguardo un particolare modello di automobile nell'arco di un mese; vale anche per n essere la percentuale dei post negativi. In questo modo si impedisce l'aumento dell'uso dei social media e il conseguente aumento di post negli anni passati per influenzare il risultato della ricerca e sfalsarne l'attendibilità. Il PNR sarà lo stesso indipendentemente dal fatto che venga calcolato con percentuale o con i numeri assoluti siccome si limita a descrivere il rapporto tra i due.

Può essere interessante analizzare le seguenti ipotesi:

**H1: il numero delle menzioni totali riguardanti un modello di auto sui social media ha una correlazione positiva con il numero delle vendite dell'auto.**

[Betaineh, 2015] concluse che tre fattori del passaparola elettronico (eWOM) hanno un significato ed un impatto positivo sulle intenzioni d'acquisto del consumatore. Conseguentemente, ipotizziamo che non solo il *sentiment* del eWOM ma anche l'attenzione come presente nel volume delle menzioni ha un'influenza positiva sul volume degli acquisti. Da qui si forma l'ipotesi successiva.

**H2: il numero delle menzioni totali riguardanti un modello di auto sui social media ha un'influenza negativa con il numero delle vendite dell'auto.**

Analogamente all'assunzione che le recensioni positive abbiano un impatto positivo sulle vendite, si assume che le recensioni negative abbiano un impatto negativo. [Lee, Park e Han, 2008] trovarono che le attitudini del consumatore verso un prodotto divengano più sfavorevoli come aumenti la proporzione delle recensioni online negative del consumatore. Da quando essi constatarono che il consumatore tende a credere ai commenti negativi più di quanto creda ai commenti positivi ci si aspetta che questa relazione sia più forte rispetto a quella tra commenti positivi e risultato delle vendite.

**H3: la percentuale negativa di menzioni riguardo un modello di automobile ha un'influenza negativa sul numero di vendite di questo modello.**

Le ipotesi precedenti si basano sull'assunzione che le persone che abbiano già comprato una macchina, abbiano anche pubblicato dei post online sui social media recensendola, così da invogliare altre persone a comprare lo stesso oggetto. Ciò non riguarda necessariamente solo le automobili di fascia alta, ad esempio le auto di lusso solitamente ricevono

un sacco di attenzione con una clientela limitata e ricercata che le acquista. È quindi possibile che un aumento dei commenti positivi da fans di automobili di fascia alta non serva ad incrementarne gli acquisti.

**H4: Più alto è il prezzo dell'automobile, più debole è la correlazione tra dati dei social media e vendite.**

Quando vengono comparati gli studi riguardo Google Trends e Twitter come anticipatori per le vendite, Twitter solitamente fornisce un valore al quadrato R più alto [Asur e Huberman, 2010; Choi e Varian, 2012]. I precedenti studi sulle previsioni delle vendite di auto tramite Google Trends eseguiti da [Barreira et al., 2013, Fantazzini e Toktamysiva, 2015] non risultarono modelli di predizione efficaci. Uno studio più recente di [Geva et al., 2017] ha individuato un modello di successo, non tenendo in considerazione le differenze tra i vari modelli di auto. Questo è problematico, perché il mercato dell'auto è altamente differenziato e tende ad essere monopolistico perciò il comportamento e le decisioni del consumatore possono essere diverse in base all'auto [Brandler e Spencer, 2015; Cattani et al., 2017; Hunt e Morgan, 1995].

**H5: La correlazione tra sentimenti e vendita di automobili per modelli con prezzi differenti è più alta rispetto alla correlazione dei relativi volumi di ricerca con vendita di automobili per modelli con prezzi differenti.**

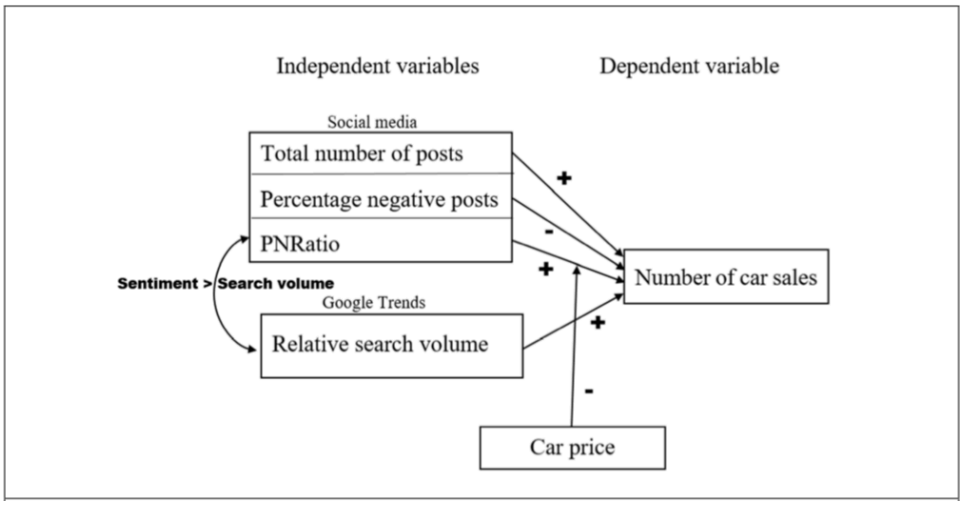


**FIG. 4**

Per testare le ipotesi di ricerca e rispondere alle domande, è importante conoscere la corrispondenza tra un gruppo di persone che esprimono un'opinione e gruppo obiettivo della

ricerca [Wijnhoven e Blomen, 2014]. Ovviamente non tutte le persone che intendono comprare una macchina hanno un account nei social media, nei Paesi Bassi questo numero corrisponde all'80% della popolazione (CBS 2015), quindi possiamo affermare che la maggior parte dei compratori di auto ha accesso ai social media.

I dati richiesti furono raccolti, per un totale di undici modelli, in un periodo di 52 mesi che oscilla da Gennaio 2012 fino ad Aprile 2016. Le auto vennero scelte in base al sistema europeo della classificazione delle automobili (Ufficio per le pubblicazioni ufficiali della comunità europea 1999) che le divide in base alla loro massa e specifiche tecniche. Le cosiddette mini-macchine sono etichettate come classe A mentre quelle più grandi sono classificate come classe B e superiori. La lista finisce con la classe F che descrive le auto di lusso, mentre le classi extra comprendono anche la classe S, ovvero quella delle auto sportive, e la classe J, ovvero quella degli off-road. Questa ricerca analizzerà due modelli ciascuno dalla classe B alla E ed un modello da classe A a F oltre alla classe S, le quali coprono le auto più comuni oltre a due auto di lusso. Per la lista completa si guardi la tabella qui sotto:

| Car model | Class | Starting price in € | Search term Coosto | Search term Google Trends |
|---|---|---|---|---|
| Fiat Panda | A | 13,675 | Fiat Panda | Fiat Panda |
| Ford Fiesta | B | 13,995 | Ford Fiesta | Ford Fiesta |
| Opel Corsa | B | 20,950 | Opel Corsa | Opel Corsa |
| Honda Civic | C | 21,050 | Honda Civic | Honda Civic |
| VW Golf | C | 21,050 | (Volkswagen Golf) OR (VW Golf) | "Volkswagen Golf" + "VW Golf" -Trends |
| Ford Mondeo | D | 29,575 | Ford Mondeo | Ford Mondeo |
| Volkswagen Passat | D | 31,450 | (Volkswagen Passat) OR (VW Passat) | "Volkswagen Passat" + "VW Passat" -Trends |
| BMW 5 Series | E | 47,990 | "BMW 5-serie" –GT | "BMW 5-serie" –GT |
| Mercedes E Class | E | 46,800 | (MB OR Mercedes-Benz OR Mercedes) "E-Klasse" | MB + Mercedes Benz + Mercedes "E-Klasse." |
| Porsche Panamera | F | 106,400 | Porsche Panamera | Porsche Panamera |
| Porsche 911 | S | 115,000 | Porsche 911 | Porsche 911 |

**FIG. 5**

Per ogni modello di auto sono state raccolte le seguenti variabili: il numero totale di post al mese, il numero di post positivi sui social media al mese, il numero di post negativi sui

social al mese, il punteggio di Google Trend di ogni mese e il numero di auto vendute ogni mese.

I dati provenienti dai social media furono analizzati tramite l'uso della *sentiment analysis* con l'uso del *tool* Coosto. Questo strumento consente di analizzare i post sui social media dei Paesi Bassi e classificare ciascun post come positivo, neutrale o negativo. Come fonti, Coosto fornisce otto siti dei social media, così come vari nuovi siti, blog, e forum: ciò significa che i dati raccolti tramite questo *software* sono meno di parte rispetto alle dinamiche social e alla popolazione di appartenenza di una piattaforma, come *Twitter* [Mislove et al., 2011; Wijnhoven e Bloemen, 2014; Wilson et al., 2012]. Coosto ha un'elevata reputazione ed è largamente utilizzato, il suo classificatore del *sentiment* ha un'accuratezza dell'80% secondo il Team Nijhuis (2013), un servizio di marketing su internet, che è ampiamente paragonato ad altri *tools* considerando i risultati della ricerca di [Serrano-Guerrero et al., 2015]. Dal momento che la ricerca è svolta per determinare il *sentiment* derivante dai social media, sono stati esclusi dall'area di ricerca blog, forum e siti nuovi. I rimanenti social in cui verrà svolta la suddetta sono *Twitter*, *Facebook*, *LinkedIn*, *YouTube*, *Google+*, *Hyves*, *Instagram*, e *Pinterest*; in totale sono stati analizzati 502.681 post.

La ricerca fu condotta inserendo il nome olandese di ciascun modello di automobile su Coosto e raccogliendo ogni mese tutti i commenti relativi e dividendoli tra positivi, neutrali e negativi tramite l'analizzatore di *sentiment* di Coosto, su *Twitter* sono inclusi sia gli originali che i *retweets*. Ciò incrementa la validità del metodo di misura dal momento che un post viene spesso *retweettato* e viene letto da molte più persone influenzandone la decisione d'acquisto.

Durante la ricerca dei dati *Google Trends*, vennero controllate espressioni e termini vari che un utente può utilizzare mentre scrive un post relativo ad una vettura, per esempio l'interrogativo per il modello Passat della Volkswagen era: "VW Passat" o "Volkswagen Passat". Inoltre, le automobili con nomi simili ma specifiche diverse vennero esplicitamente escluse dalla ricerca, come ad esempio la BMW 5 serie GT. I post riguardanti questa vettura sarebbero stati messi in evidenza lo stesso quando si sarebbe andati a cercare la BMW serie 5, tuttavia la GT differisce dalla BMW 5 standard il che significa che

è improbabile che i post riguardanti questa versione influenzino i consumatori a comprare la versione standard. La domanda venne pertanto definita come "BMW serie 5"-GT. Come fattore di comparazione, il volume di ricerca relativi per mese di Google Trends venne collezionato per lo stesso modello di vettura usando gli stessi termini di ricerca. Questi a volte dovettero essere aggiustati di poco in modo da accordare il linguaggio di ricerca di Google Trends, per esempio sostituendo il termine di ricerca booleano OR con il simbolo +. Google Trends fornisce solo il volume di ricerca relativo per mese, che è la parte di domanda del termine ricercato. La parte di domanda viene calcolata dividendo la parte di volume del termine ricercato con il numero totale di ricerche nella regione specifica. Il mese con il più alto volume di ricerche relative nella regione specifica viene in seguito normalizzato a 100. Facendo riferimento a questa ricerca ciò significa che il volume di ricerca di un particolare modello di automobile venne diviso per tutte le ricerche nei Paesi Bassi tra Gennaio 2012 e Aprile 2016. Dopo questo, il mese con la più alta media riceve quindi il ponteggio mentre i punteggi degli altri mesi sono aggiustati in accordo a quello massimo [Choi e Varian, 2012]. Questa normalizzazione implica che il punteggio di 100 rappresenti una parte di domanda differente per ogni modello, dipendendo da cosa era il massimo delle ricerche del mese.

Per la variabile dipendente, il numero di vetture vendute mensilmente venne dedotto dal sito internet BOVAG, la Federazione di commercio di automobili e detentori di garage olandese; il sito di ricerca è https://www.bovag.nl/pers/personenauto/verkoopcijfers-personenauto-s-naar-merk-model-per (10 Maggio 2016).

Uno studio ha rivelato che il 60% dei compratori normali avesse bisogno tra uno e sei mesi di tempo dal primo pensiero d'acquisto e l'effettiva azione di comprarla mentre solo il 16% abbia bisogno di meno di un mese per questa scelta. Solo il 9% aveva bisogno di più di un anno per comprare un nuovo veicolo [Putsis e Srinivasan, 1994-1995]. La durata della decisione, ovvero il lasso di tempo nel nostro modello, sarà perciò testato fino ad un massimo di dodici mesi dal momento della vendita.

# CONCLUSIONE

Ogni giorno un consumatore ordinario si trova a dover affrontare delle scelte di consumo, poiché nel mercato si può trovare una vasta gamma di prodotti, e necessita di acquisire delle informazioni a riguardo per risolvere le insicurezze sull'acquisto. Per aiutare il consumatore con questa decisione sono stati pensati dei metodi di sintetizzazione di informazione appartenenti al campo della *sentiment analysis* o *opinion mining*. Abbiamo visto come la ricerca su questi temi è di rilevante importanza se volta a risolvere i problemi tecnici vincolati all'estrazione, sintesi ed analisi però può contemporaneamente avere lo scopo di prevedere fenomeni sociali. Le sue applicazioni trovano uso specialmente in marketing e scienze sociali, servendosi dei social network come mezzo di raccolta dati. Una difficoltà che si può riscontrare in questo settore è la "cattiva informazione", poiché garantendo nelle piattaforme social all'utente di potersi esprimere liberamente senza alcuna barriera, si rischia di rendere uguali tutte le opinioni che vengono espresse ed abbassare inesorabilmente la qualità dell'informazione digitale.

Possiamo affermare quindi che la *sentiment analysis* è quel campo di studi che analizza le opinioni, i sentimenti, le valutazioni, gli apprezzamenti, le attitudini e le emozioni delle persone in merito a entità come prodotti, servizi, organizzazioni, individui, problemi, eventi e loro attributi.

Questo strumento ha una serie di caratteristiche vincenti che altrimenti sarebbero difficilmente acquisibili, ad esempio l'acquisizione in tempo reale di informazioni su un campione vastissimo in forma geo-localizzata. In molti sono ancora molto scettici sull'argomento in quanto ritengono che i sondaggi tradizionali e la figura dell'intervistatore non debba essere sostituita, anche se il progresso tecnologico riserva dei vantaggi non indifferenti.

Queste forme di ricerca si servono di parole chiave che aiutano i server ad identificare la presenza di opinioni e giudizi e vengono chiamate *sentiment words* o *opinion words*. Gli studi ci hanno però esplicato che questi indicatori non bastano a capire questo tipo di analisi in quanto ci sono dei problemi da affrontare sull'interpretazione che una parola assume nel contesto della frase e il sarcasmo è molto difficile da decifrare.

Avendo visto in precedenza tutte i vantaggi e le limitazioni di cui la *sentiment analysis* gode, la vera sfida per questa disciplina resta il riconoscere automaticamente l'espressione di sentimenti ed emozioni a partire dal testo. Ci si trova davanti infatti ad una serie di difficoltà da affrontare, già partendo dalla granularità dell'analisi, che può dividersi in

documento, frase, entità o caratteristica, si può intuire quanto sia possibile scendere in profondità in un testo e rendere l'analisi sempre più complessa. Aggiungendo poi le difficoltà di natura semantica, come l'ambiguità che possono assumere le frasi e le parole a seconda dell'utilizzo, e di natura pragmatica, come ad esempio la *sarcasm detection*, si può immaginare che un metodo di analisi automatico sia in difficoltà nell'interpretazione delle reali opinioni che il consumatore desidera esprimere.

Abbiamo ampiamente trattato anche il tema dell'*opinion spam detection* nel secondo capitolo, analizzando i comportamenti degli *opinion spammer* e i vari modi di spam. Tenendo conto che ci siamo particolarmente soffermati sull'analisi delle recensioni fittizie, in quanto le recensioni che commentano il brand o la manifattura e la banale pubblicizzazione dei prodotti sono facilmente riconoscibili dalla maggioranza, si può affermare che sia fondamentale saper riconoscere questo problema in quanto si possono celare fini diversi dalla reale esperienza del consumatore quali lucro o promozione e screditamento di alcune attività o prodotti. Anche la connessione tra le varie lingue può risultare difficoltosa, in quanto esprimere un concetto dando la stessa enfasi in un altro linguaggio può risultare complicato per un sistema di traduzione automatica.

Nonostante tutte queste difficoltà, abbiamo anche visto tramite un esempio che l'utilizzo della *sentiment analysis* può essere una risorsa più che vantaggiosa per prevedere l'andamento delle vendite in base alle opinioni dei consumatori e, allocando le risorse in maniera più oculata, i manager delle varie imprese di produzione di automobili avrebbero potuto soddisfare meglio la domanda di mercato. Riuscire quindi a utilizzare con maggior accuratezza il potere predittivo dei sentimenti è uno stratagemma che, se valorizzato, potrebbe diventare una chiave di svolta in termini di precisione e accuratezza delle vendite a cui non si sarebbe potuti arrivare in precedenza.

Resta un ultimo dilemma da affrontare nel sentiero appena percorso riguardo questa disciplina: usare le opinioni che i consumatori scrivono in maniera del tutto disinteressata e fiduciosi del fatto che verranno trattate con riservatezza è eticamente corretto? Attualmente uno dei temi più trattati e discussi sono proprio le condizioni di privacy a tutela del cittadino, nel nostro caso riflettiamo sulle esigenze del pubblico social. Non è oramai un mistero che la gratuità dell'uso dei social network venga poi ripagata dalle informazioni e dai pensieri degli utenti che ne rappresentano la vera ricchezza. Nonostante la discutibile legittimità dell'uso di queste informazioni da parte sia delle multinazionali che dagli stessi *twitter, Facebook* e tutte le altra piattaforme virtuali, possiamo affermare che questo

fenomeno sociale riguardante i dati online debba essere maggiormente tutelato e controllato.

# BIBLIOGRAFIA


Asur, S., and Huberman, B. A. 2010. "*Predicting the future with social media*," in Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on (Vol. 1), IEEE, pp. 492–499.

Balamurali, A., Joshi, A., and Bhattacharyya, P. (2012). *"Cross-Lingual Sentiment Analysis for Indian Languages using Linked WordNets"* pp. 74-80 *In Proceedings of COLING*, 2012.

Balamurali, A., Haffari, G., and Bhattacharyya, P. (2013). *"Leveraging Unlabelled Corpora for Sentiment Analysis".*

Bar-Haim, Roy, Elad Dinur, Ronen Feldman, Moshe Fresko, and Guy Goldstein. *Identifying and Following Expert Investors in Stock Microblogs*. in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2011)*. 2011.

Barreira, N., Godinho, P., and Melo, P. 2013. "*Nowcasting unemployment rate and new car sales in southwestern Europe with Google Trends*," NETNOMICS: Economic Research and Electronic Networking (14:3), Springer, pp. 129–165.

Bing, L., Chan, K. C. C., and Ou, C. 2014. *"Public sentiment analysis in Twitter data for prediction of a company's stock price movements,"* 2014 Ieee 11th International Conference on E-Business Engineering (Icebe), pp. 232–239 (doi: 10.1109/icebe.2014.47).

Bollen, Johan, Huina Mao, and Xiao-Jun Zeng. *Twitter mood predicts the stock market*. Journal of Computational Science, 2011.

Brander, J. A., and Spencer, B. J. 2015. *"Intra-industry trade with Bertrand and Cournot oligopoly: The role of endogenous horizontal product differentiation,"* Research in Economics (69:2), Elsevier, pp. 157–165.

Castellanos, Malu, Umeshwar Dayal, Meichun Hsu, Riddhiman Ghosh, Mohamed Dekhil, Yue Lu, Lei Zhang, and Mark Schreiman. *LCI: a social channel analysis platform for live customer intelligence.* in *Proceedings of the 2011 international conference on Management of data (SIGMOD2011).* 2011.

Castillo, Carlos and Brian D. Davison. *Adversarial web search*. Foundations and Trends in Information Retrieval, 2010. 4(5): p. 377-486.

Cattani, G., Porac, J. F., and Thomas, H. 2017. *"Categories and competition,"* Strategic Management Journal (38:1), Wiley Online Library, pp. 64–92.

Ceron, A., Curini, L., Iacus, S. M., and Porro, G. 2013. *"Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France,"* New Media & Society (16:2), pp. 340–358 (doi: 10.1177/1461444813480466).

Choi, H., and Varian, H. 2012. *"Predicting the present with Google Trends,"* Economic Record (88: special issue SI), pp. 2–9.

Das, Sanjiv and Mike Chen. *Yahoo! for Amazon: Sentiment extraction from small talk on the web.* Management Science, 2007. 53(9): p. 1375-1388.

Dave, Kushal, Steve Lawrence, and David M. Pennock. "*Mining the peanut gallery: Opinion extraction and semantic classification of product reviews".* in *Proceedings of International Conference on World Wide Web (WWW2003*). 2003.

Dijkman, R., Ipeirotis, P., Aertsen, F., and van Helden, R. 2015. *"Using twitter to predict sales: a case study,"* Beta Research School, Eindhoven (available at https://arxiv.org/ftp/arxiv/papers/1503/1503.04599.pdf).

Duh, Kevin, Akinori Fujino, and Masaaki Nagata. *Is machine translation ripe for cross-lingual sentiment classification?* in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers (ACL-2011).* 2011.

Epstein, R., and Robertson, R. E. 2015. *"The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections.,"* Proceedings of the National Academy of Sciences of the United States of America (112:33), pp. E4512-21 (doi: 10.1073/pnas.1419828112).

Feldman, Ronen, Benjamin Rosenfeld, Roy Bar-Haim, and Moshe Fresko. *The Stock Sonar - Sentiment Analysis of Stocks Based on a Hybrid Approach*. *in Proceedings of 23rd IAAI Conference on Artificial Intelligence (IAAI-2011).* 2011.

Gensler, S., Neslin, S. A., and Verhoef, P. C. 2017. *"The Showrooming Phenomenon: It's More than Just About Price,"* Journal Of Interactive Marketing (38), 360 Park Ave South, New York, Ny 10010-1710 Usa: Elsevier Science Inc, pp. 29–43.

Geva, T., Oestreicher-Singer, G., Efron, N., and Shimshoni, Y. 2017. *"Using Forum and Search Data for Sales Prediction of High-Involvement Products,"* Management Information Systems Quarterly (41:1), pp. 65–82.

Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M., and Watts, D. J. 2010. *"Predicting consumer behavior with web search,"* Proceedings of the National Academy of Sciences of the United States of America (107:41), National Academy of Sciences, pp. 17486–17490 (available at http://www.jstor.org/stable/20780485).

Groh, Georg and Jan Hauffa. *Characterizing Social Relations Via NLPbased Sentiment Analysis.* in *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM-2011).* 2011.

Hancock, Jeffrey T., Lauren E. Curry, Saurabh Goorha, and Michael Woodworth. *"On lying and being lied to: A linguistic analysis of deception in computer-mediated communication."* Discourse Processes, 2007. 45(1): p. 1-23.

Hatzivassiloglou, Vasileios and Kathleen R. McKeown. "*Predicting the semantic orientation of adjectives*." in *Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL-1997)*. 1997.

Hearst, Marti. "*Direction-based text interpretation as an information access refinement*," in *Text-Based Intelligent Systems*, P. Jacobs, Editor 1992, Lawrence Erlbaum Associates. p. 257-274.

Hong, Yancheng and Steven Skiena. "*The Wisdom of Bookies? Sentiment Analysis vs. the NFL Point Spread."* in *Proceedings of the International Conference on Weblogs and Social Media (ICWSM-2010).* 2010.

Hu, Minqing and Bing Liu. "*Mining and summarizing customer reviews*." in *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004).* 2004.

Hunt, S. D., and Morgan, R. M. 1995. *"The comparative advantage theory of competition,"* The Journal of Marketing, JSTOR, pp. 1–15.

Jindal, Nitin and Bing Liu*." Identifying comparative sentences in text documents."* in *Proceedings of ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR-2006).* 2006a.

Jindal, Nitin and Bing Liu. "*Mining comparative sentences and relations."* in *Proceedings of National Conf. on Artificial Intelligence (AAAI-2006)*. 2006b.

Jindal, Nitin and Bing Liu. "*Opinion spam and analysis."* in *Proceedings of the Conference on Web Search and Web Data Mining (WSDM-2008).* 2008.

Joshi, Mahesh, Dipanjan Das, Kevin Gimpel, and Noah A. Smith*. "Movie reviews and revenues: An experiment in text regression." in Proceedings of the North American Chapter of the Association for Computational Linguistics Human Language Technologies Conference (NAACL 2010).* 2010.

Kandaswami, K., and Tiwar, A. 2014*. "Deloitte. Driving through the consumer's mind: Steps in the buying process,"* Deloitte Touch India (available at http://www2.deloitte.com/content/dam/Deloitte/in/Documents/manufacturing/in-mfg-dtcmsteps-in-the-buying-process-noexp.pdf).

Karlgren, J., Sahlgren, M., Olsson, F., Espinoza, F., and Hamfors, O. 2012. *"Usefulness of sentiment analysis,"* Lecture Notes in Computer Science (7224), pp. 426–435.

Kennedy, Christopher. "*Comparatives, Semantics of," in Encyclopedia of Language and Linguistics, Second Edition,* 2005, Elsevier.

Kotler, P. J. 1994. *"Marketing management: analysis, planning, implementation, and control"* (8th ed.), Englewood Cliffs, N.J.: Prentice Hall.

Lassen, N. B., Madsen, R., and Vatrapu, R. 2014. *"Predicting iPhone sales from iPhone tweets,"* in 2014 IEEE 18th International Enterprise Distributed Object Computing Conference, IEEE (doi: 10.1109/edoc.2014.20).

Lee, J., Park, D.-H., and Han, I. 2008. *"The effect of negative online consumer reviews on product attitude: An information processing view,"* Electronic Commerce Research and Applications (7:3), pp. 341–352 (doi: 10.1016/j.elerap.2007.05.004).

Liu, Jingjing, Yunbo Cao, Chin-Yew Lin, Yalou Huang, and Ming Zhou. "*Low-quality product review detection in opinion summarization*." in *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL-2007)*. 2007.

Liu, B. 2012. *"Sentiment analysis and opinion mining"* Synthesis Lectures on Human Language Technologies (5:1), pp. 1-167 (doi:10.2200/ S00416ED1V01Y201204HLT016).

Liu, Bing. "*Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data,"* 2006 and 2011: Springer.

Mahyoub F., Siddiqui M., Dahab M., (2015) *"Building an Arabic Sentiment Lexicon using Semi-supervised Learning".*

McGlohon, Mary, Natalie Glance, and Zach Reiter. "*Star quality: Aggregating reviews to rank products and merchants."* in *Proceedings of the International Conference on Weblogs and Social Media (ICWSM2010).* 2010.

Mihalcea, Rada and Carlo Strapparava. *"The lie detector: Explorations in the automatic recognition of deceptive language."* in *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers.* 2009.

Miller, Mahalia, Conal Sathi, Daniel Wiesenthal, Jure Leskovec, and Christopher Potts. "*Sentiment Flow Through Hyperlink Networks*." in *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM-2011).* 2011.

Mislove, A., Lehmann, S., Ahn, Y., Onnela, J., and Rosenquist, J. 2011. *"Understanding the Demographics of Twitter Users,"* Fifth International AAAI Conference on Weblogs and Social Media (N. Nicolov and J. Shanaha, eds.), Barcelona: AAAI digital library, pp. 17–21.

Mohammad, Saif. "*From Once Upon a Time to Happily Ever After: Tracking Emotions in Novels and Fairy Tales."* in *Proceedings of the ACL 2011 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH).* 2011.

Mohammad, Saif and Tony Yang. "*Tracking Sentiment in Mail: How Genders Differ on Emotional Axes."* in *Proceedings of the ACL Workshop on ACL 2011 Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA-2011)*. 2011.

Morinaga, Satoshi, Kenji Yamanishi, Kenji Tateishi, and Toshikazu Fukushima. "*Mining product reputations on the web." in Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002).* 2002.

Nasukawa, Tetsuya and Jeonghee Yi. "*Sentiment analysis: Capturing favorability using natural language processing."* in *Proceedings of the KCAP-03, 2nd Intl. Conf. on Knowledge Capture.* 2003.

Nguyen, T. H., Shirai, K., and Velcin, J. 2015. *"Sentiment analysis on social media for stock movement prediction,"* Expert Systems with Applications (42:24), pp. 9603–9611 (doi: 10.1016/j.eswa.2015.07.052).

Pang, B., and Lee, L. 2008. *"Opinion Mining and Sentiment Analysis,"* Foundations and Trends in Information Retrieval (2:2), pp. 91–231 (doi: 10.1561/1500000001).

Pang, B., Lee, L., and Vaithyanathan, S. 2002. "*Thumbs up? Sentiment classification using machine learning techniques*," Proceedings of the Conference on Empirical Methods in Natural, pp. 79–86.

Pedrycz W., and Chen Shyi-Ming, (2016) "*Sentiment Analysis and Ontology Engeneering".*

Putsis, W. P., and Srinivasan, N. 1994. *"Buying or just browsing? The duration of purchase deliberation,"* Journal of Marketing Research (31:3), US: American Marketing Association, pp. 393–402 (doi: 10.2307/3152226).

Putsis, W. P., and Srinivasan, N. 1995. *"So, how long have you been in the market? The effect of the timing of observation on purchase,"* Managerial and Decision Economics (16:2), John Wiley & Sons, Ltd., pp. 95–110 (doi: 10.1002/mde.4090160202).

Redondo J., Fraga I., Padròn I., and Comesana M., (2007) *"The Spanish adaptation of ANEW".*

Sadikov, Eldar, Aditya Parameswaran, and Petros Venetis. "*Blogs as predictors of movie success."* in *Proceedings of the Third International Conference on Weblogs and Social Media (ICWSM-2009).* 2009.

Sakunkoo, Patty and Nathan Sakunkoo. "*Analysis of Social Influence in Online Book Reviews."* in *Proceedings of third International AAAI Conference on Weblogs and Social Media (ICWSM-2009).* 2009.

Salameh M., Kiritchenko S., Mohammad Saif M., (2015) "*Sentiment after Traslation: A Case on Arabic Social Media Posts"*

Serrano-Guerrero, J., Olivas, J. A., Romero, F. P., and Herrera-Viedma, E. 2015. *"Sentiment analysis: A review and comparative analysis of web services,"* Information Sciences (311: August 2015), pp. 18– 38 (doi: 10.1016/j.ins.2015.03.040).

Suthikarnnarunai, N. 2008. *"Automotive supply chain and logistics management,"* Imecs 2008: International Multiconference of Engineers and Computer Scientists, Vols I and Ii, pp. 1800–1806.

Tong, Richard M. "*An operational system for detecting and tracking opinions in on-line discussion."* in *Proceedings of SIGIR Workshop on Operational Text Classification.* 2001.

Tuarob, S., Tucker, C. S., and Asme. 2014. "*Fad or here to stay: predicting product market adoption and longevity using large scale, social media data*," Proceedings of the Asme International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, 2013, Vol 2b (doi: V02bt02a012).

Tumasjan, Andranik, Timm O. Sprenger, Philipp G. Sandner, and Isabell M. Welpe. "*Predicting elections with twitter: What 140 characters reveal about political sentiment."* in *Proceedings of the International Conference on Weblogs and Social Media (ICWSM-2010).* 2010.

Turney, Peter D. "*Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews."* in *Proceedings of Annual Meeting of the Association for Computational Linguistics* (ACL-2002). 2002.

Wan, Xiaojun. "*Using bilingual knowledge and ensemble techniques for unsupervised Chinese sentiment analysis*." in *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-2008).* 2008.

Wiebe, Janyce. *"Identifying subjective characters* in *narrative."* in *Proceedings of the International Conference on Computational Linguistics (COLING-1990)*. 1990.

Wiebe, Janyce. "*Learning subjective adjectives* from *corpora."* in *Proceedings of National Conf. on Artificial Intelligence (AAAI-2000)*. 2000.

Wiebe, Janyce, Rebecca F. Bruce, and Thomas P. O'Hara. *"Development and use of a gold-standard data set for subjectivity classifications." in Proceedings of the Association for Computational Linguistics (ACL-1999)*. 1999.

Wiebe, Janyce. *"Tracking point of view* in *narrative."* Computational Linguistics, 1994. 20: p. 233–287.

Wijnhoven, F., and Bloemen, O. 2014. *"External validity of sentiment mining reports: Can current methods identify demographic biases, event biases, and manipulation of reviews?"* Decision Support Systems (59:1), pp. 262–273 (doi: 10.1016/j.dss.2013.12.005).

Wilson, R. E., Gosling, S. D., and Graham, L. T. 2012. "*A Review of Facebook Research in the Social Sciences,*" Perspectives on Psychological Science (7:3), pp. 203–220 (doi: 10.1177/1745691612442904).

Yang, X., Pan, B., Evans, J. A., and Lv, B. F. 2015. *"Forecasting Chinese tourist volume with search engine data,"* Tourism Management (46), pp. 386–397 (doi: 10.1016/j.tourman.2014.07.019).

Yano, Tae and Noah A. Smith. "*What's Worthy of Comment? Content and Comment Volume in Political Blogs."* in *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM 2010).* 2010.

Yu, Sheng e Subhash Kak, (2012) "*A Survey of Predicting Social Media."*

Zabin, J.,e Jefferies, A., (2008) "*Nielsen Online's BuzzMetrics custumers approach and many attain, best in class status."*

Zhang, Wenbin and Steven Skiena. "*Trading strategies to exploit blog and news sentiment."* in *Proceedings of the International Conference on Weblogs and Social Media (ICWSM-2010).* 2010.