UNIVERSITÀ CA' FOSCARI VENEZIA

Corso di Laurea Magistrale in Marketing e Comunicazione

ordinamento ex D.M. 270/2004

Tesi di Laurea

Machine learning e fattore umano nella sentiment analysis. Il caso Starbucks a Milano

Relatore:
Ch. Prof. Carlo Gaetan

Laureanda:
Consuelo Angioni

Matricola: 988064

Anno Accademico 2016/2017

Alla mia famiglia.

A Dan.

Introduzione

I thought, you know, I am a storyteller. I'm a qualitative researcher. I collect stories; that's what I do. Maybe stories are just data with a soul.

Brené Brown

Il presente lavoro di tesi nasce dall'interesse di chi scrive verso le due dimensioni principali che caratterizzano la *sentiment analysis*: il *data mining*, inteso come l'estrazione di significato da grandi quantità di dati, e l'analisi dell'espressione umana, che rende conto da una parte del linguaggio come veicolo di definizione e di affermazione del sé, dall'altra dell'evoluzione della comunicazione verso nuove forme di dialogo tra persone e comunità. La *sentiment analysis* è uno strumento che da diversi anni trova, oltre che nelle discipline della sociologia, dell'economia, della politica, ampia applicazione nell'ambito del marketing e del *brand management*. Il termine fa riferimento all'insieme di tecniche di raccolta, elaborazione e analisi dei dati testuali con il fine di rilevare valutazioni, opinioni e, più in generale, espressioni soggettive. Nel contesto del web, l'affermazione della *sentiment analysis* come strumento di ricerca di marketing è cresciuta di pari passo con la capacità di raccogliere dalle piattaforme virtuali le varie tipologie di contenuti scritti dagli utenti nelle piattaforme di *social networking*, in tempi rapidi e in enormi

quantità. La raccolta dei dati e la necessità di produrne analisi velocemente ha fatto crescere in questa direzione anche la ricerca nel campo dell'apprendimento automatico, o *machine learning*. Nell'ambito del *text mining* e, più nello specifico, della *sentiment analysis*, questo ha comportato lo sviluppo di metodologie che permettano ai computer di confrontarsi con il linguaggio umano, di elaborarlo e di comprenderlo. Proprio su questo ambito di applicazione si inserisce la ricerca che segue. Lungi dall'essere una trattazione esaustiva delle tecniche di *machine learning* e di *Natural Language Processing*, si pone l'obiettivo di descrivere e quindi di sfruttare alcune di queste tecniche per l'analisi di un caso reale di monitoraggio della *brand reputation* e di *brand management*. Per raggiungere questo obiettivo, si è deciso di partire dalla prima dimensione della *sentiment analysis*, ovvero dal suo essere analisi del linguaggio umano e della sua componente soggettiva. I primi capitoli si soffermano quindi sulla sfida delle tecniche di elaborazione automatica del testo, rendendo conto della peculiarità del linguaggio naturale come strumento di espressione e di conversazione, così come della sua evoluzione ai tempi di Internet: sia nei riguardi del trasformazione della forma, sia nei riguardi dei contenuti e dei contesti in cui i messaggi vengono trasmessi. L'introduzione sul linguaggio naturale apre anche all'analisi dell'evoluzione della comunicazione, di cui il linguaggio è ancora oggi il principale strumento: si offrirà una panoramica del passaggio da comunicazione di massa a mercato conversazionale, rendendo conto della nozione di Web 2.0 e delle opportunità che i contenuti generati dagli utenti (*User Generated Content*) rappresentano per le aziende nelle arene online. I successivi capitoli si occupano più specificamente della *sentiment analysis*. Nel secondo capitolo, si offre una panoramica degli strumenti e delle nozioni relative a questo tipo di analisi testuale nello specifico contesto della ricerca di marketing. A partire dalla definizione

di *sentiment analysis* come sintesi degli approcci qualitativi e quantitativi, si illustrano brevemente gli aspetti metodologici riprendendo la letteratura su questa tematica, offrendo una descrizione delle differenze di approccio, quindi spiegando gli strumenti che verranno impiegati nella parte successiva: il social network Twitter e l'ambiente R, dove verrà effettuata gran parte dell'analisi sfruttando modelli basati su *machine learning*. Infine, nel terzo capitolo, si affronterà la parte di applicazione di *sentiment analysis* su un caso reale. Utilizzando i dati ricavati da Twitter tra gennaio e febbraio 2017, si analizzerà la reputazione del marchio Starbucks per gli utenti italiani in relazione alla notizia dell'apertura del primo negozio a Milano e alla campagna di posa delle palme in piazza Duomo. L'analisi sarà dapprima affrontata con alcuni dei modelli classici del *machine learning* per la classificazione supervisionata, utilizzando il software R e il codice sorgente che viene riportato in appendice. In seguito, invece, si approfondirà l'indagine impiegando il modello iSA della piattaforma Voices Analytics, in particolare sviluppando l'analisi da *sentiment* a *opinion mining*, quest'ultima intesa come evoluzione rispetto alla mera identificazione della classificazione dell'espressione soggettiva verso una più comprensiva rappresentazione delle dimensioni collegate a quell'espressività. In questa fase si cercherà di sottolineare il limite di una *sentiment analysis* che sia ridotta alla sola classificazione della polarità e si mostrerà l'importanza di sviluppare la ricerca partendo prima dalla rilevanza del contesto e dalla conoscenza approfondita del fenomeno. Si vedrà in questo senso come il ruolo del ricercatore (il "fattore umano" citato nel titolo) risulti fondamentale per non standardizzare l'approccio e permettere una corretta impostazione dell'indagine, assicurando che la parte di processo affidata alla macchina sia stata correttamente indirizzata.

Indice

Capitolo 1

L'EVOLUZIONE DEL CONTENT

"Meow" means "woof" in cat.

George Carlin

1.1 Il linguaggio

1.1.1 Il linguaggio naturale

Quando si parla di *sentiment analysis* si entra nel campo del *Natural Language Processing*, ovvero di quel settore dell'Intelligenza Artificiale e delle discipline informatiche che si occupa del rapporto dei computer con il linguaggio umano, o "linguaggio naturale". La comprensione del linguaggio naturale è, infatti, la principale sfida dell'AI e delle tecniche di *deep learning*, attraverso le quali è possibile insegnare alle macchine a riconoscere parole, comprendere testi, comunicare con gli umani [10]. Secondo la linguistica e la filosofia del linguaggio, il linguaggio naturale è un qualsiasi linguaggio che

sia evoluto in modo spontaneo nel contesto umano, attraverso il suo stesso uso. Dopo molti secoli di linguaggio mediato quasi esclusivamente in forma scritta tradizionale o verbale, oggi abbiamo una molteplicità di mezzi con cui veicoliamo messaggi, usando anche forme iconografiche e combinazioni simboliche complesse che sono rese possibili da strumenti di comunicazione molto più pervasivi e diffusi rispetto al passato. Appare chiara allora l'utilità di rendere conto, nell'ambito dell'analisi del *sentiment*, dell'evoluzione di questo linguaggio: solo comprendendo questa evoluzione si possono infatti porre le basi di questo studio e vedremo nei prossimi paragrafi come questa riguardi tanto la forma (*come* comunichiamo) quanto il contenuto (*di cosa* parliamo).

Le parole che usiamo e le strutture grammaticali che costruiscono il nostro linguaggio hanno un profondo effetto sul modo stesso in cui pensiamo. Secondo alcuni studiosi, dal momento in cui il pensiero è di per sé costituito di parole e frasi, le persone non sono in grado di concepire pensieri che includano cose per le quali non esistano nomi che le identifichino: sarebbe quindi il linguaggio a definire e causare il pensiero, non il contrario, come suggerisce Pinker [77]. Sono numerosi i filosofi del linguaggio che affermano l'essere umano sia definito all'interno del linguaggio che usa e, di conseguenza, profondamente influenzato da esso. "Il limite del mio linguaggio costituisce il limite del mio stesso mondo": secondo il punto di vista di Wittgenstein, il linguaggio precederebbe il pensiero, di cui tuttavia è espressione [104].

L'indagine sul linguaggio naturale si complica se lo esploriamo nel contesto della conversazione. Il linguaggio della collettività è, come detto all'inizio, costruito sullo stesso uso che se ne fa. Questo implicherebbe che non esista un linguaggio che non sia intrinsecamente legato alla comunità che lo utilizza e, con l'uso, legittima: "Ogni linguaggio è sempre un fatto sociale: ogni

soggetto impara a parlare un linguaggio, distinguendo ciò che è grammaticalmente e lessicalmente corretto da ciò che non lo è, sulla base dei segnali positivi e negativi (di approvazione o disapprovazione, di accordo o disaccordo) che riceve nella propria comunità" [104]. In questo senso, il linguaggio naturale è da distinguere rispetto al linguaggi formali, come i linguaggi di programmazione o le regole linguistiche utilizzate nell'ambito della logica. Questa distinzione è, ad avviso di molti studiosi di linguistica, filosofia del linguaggio e neuropsicologia, fondamentale per comprendere la principale difficoltà di insegnare ad un calcolatore come processare il linguaggio ordinario. Proprio per la sua natura di strumento la cui nascita ed evoluzione sono basati sull'uso e la consuetudine, il linguaggio umano sfugge ad una coerenza formale e perfettamente finita. L'analisi logica del linguaggio contiene in questo senso, avverte Heisenberg, "il pericolo di una eccessiva semplificazione" [49]. L'importanza che il significato secondario di una parola riveste nella comprensione di un'intera frase è un tema di cui un approccio basato sulla sola logica del linguaggio non può rendere conto: "Il fatto che ogni parola può produrre molteplici movimenti, più o meno coscienti, nella nostra mente, può essere usato per rappresentare, attraverso il linguaggio, alcune parti della realtà molto più chiaramente di quanto non avvenga attraverso l'uso degli schemi logici" [49].

Alcuni critici si sono definiti particolarmente scettici nei confronti di una possibilità da parte dei computer di comprendere il linguaggio umano, a cominciare dal fatto che per gli stessi umani, come sostenuto da Chomsky, i meccanismi sottostanti al linguaggio sono ancora stati finora solo parzialmente compresi [22]. La complessità del linguaggio, infatti, ha a che fare anche con i modi che permettono agli umani di capirsi gli uni con gli altri, che non si riducono alla mera espressione verbale: è, questa, una delle differenze più

significative del linguaggio naturale rispetto al linguaggio artificiale. Come afferma ancora Pinker [77], infatti, il linguaggio è compreso a molteplici livelli, invece che come diretta analisi del contenuto di una frase. Nella vita di tutti i giorni ci aspettiamo la capacità dell'interlocutore di "leggere tra le righe", intendere i sottesi, cogliere l'ironia e le altre figure retoriche che non sempre il testo scritto veicola in modo efficace. In alcuni casi, invece, il lingaggio presenta dei problemi di interpretazione per gli uomini stessi. Uno degli aspetti più interessanti rilevati dagli studi di *Natural Language Processing*, che emerge anche nell'ambito della *sentiment analysis* stessa come vedremo in fase di applicazione, è la difficoltà di attribuire con certezza e univocità un certo significato ad una proposizione umana: un tema, questo, su cui si sono confrontati numerosi pensatori del linguaggio nel corso dei secoli. Uno dei problemi del dialogo con l'altro è fondato sulla impossibilità di avere l'assoluta certezza che la comunicazione sia efficace, anche laddove sembri di parlare un linguaggio comune: "Le conversazioni si riducono sovente a monologhi paralleli [...] Si crede di scambiare idee e si ha solo uno scambio di parole, e le parole percepite non ci comunicano le idee di coloro che ce le offrono, risvegliano in noi solo le nostre. Non ci viene mai dato se non quello che avevamo" [12]. La impossibilità della certezza di efficacia della conversazione si avverte ancora più se questa è scritta e quindi privata di quegli elementi di contesto che, nel linguaggio orale, diventano indizi per permettere la comprensione: si pensi al tono della voce e alla gestualità che consentono all'umano di riconoscere l'ironia e il sarcasmo, esempi tipici di ambiguità in cui il senso letterale del testo non è sufficiente ed, anzi, produce l'effetto opposto a quanto l'emittente del messaggio vuole comunicare [77].

Ci sono poi i problemi derivati dalle "grammatiche ambigue" [61]: "I segni hanno sempre lo stesso significato, mentreché le parole ne hanno parecchi". Si

riporta l'esempio classico tra i linguisti della frase "la vecchia porta la sbarra": non c'è modo di attribuire il corretto significato a questa proposizione se non con un minimo di contesto [61]. La scelta di fronte all'ambiguità richiede sempre, quindi, capacità di cogliere il sotteso. Se l'ambiguità del linguaggio naturale è ciò che permette l'esistenza di giochi e indovinelli basati ad esempio sulla *polisemia*, il contesto sarebbe ciò che permette di risolverli. La nozione di contesto è però a sua volta vaga: può andare dalle frasi immediatamente adiacenti fino all'intero discorso in cui la frase è inserita; per questo motivo delimitare il contesto necessario e sufficiente a permettere la comprensione di una frase è a sua volta un'impresa di difficile formalizzazione, che include una certa dose di ragionamento [61].

1.1.2 L'evoluzione della forma

Parallelamente agli studi sull'Intelligenza Artificiale, la natura di cambiamento incessante che è caratteristica propria del linguaggio umano ha continuato ad occupare i linguisti, muovendosi di pari passo con l'evolversi della società. Come avverte Niola [70], ad ogni cambio epocale si accompagna "un sobbalzo della lingua" e di questo cambiamento lo specchio più evidente è offerto dalla comunicazione sul web.

L'evoluzione riguarda, prima ancora che i contenuti, la forma. Negli ultimi anni, Internet è diventato un laboratorio per lo studio della lingua: non solo offre un gigantesco *corpus* di linguaggio reale usato da persone reali, ma agisce anche come veicolo incredibilmente efficiente per la trasmissione delle idee, permettendo di evidenziare esempi di linguaggio che le persone trovano sufficientemente interessanti da condividere tra di loro [77]. La creazione di nuove parole, strutture e grammatiche di cui l'uso e la ripetizione permettono la legittimità è qualcosa che è sempre esistito e che, come anticipato prece-

dentemente, fa parte della natura stessa del linguaggio umano. Tuttavia, l'esistenza di Internet sembra aver rinforzato e accelerato questo processo. L'innovazione del linguaggio passa per l'introduzione di nuove parole, per l'attribuzione di nuovi significati a parole già esistenti, per la rottura delle regole e per l'invenzione di nuovi veicoli di significato e di comunicazione che vanno oltre le parole stesse (si pensi, ad esempio, ai *meme*).

Secondo alcuni studiosi, non è il linguaggio di per sé che sta evolvendo più rapidamente rispetto al passato, ma la tecnologia che permette la trasmissione di nuove parole e nuovi significati da gruppi di persone ad altri in modo più veloce [23]. La tecnologia cui fa rifierimento l'autore è quella degli strumenti con cui ci connettiamo ad Internet e alla rete di contatti web per comunicare gli uni con gli altri, attraverso il *texting*. Questa forma di comunicazione è la principale fonte di diffusione di nuovi termini e nuove strutture. McWorther [64] a questo proposito afferma che il *texting* ha ormai sviluppato la sua propria grammatica, da non intendersi come sostituiva di quella tradizionale. L'evoluzione del linguaggio attraverso Internet e i social network avrebbe infatti a che fare con il progressivo avvicinarsi della scrittura al parlato. McWhorter [64] in questo senso definisce il *texting* come *"fingered speech"*. La brevità e la velocità permesse dagli strumenti che utilizziamo per comunicare hanno rapidamente consentito l'abbandono di molti dei vincoli che la scrittura formale ha sempre richiesto, considerati per certi versi superflui: la punteggiatura, ad esempio, o le regole di ortografia come la lettera maiuscola dopo il punto fermo. Questo processo finirebbe in realtà per avvicinare il linguaggio di Internet al linguaggio del parlato: nessuno, sottolinea McWorther, pensa ai punti o alle lettere maiuscole mentre sta parlando. Il *texting* in questo senso è *"speech"*, dialogo, che rende il più possibile fluida e spontanea la conversazione snellendo la struttura delle frasi e accorciando il

tempo tra formulazione del messaggio, ricezione, risposta [64]. Questo, avverte McWorther, non significa che la struttura del linguaggio verrà rimossa, ma che ci troveremo ad avere a che fare con una struttura nuova e quindi con una nuova complessità. In questo senso, il *texting* è da intendere come forma di comunicazione diversa, non invece come una evoluzione della scrittura.

1.1.3 L'evoluzione del contenuto

Oltre all'aspetto esteriore del linguaggio, l'oggetto della comunicazione è evoluto con e attraverso Internet – ed è proprio sul contenuto che l'analisi del *sentiment* è chiamata a confrontarsi. Oggi lo scorrere della comunicazione porta a una sorta di "flusso di coscienza individuale e collettivo insieme", scrive Niola [70]: il flusso è registrato nelle piattaforme web e il suo accesso è permesso dall'esistenza delle parole chiave e del concetto di *tag*. "L'*hashtag* cambia il regime di senso alla parola": si parla a questo proposito di *augmented word*, una parola come "chiave" che "apre la porta di un mondo di connessioni e significati. [...] Gli *hashtag* quindi sono delle schegge generate dal big bang dell'universo digitale. Proprio da questi nascono delle combinazioni e classificazioni inedite [...] Se la rete è disseminazione e condivisione, allora il tag è il verbo del moderno internet"[70]. Il flusso di comunicazione, catturato dal concetto di Web 2.0 [102], vede l'individuo più disposto a condividere informazioni personali e a comunicarle ad audience più ampie. Lo stile più informale ed aperto di cui si è accennato precedentemente può essere visto come lo specchio di una informalità ed apertura che riguardano anche i contenuti che vengono condivisi nel web: succinti e più diretti. La disponibilità dell'utente ad esprimersi sui temi più disparati è oggetto di numerosi studi che riguardano la dualità reale-virtuale, il tema dell'identità online, la questione della affidabilità e della autenticità delle espressioni del

sé che avvengono nel contesto del web.

È però in particolar modo nei riguardi dell'espressione della soggettività, cioè della affermazione di una opinione o di uno stato d'animo, che la capacità di comprendere il significato di un testo umano diventa rilevante nell'ambito della ricerca di marketing. Il *sentiment*, inteso appunto come risposta connotata in senso positivo o negativo e capace di includere componenti sia razionali che emotive, non è nuovo come oggetto di studio del marketing: è anzi uno degli indicatori principali con cui analizzare i comportamenti di acquisto. Le modalità con cui gli esseri umani danno espressione delle proprie percezioni soggettive e delle proprie opinioni sono state profondamente influenzate dal web. Di questo, tra tutte le forme di applicazione di Intelligenza Artificiale, deve rendere conto la *sentiment analysis*, e vedremo nei prossimi paragrafi in che modo l'evoluzione degli strumenti di comunicazione abbia aperto la strada all'opportunità che questo tipo di ricerca offre, tra i vari ambiti di applicazione, al marketing.

1.2 La comunicazione

1.2.1 Nuovi modelli di comunicazione

L'analisi dell'evoluzione della comunicazione di massa, intesa come classe dei fenomeni comunicativi che si basa sull'uso dei media [71] è da ritenersi propedeutica ad una comprensione dello scenario di riferimento di questo lavoro di tesi. In particolare, gli strumenti utilizzati dalle imprese per la comunicazione aziendale si sono evoluti parallelamente con l'emergere di quello che è stato definito in precedenza "Web 2.0". In questo nuovo contesto, anche il concetto di "comunicazione di massa" è cambiato, con una rottura di para-

digma che ha per certi versi messo in discussione anche modelli e ruoli della comunicazione di massa stessa.

Negli anni '30, Harold Lasswell concepisce la *bullet theory* [97], tra i capisaldi degli studi di comunicazione dei media di massa. Il modello su cui è basato il filone di pensiero di cui Lasswell è precursore vede la dinamica comunicativa come sostanzialmente unidirezionale: l'individuo viene colpito singolarmente da un messaggio come da una pallottola, in un moto da emittente a destinatario in cui il soggetto che riceve il messaggio svolge un ruolo di totale passività. L'altro presupposto di questa teoria della comunicazione di massa è che gli individui siano colti in modo separato e non, invece, come nodi di una rete in cui il messaggio, circolando, viene richiamato e riprodotto dagli individui che a loro volta contribuiscono alla sua diffusione ed alla sua trasformazione. Il modello di Lasswell, pur fondamentale nel costituire un primo approccio allo studio della relazione tra media e singoli, è stato in seguito messo in discussione anche alla luce dell'evoluzione dei media stessi. Il presupposto da cui partiva Lasswell, in cui l'individuo è slegato dai rapporti con gli altri membri della società che sono a loro volta destinatari dei messaggi lanciati dai media, semplifica fortemente il modello comunicativo e risulta riduttivo nello spiegare le dinamiche che oggi, più di prima, costituiscono la base della comunicazione al tempo del World Wide Web. Questo punto, ovvero la rilevanza delle relazioni con gli altri individui nella comunicazione, assieme alla messa in discussione della totale passività del destinatario rispetto all'emittente (approccio *behaviorista* che ricalca il modello "stimolo-risposta") sono la base delle teorie che spiegano l'evoluzione della comunicazione moderna e i fenomeni di comunicazione più recenti, tra cui la crescente rilevanza dei social media e la nascita dello *User Generated Content* [38].

La scuola di Lazarsfeld metterà in discussione gli assunti di base della *bullet theory*, rinobilitando il ricevente non più come bersaglio passivo, ma come consumatore attivo e, soprattutto, non isolato. In quest'ottica, l'individuo è a sua volta attore della comunicazione e compone, assieme agli altri individui dell'ambiente sociale, la rete di rapporti che non solo permette ai messaggi della comunicazione di circolare, ma anche di essere integrati e alimentati dai contributi individuali degli attori che compongono la rete stessa. Il potere del media viene in questo senso ridimensionato e affiancato al potere forte che assume invece la "massa". Anche il concetto di massa viene riconsiderato non come insieme indifferenziato e uniforme di tanti singoli, bensì come presenza contemporanea di più segmenti e comunità. Grazie ai contributi della scuola di Lazersfeld, Berelson e Gaudet, gli studi sulla comunicazione rivalutano l'importanza dei contatti personali e dell'influenza dei singoli sulla circolazione dei messaggi e sulla persuasione degli individui [38]. Anche in quest'ottica si assiste alla crescente rilevanza della conversazione, intesa come dialogo e scambio, evoluzione rispetto alla comunicazione, intesa come mera trasmissione di un messaggio. Tra i presupposti di questo nuovo approccio c'è anche quello della differenziazione dei nodi che compongono la rete sociale e quindi del diverso peso da attribuire ad alcuni nodi rispetto ad altri. Richiamando le nozioni proprie della Analisi delle Reti Sociali [26], l'attribuzione di una diversa capacità di comunicazione e di persuasione ad alcuni individui rispetto ad altri, e ancor più la maggiore influenza di alcuni individui rispetto anche ai media di massa, sposta l'attenzione verso gli *opinion leader* e gli *influencer*, complicando ulteriormente il modello delle dinamiche comunicative in cui le stesse aziende si trovano a lanciare i loro messaggi [26]. Lo sviluppo delle reti sociali online tipiche del Web 2.0 ha dimostrato l'abilità di generare comunità online a crescita molto rapida, dove

gli utenti comunicano, condividono informazione e si mantengono in contatto, spesso anche senza conoscersi direttamente gli uni con gli altri. Gli studi riguardanti il flusso dell'informazione all'interno della società (prima ancora che nel web), cominciarono già negli anni '50, in cui venne avanzata ed analizzata l'ipotesi che questo procedesse partendo dai *mass media*, arrivando direttamente agli *opinion leader* e poi, da questi, al resto della popolazione. Si parla in questo senso del *"two-step communication flow"* [56] e di come l'*influencer* sia un soggetto decisivo nel definire le decisioni di acquisto dei consumatori [26].

Indipendentemente dalla maggiore o minore rilevanza degli *influencer* nell'ambito della creazione e diffusione delle idee [1], appare chiaro che un modello a flusso unidirezionale, così come qualsiasi altra rappresentazione che non tenga conto di tutti gli attori della rete e della natura composita delle sue diramazioni, risulti riduttivo.
Prendendo ad esempio a riferimento il modello di McQuail e Windahl, si individuano due processi paralleli: la ricezione del messaggio e la risposta allo stesso, in cui la ricezione non equivale ad una risposta così come la non-ricezione non implica necessariamente una mancanza di risposta [63]. Secondo questo modello, il destinatario esposto al messaggio può esprimere una risposta che, quando rilevata dalla fonte, si trasforma in *feedback*, ovvero una nuova comunicazione "di ritorno". È grazie all'esistenza del *feedback*, ovvero

[1] Come riporta D'Adda [26], studi più recenti limitano al contrario il ruolo degli *influencer* all'interno di un network, indicando invece come primarie le relazioni interpersonali che intercorrono tra utenti ordinari e quindi la prontezza e la predisposizione di una società nell'adottare un'innovazione.

alla capacità del mittente del messaggio di recepire la reazione del destinatario, che la comunicazione evolve rispetto al modello della comunicazione di massa delineato precedentemento, verso un processo di natura interattiva e multidirezionale. Secondo questo nuovo modello, il messaggio originario si distribuisce sulla rete sociale in modo non controllato e imprevedibile: più di prima diventano fondamentali le diverse rilevanze degli individui di fronte alle campagne dei media, come attori a loro volta del processo comunicativo. Con la diffusione di Internet nasce e cresce il ruolo del contenuto creato e pubblicato sul web dai suoi utilizzatori, nella forma di post, microblog, immagini, file audio e video. La condivisione in prima persona da parte di chi prima era considerato solo ricevente di messaggi, unita alla possibilità per ciascun singolo di raggiungere audience molto più ampie della propria cerchia di relazioni personali, fa parte dei cambiamenti portati dalla digitalizzazione di cui la comunicazione aziendale inizia ad occuparsi con maggiore interesse. Si può rivedere il cambiamento di paradigma rispetto a quello della comunicazione di massa descritta da Lasswell considerando lo schema delle variabili di comunicazione [75]:

a. La tipologia del flusso: unidirezionale o bidirezionale. La prima comporta la trasmissione di un messaggio da un emittente a uno o più riceventi senza contemplare una risposta (*i.e.* televisione, radio, stampa ecc), mentre la seconda prevede un'interazione di tipo circolare. Nell'evoluzione rispetto alla comunicazione di massa descritta da Lasswell, il flusso adesso è bidirezionale, in quanto prevede una comunicazione di ritorno.

b. Il modello di comunicazione: *one-to-many*, *one-to-one*, *many-to-many*. Il modello *one-to-many* è tipico della comunicazione di massa e comporta il contatto a distanza tramite un mezzo tra una fonte e una pluralità

di soggetti simultaneamente: il flusso è pacchettizzato, il contenuto è precodificato e l'audience non partecipa. Il Web 2.0 ha consentito un'evoluzione della comunicazione verso gli altri due modelli: il modello *one-to-one* considera invece un flusso a due vie, interattivo, personalizzato tra una fonte e singolo destinatario; il modello *many-to-many* prevede un'interazione tra una pluralità di soggetti, che scambiano tra loro le informazioni, creando contenuti (forum, newsgroup, comunità virtuali).

1.2.2 Il monitoraggio del *content*

Il contenuto generato dagli utenti e l'evoluzione della comunicazione portano ad una rivalutazione del mondo del web non solo come piattaforma di scambio di contenuti, ma anche di potenziale rivoluzione del consumo e del rapporto con le aziende. I brand si inseriscono negli ambienti di produzione dal basso proprio alla luce di questa evoluzione della comunicazione, che venne concettualizzata nel 1999 da Rick Levine, Christopher Locke, Doc Searls e David Weinberger: Internet come uno spazio di diffusione delle informazioni nuovo, capace di stravolgere completamente la natura della comunicazione di marketing, sarà la premessa teorica del *Cluetrain Manifesto* [59]. In questa sede non ci occuperemo di approfondire gli studi sulla rilevanza del consumatore *prosumer* ed il concetto del "mercato conversazionale". Interessa però sottolineare l'impatto che la creazione del cosiddetto "Web 2.0" ha in particolare nell'espandersi degli studi e degli strumenti che rientrano nel campo dell'AI e dell'elaborazione del linguaggio naturale. Indipendentemente dalle conclusioni tratte da Locke, Searls e Weiberger, le premesse mettono in luce quantomeno un punto di partenza difficilmente discutibile: il pensiero e le opinioni dei consumatori sono più accessibili rispetto a prima. Questa ac-

cessibilità, permessa dallo spontaneo "avere una voce" di cui Internet dota ogni suo utente, non è da considerare soltanto nell'ambito della produzione di recensioni e commenti in merito ai prodotti e ai servizi offerti dalle aziende, o alla risposta a iniziative promozionali e campagne. Riguarda, invece, qualsiasi forma di espressione del consumatore, delle sue opinioni e dei suoi sentimenti, non necessariamente in risposta ad un sollecito specifico che venga dal brand nell'ambito di una campagna di comunicazione online. Il potenziale in termini di capacità di catturare "ciò che pensa il consumatore" è quindi ovviamente più ampio. Alcuni tratti della prossima evoluzione del Web 2.0 sono già oggetto di studio e di ricerca: la trasformazione di Internet come enorme database "semantico" [52] i cui documenti sono inseriti in connessioni che non si riducono al collegamento ipertestuale, quindi del *"data web"* come struttura di partenza per il diffondersi dell'Intelligenza Artificiale.

L'idea di un Internet "evoluto", capace di apprendere il linguaggio umano e di sfruttarlo per migliorare l'esperienza di navigazione dell'utente, presenta ovvie implicazioni anche nel mondo della *sentiment analysis*. Un'analisi avanzata, capace di andare oltre le parole chiave e di comprendere i significati e i pensieri espressi dai contenuti semantici che popolano il web, permetterebbe di identificare il sentimento dei testi prodotti dagli utenti e, ancor più, di usarlo in tempo reale. La rilevanza del monitoraggio del *content* si vede allora in più di un momento del ciclo di vita della comunicazione: dall'iniziare una campagna, al correggere il tiro di una già esistente, al comprendere in tempo reale gli effetti di un evento, fino al prevederli. Dal punto di vista del marketing, monitorare l'influenza che il contenuto prodotto dal proprio brand ha sugli utenti, così come quella che che il contenuto prodotti dagli utenti ha sul brand, diventa cruciale nell'ambito di un *brand management* evoluto: laddove la costruzione di un marchio e il suo posizionamento quantomeno

virtuale sono diventati più veloci e accessibili, il suo controllo e monitoraggio sono invece più difficili e richiedono strumenti capaci di competere con la velocità e la quantità dei dati prodotti nel web [82].

Capitolo 2

BREVE INTRODUZIONE ALLA SENTIMENT ANALYSIS

2.1 La *sentiment analysis* in teoria

2.1.1 Metodi quantitativi e qualitativi nella ricerca di marketing

Nell'ambito della ricerca di marketing, è solito distinguere il metodo quantitativo da quello qualitativo [58]. Mentre il primo verte sulla rappresentatività di un'intera popolazione, estendendo i risultati ottenuti su un campione estratto dal segmento che si vuole analizzare, il secondo si concentra invece sull'esplorazione di un fenomeno in modo destrutturato e più libero, al fine di ricavare *insight* che riguardano sentimenti, impressioni e opinioni di un ristretto gruppo di soggetti su cui la ricerca viene condotta. La tipologia di informazione che il ricercatore ottiene è quindi soggettiva e non traducibile

numericamente; il ricercatore rileva direttamente i dati attraverso interviste, *focus group* e tecniche che implicano il contatto diretto con il soggetto. Le ricerche quantitative, al contrario, hanno l'obiettivo di misurare un fenomeno definendo delle variabili di interesse e generalizzando sull'intera popolazione i risultati ottenuti su un ampio campione. La chiave della ricerca quantitativa è l'approccio matematico e formalizzato al problema che si basa su un'automatizzazione dell'analisi, in cui la raccolta dei dati e l'estrazione di significato dagli stessi è affidata ad una sequenza di istruzioni standardizzate. Questo limita particolarmente l'intervento della soggettività del ricercatore e il suo ruolo nell'interpretazione dell'analisi [58].

Tra i due approcci è storicamente aperto un dibattito in merito ai loro limiti e alla loro validità [80]. Se da una parte i dati raccolti con i metodi di ricerca qualitativa rischiano di essere viziati dalla interpretazione soggettiva del ricercatore e vengono accusati di non fornire informazioni traducibili in termini numerici e statistici, dall'altra la ricerca di tipo qualitativo ha proprio nei suoi limiti anche la sua forza rispetto ai metodi quantitativi. Questa forza è riassumibile nel concetto del "fattore umano", inteso proprio come la capacità del ricercatore di cogliere le sfumature e le possibili variazioni di significato che l'approccio quantitativo tende invece a ridurre o eliminare [80]. Vedremo più avanti come il dibattito, soprattutto nell'ambito dell'analisi del *sentiment*, verta proprio sulla stessa considerazione del "fattore umano" come elemento di forza o di debolezza rispetto alla ricerca: parte della letteratura considera i metodi quantitativi come fallimentari nel comprendere in modo corretto i dati proprio in quanto l'interpretazione umana, che viene esclusa, è vista invece come fondamentale nel catturare il messaggio contenuto in un testo.

2.1.2 La ricerca di marketing nel contesto del web

Con l'avvento di Internet e del World Wide Web, la raccolta dei dati su cui operare con i metodi tradizionali di ricerca, sia quantitativi che qualitativi, è cambiata. Come accennato nel Capitolo 1 di questa tesi, l'accesso al contenuto generato dagli utenti è permesso in modo immediato e spontaneo: questo si traduce nella possibilità di raccogliere informazioni sull'opinione di una popolazione, ovvero sulla espressione soggettiva e personale di un'idea. Il web e in particolar modo le piattaforme di *social networking* sopracitate offrono al marketing velocità di ricerca e grandi quantità di dati a disposizione. Anche per la ricerca qualitativa, gli strumenti tradizionali come il *focus group* e i questionari, se effettuati online, consentono di raccogliere volumi di dati decisamente maggiori in minor tempo.

Dall'altra parte, la ricerca di marketing sul mondo online presenta ancora almeno un limite significativo, avverte Kotler: la popolazione che ha accesso ad Internet e che utilizza il web come canale di espressione dei suoi pareri, pensieri e intenzioni non è certamente rappresentativa dell'intera popolazione cui potrebbe volersi rivolgere il ricercatore [58]. Secondo il *Digital in 2017 report* [30], durante il 2016 il numero di persone che si sono connesse in Italia a Internet è cresciuto del 4% rispetto all'anno precedente (39.21 milioni di persone), e dell'11% quello relativo all'uso dei social media: un totale di 28 milioni, che corrisponde a una penetrazione del 47%. Per quanto riguarda il nostro Paese, nonostante l'aumento di italiani che navigano su Internet abitualmente ed utilizzano piattaforme social per comunicare, è da tenere in considerazione il fatto che questo rappresenti un sottogruppo dell'intera popolazione: il totale di persone residenti in Italia che accedono mensilmente a piattaforme social è infatti corrispondente al 52% [30]. Come sottolinea ancora Kotler [58], la ricerca di marketing effettuata online non è adatta per

ogni compagnia o prodotto: la sua rappresentatività è strettamente legata a variabili come il grado di informatizzazione della popolazione, la tipologia di consumatore che utilizza i social per comunicare, il tipo di piattaforma che viene considerata nell'indagine.

Un altro aspetto di cui tenere conto quando ci si approccia ad un'analisi sul web è che, a differenza dell'indagine offline, i messaggi pubblicati online sono scritti spontaneamente dagli utenti e pervenuti al ricercatore in modo sporco e non ordinato. Ancora, è da evidenziare il valore e allo stesso tempo il limite di questo tipo di dati: da una parte, la loro natura non strutturata che richiede uno sforzo superiore rispetto alla tipica indagine offline che segue un copione standardizzato; dall'altra, la spontaneità del dato ricevuto dall'utente e non indirizzato dal ricercatore, che consente di raccogliere opinioni libere e non guidate, rivelando anche collegamenti e informazioni non inizialmente preventivati [32]. Ancora, le opinioni raccolte possono provenire da utenti che una tipica indagine offline non raggiungerebbe: si pensi ai soggetti che eludono le interviste e i questionari, o che rispondono in modo non completamente sincero agli input ricevuti dall'intervistatore. In questa sede, vedremo appunto la *sentiment analysis* come sintesi tra il metodo quantitativo e il metodo qualitativo in un contesto di ricerca online, ovvero di estrazione di informazioni da Internet, con la consapevolezza dei limiti, ma anche dei vantaggi già citati, che un'indagine su una popolazione esclusivamente online ovviamente presenta.

2.1.3 La *sentiment analysis* online come sintesi di metodi qualitativi e quantitativi

L'orientamento di questo lavoro di ricerca è quello di proporre la *sentiment analysis* come metodo capace di ovviare ai limiti dei metodi quantitativo e

qualitativo di ricerca, sfruttando i vantaggi della ricerca di marketing online e, in particolar modo, della ricchezza dei contenuti prodotti spontaneamente dagli utenti nelle piattaforme di *social networking*. L'analisi del sentimento della popolazione di Internet raccoglie infatti la natura dell'informazione che è tipica delle ricerche qualitative, ovvero sentimenti, impressioni, opinioni dei consumatori su un brand, un prodotto o un messaggio; dall'altro, rimuove i limiti di quegli stessi metodi riducendo al minimo la soggettività e sfruttando il rigore degli strumenti statistici che sono tipici dei metodi quantitativi [80]. Pur sottolineando nuovamente i confini di rappresentatività nel considerare la sola popolazione online, l'applicazione dell'analisi sul contesto del web permette l'accesso ad un volume di dati che aumenta la forza dello strumento di catturare un campione molto più ampio dei metodi tradizionali qualitativi. I dati sul *sentiment* e sull'opione degli utenti, raccolti in modo passivo e successivo rispetto a quando vengono generati, quindi non seguenti ad un input ricevuto dall'intervistatore (in questo senso, dunque, non condizionati) sono poi processati con i vantaggi tipici dei metodi quantitativi, ovvero quelli di misurabilità e oggettività [80].

2.1.4 La *sentiment analysis* e il fattore umano

Torna a questo punto opportuno richiamare l'importanza del "fattore umano" di cui sopra. Uno dei testi di riferimento per questa tesi è il volume di Iacus, Ceron e Curini [20]. Nel loro lavoro di ricerca, gli autori propongono una lettura della *sentiment analysis* come tecnica capace di estrarre informazioni dai social network online in una modalità che recupera la soggettività del ricercatore, altrimenti eliminata dal metodo di ricerca puramente quantitativo, limitandola però ad una fase in cui il "fattore umano" assume maggiore rilievo nella corretta interpretazione del dato. In particolare l'a-

nalisi, parafrasando gli autori, deve essere in grado di estrarre il *sentiment* nello stesso modo in cui questo verrebbe individuato ricorrendo ad un *focus group* o ad un'intervista individuale, interpretando cioè nel modo più esatto l'input ricevuto dal soggetto. Del resto, come si è provato a dar conto già nel primo capitolo, "la complessità del linguaggio è così vasta che qualunque metodo completamente automatico non può che fallire"[20]. Vedremo allora in seguito come la sfida maggiore della *sentiment analysis* come tecnica di indagine che segue una procedura automatizzata e standardizzata risieda proprio nel superamento dei limiti della capacità di un modello statistico e quantitativo di cogliere correttamente l'espressione veicolata attraverso il linguaggio umano.

2.2 Strumenti e metodi per la *sentiment analysis*

2.2.1 Confronto tra gli approcci *machine learning* e *lexicon-based*

Le tecniche e le tipologie di *sentiment analysis* descritte in letteratura sono molteplici. Una prima distinzione riguarda i due approcci più frequenti con cui si può costruire il processo di analisi: l'approccio *machine learning* e l'approccio *lexicon-based*, cioè quello basato su dizionari [96].

Nel primo approccio, l'analisi del *sentiment* viene trattata come un problema di classificazione del testo e quindi risolta con tecniche di *machine learning*. Indipendentemente dall'algoritmo impiegato per la classificazione, tutte le tecniche includono il *training* di un modello su un campione di dati, ovvero di testi cui è già stato assegnato un giudizio di valore (*sentiment* positivo o

sentiment negativo), e quindi l'impiego del modello su dati nuovi. Nel secondo approccio invece, la definizione del *sentiment* è basata sull'analisi delle singole parole o frasi sfruttando dizionari di parole "emotive" (e.g. buono, cattivo, fantastico, terribile..) cui è assegnato un peso in termini di positività o negatività; le stesse parole vengono cercate nei testi di cui si vuole indagare il *sentiment* e il risultato dipende dalla frequenza di quelle parole nel testo [96]. A differenza del *machine learning approach*, dunque, il *lexicon-based* richiede, nella misurazione del sentimento, un dizionario di riferimento sul quale il testo di cui si vuole estrarre il *sentiment* viene confrontato. Nella pratica, un dizionario potrebbe contenere la parola "ottimo", cui viene assegnato il punteggio +3, e la parola "cattivo", cui viene assegnato il punteggio -1. Nella più semplice implementazione di un modello basato sull'approccio *lexicon-based*, quindi, tutte le parole del documento da analizzare sono confrontate con le parole già "pesate" del dizionario: ogni volta che una parola del documento da analizzare corrisponde ad una presente nel dizionario, il punteggio associato a quella parola viene sommato al punteggio complessivo del *sentiment* del documento. Il risultato finale sarà nient'altro che una somma di tutti i punteggi [96].

Entrambi gli approcci presentano vantaggi e svantaggi: il *lexicon-based* non richiede una precedente operazione di etichettatura dei testi in entrata, ma si basa su dizionari già esistenti. Questo può però rappresentare un duplice limite per l'indagine: da un lato la non scontata disponibilità di dizionari sufficientemente complessi nella lingua del testo su cui l'indagine è svolta, dall'altro la difficoltà di tenere conto del contesto dell'indagine e quindi dell'adeguatezza del dizionario che viene utilizzato, dal momento che lo stesso dizionario potrebbe non avere la stessa efficacia su contesti d'indagine differenti. Ancora, le collezioni di documenti con cui viene confrontato il testo

nel caso dell'approccio basato su dizionario sono quasi esclusivamente composte da testo scritto correttamente e senza errori di battitura. Nell'ambito di una analisi effettuata su dati raccolti da Internet, i testi sono perlopiù sporchi, spesso sgrammaticati e intervallati da un utilizzo "improprio" della punteggiatura che peraltro, in alcuni contesti, è a sua volta portatrice di significato (si pensi al caso delle *emoticon*). Dall'altro lato, l'approccio *machine learning* non richiede solitamente il ricorso ad un dizionario e sfrutta l'accuratezza dei metodi di classificazione utilizzati anche in altri strumenti di indagine. Tuttavia, questa accuratezza dipende fortemente da una corretta etichettatura dei testi utilizzati per il *training* e da una attenta selezione delle *feature* considerate dall'algoritmo [13]. Affronteremo meglio nel corso dell'indagine questi aspetti.

Vale la pena sottolineare che i due approcci non sono necessariamente alternativi: in letteratura sono state proposte anche tecniche ibride che combinano il *lexicon-based* con il *machine learning* [69]. Nell'ambito di questa ricerca, le tecniche di *sentiment analysis* che verranno descritte ed utilizzate non includono quelle ibride, né quelle fondate sull'approccio basato su un dizionario. Nel Capitolo 3, entrerò nel dettaglio dell'approccio *machine learning*, fornendo una panoramica degli strumenti a disposizione per effettuare l'analisi; seguirà quindi l'applicazione di un modello su un dataset acquisito da una piattaforma di *social networking* online.

2.2.2 Apprendere automaticamente: il *machine learning*

Il *machine learning* è quell'insieme di tecniche che permettono ad una macchina di apprendere e di perfezionarsi in una determinata abilità. La macchina

è intesa come un computer o un software che utilizza gli algoritmi ai quali si fa apprendere quella abilità: per usare la nota definizione di Andrew Ng [7], *"Machine learning is the science of getting computers to act without being explicitly programmed"*. L'apprendimento delle macchine è anche chiamato "apprendimento automatico" in quanto gli algoritmi diventano capaci di eseguire il compito automaticamente e indipendentemente dalle istruzioni del ricercatore umano, apprendendo invece dai dati stessi. Applicazioni comuni di *machine learning* che sono parte del nostro quotidiano sono ad esempio i sistemi di traduzione automatica, che traducono un testo da una lingua all'altra in modo automatizzato; i sistemi di raccomandazione tipici di piattaforme come Netflix o Amazon, che permettono di consigliare all'utente elementi di probabile interesse basandosi sullo storico di quelli precedentemente apprezzati; o ancora i filtri *antispam*, che identificano i messaggi pubblicitari basandosi sulla compresenza di alcune caratteristiche riconosciute precedentemente come posta indesiderata.

In questa sede ci si limita a fornire una definizione di *machine learning* sufficiente a comprendere il resto dei passaggi di metodo che riguardano l'applicazione di una tecnica in ambito di analisi del *sentiment*. Prima però di addentrarsi nell'applicazione di tecniche di *machine learning* per cogliere l'espressione soggettiva di un documento, potrebbe essere opportuno tenere a mente una considerazione legata a questo tipo di approccio, ovvero il fatto che la macchina non possieda una conoscenza concettuale degli elementi che sta analizzando. Il *machine learning* è un sistema che apprende dai dati, basandosi su una analisi statistica di un campione di elementi "storici" ed ottenendo successivamente una approssimazione sui dati nuovi, senza però comprendere la natura del dato che sta esaminando [32]. Torna quindi la considerazione già accennata più sopra, in merito al fatto che i metodi auto-

matici non possono – ancora - mirare al livello di correttezza di comprensione di un testo cui può invece ambire l'umano: essi, citando di nuovo Iacus et al (2014), possono solo "velocizzare alcune operazioni di analisi testuale e permettere l'analisi su larga scala di milioni di testi [...] Possono essere considerati uno strumento che aumenta le capacità umane (come un telescopio o una leva) ma non certo lo strumento che sostituisce l'uomo" [20].

2.2.3 La classificazione supervisionata

La classificazione è lo stadio della *sentiment analysis* in cui entrano in gioco le tecniche di *machine learning*. Un algoritmo di *machine learning*, infatti, esamina elementi che sono già stati classificati e costruisce un modello statistico che permette di classificare i nuovi elementi, basandosi sulle caratteristiche (*feature*) possedute. Si può pensare alle caratteristiche come a delle variabili esplicative e alla classe posseduta dagli elementi come a etichette o variabili target: i modelli di classificazione si propongono quindi di individuare i legami che intercorrono tra le *feature* (o variabili explicative) corrispondenti ad un elemento e l'appartenenza o meno di quell'elemento ad una classe. Questi legami vengono quindi tradotti come regole di classificazione, che vengono impiegate per assegnare la classe agli elementi considerati. Nell'impiego del *machine learning*, gli approcci che si possono utilizzare sono riassumibili in classificazione non supervisionata e classificazione supervisionata. Per quanto riguarda la classificazione dei testi sulla base del *sentiment*, la non supervisionata è più frequentemente utilizzata nell'approccio *lexicon-based* [74] mentre la classificazione supervisionata è quella più ricorrente nell'uso dell'approccio *machine learning*. In questa sede ci occuperemo quindi soltanto della classificazione supervisionata.

I metodi di classificazione supervisionata sono quelli che prevedono un

maggiore controllo da parte del ricercatore, in quanto richiedono che siano note, prima di procedere nell'analisi, le classi "finali" cui vengono assegnati gli elementi. Nello specifico, per la *sentiment analysis* si utilizeranno quindi testi che sono stati precedentemente etichettati, ovvero cui sia stato assegnato un valore relativo al *sentiment* (positivo o negativo, seguendo una classificazione base). Di questa fase vedremo in seguito una descrizione di dettaglio e una applicazione diretta. Come in tutte le tecniche di *machine learning*, il processo si può scomporre sostanzialmente in 3 fasi:

a. *Fase di training*

   Un set di elementi testuali, sottoinsieme dell'intera popolazione (ovvero del totale dei testi da analizzare), viene utilizzato per insegnare al modello di classificazione a dedurre le regole che consentono di attribuire ad ogni elemento la corretta "etichetta".

b. *Fase di test*

   Le regole prodotte nella fase precedente vengono impiegate per classificare gli elementi che non sono parte del sottoinsieme precedente, ovvero quelli non ancora classificati.

c. *Implementazione*

   Il modello viene utilizzato per classificare i nuovi elementi, ovvero i contenuti testuali futuri che il ricercatore vorrà analizzare. Il modello utilizzerà quindi le regole generate in fase di *training* e "validate" dalla fase di test.

Nello specifico ambito della *sentiment analysis*, abbiamo detto come la classificazione riguardi l'assegnare ad un testo l'etichetta (classe) che ne definisce il sentimento (positivo o negativo, o un range compreso tra questi due poli). L'obiettivo delle tecniche di *machine learning* è quello di produrre

un modello capace di apporre questa etichetta in modo automatico, senza l'intervento del ricercatore [39]: il modello così generato si può considerare predittivo, nel senso che è capace di predire correttamente il *sentiment* del testo (la sua classe di appartenenza) a partire da una serie di caratteristiche possedute dal testo (le sue *feature*, o variabili esplicative). La classificazione è detta supervisionata in quanto le possibili classi sono decise a priori dal ricercatore, il quale governa l'indagine definendo in partenza quali sono le dimensioni su cui verterà l'analisi. Usare un approccio supervisionato richiede quindi che, nella fase di *training* (a), il modello si alleni su un set di dati precedentemente classificati, ovvero cui sia già stata assegnata l'etichetta del *sentiment*. Nella fase di utilizzo del modello su un caso reale, vedremo le tecniche propedeutiche al *training* del modello: queste includono anche la preparazione del dataset e quindi l'applicazione delle etichette al *training set* su cui il modello apprenderà e genererà le regole.

2.2.4 Twitter e le API

La scelta di posizionare la descrizione di Twitter e delle dinamiche di raccolta dei dati solo dopo aver anticipato le tecniche di classificazione potrebbe sembrare controintuitiva, dal momento che la fase di raccolta dei dati è temporalmente precedente a quella di implementazione del modello. La decisione è in realtà dovuta all'arbitrarietà della scelta che si opera nel momento in cui si predilige Twitter come campo di raccolta e analisi dei dati. Quando parliamo di *sentiment analysis* online, possiamo fare riferimento più genericamente a tutti i contesti in cui gli utenti di Internet esprimono la loro opinione. Questo include blog, forum, chatroom, microblogging e, ovviamente, piattaforme di *social networking* tra cui Facebook, Twitter, Quora, Reddit. Non solo: quando parliamo di *sentiment analysis*, ricordiamo che questo strumento è

solo uno delle possibili applicazioni della *social media mining*, intesa come l'analisi sistematica delle informazioni generate dai social media. Le stesse tecniche che vedremo applicate su un caso reale utilizzando dati raccolti da Twitter possono essere impiegate anche in altri contesti di monitoraggio del *sentiment* online. La scelta di concentrare l'indagine su Twitter segue però le stesse motivazioni che vedono ancora questa piattaforma, nella quantità di lavori di ricerca reperbili ad oggi, come quella privilegiata su cui effettuare analisi di *social media mining*. Riassumendo quanto ben riportato da Heimann e Dainemann [48], le ragioni per scegliere Twitter sono sostanzialmente riconducibili a) alla tipologia di trasmissione delle informazioni che caratterizza la struttura della rete di questo social network, b) all'utilizzo che la comunità online fa di Twitter rispetto ad altri social network.

Il primo punto afferisce alla capacità di Twitter di favorire le connessioni di secondo ordine, o *"weak ties"* [43]. Per come è concepita la rete di Twitter, i legami non sono biunivoci e paritari: questo comporta che si possa essere *follower*, o seguace, di un utente senza essere a propria volta seguiti dallo stesso. Secondo questo approccio, è possibile per qualsiasi utente creare un legame (benché debole) con il Presidente degli Stati Uniti, mentre ben più difficile sarebbe diventargli amico su Facebook. Per lo stesso principio, che coinvolge anche il secondo punto, Twitter è anche in grado di veicolare un maggiore scambio di informazioni tra utenti che condividono contenuti frequentemente con utenti meno attivi: l'uso della piattaforma è per molti finalizzata all'ascoltare, più che al condividere, quindi al rimanere esposti e aggiornati sulle attività, le notizie e le idee diffuse da altri. Il secondo punto afferisce proprio all'uso di Twitter per raccogliere informazioni e commenti o per contattare le aziende per richieste o reclami sui loro prodotti e servizi. Nel Capitolo 3 di questa tesi vedremo più da vicino la natura dei dati raccolti

da Twitter e ci soffermeremo sulla descrizione della tipologia di informazioni allo stadio zero, ovvero prima della fase di preparazione del dataset su cui effettueremo l'analisi. Nel paragrafo successivo, verrà invece fornita una prima descrizione del linguaggio di programmazione e del software che è stato scelto per applicare la *sentiment analysis* sui dati raccolti.

2.2.5 R

La scelta di R per i fini di questa tesi è secondaria rispetto alle considerazioni teoriche e metodologiche sull'analisi del *sentiment*. Facendo però riferimento di nuovo al lavoro di Heimann e Danemman [48], l'utilizzo di R come ambiente su cui condurre un'analisi di *social media mining* comporta diversi vantaggi. Oltre a svolgere la funzione di semplice calcolatore, la flessibilità di R consente di assistere l'utente nella manipolazione anche di grandi dataset, sia per il calcolo di funzioni base sia per l'applicazione di algoritmi e operazioni matematiche complesse, nonché elaborazioni statistiche e produzione di grafici. Possiede numerose funzioni dedicate al calcolo matematico avanzato e all'analisi statistica e consente di realizzare nuove funzioni facilmente richiamabili all'occorrenza dall'utente. R [79] consente inoltre l'accesso a librerie di funzioni e programmi utilizzabili dagli utenti. Insieme al software si possono scaricare dal sito principale e dai numerosi siti collegati sia i manuali d'uso sia i pacchetti aggiuntivi. Nuove librerie vengono continuamente create dagli sviluppatori e rese accessibili in modo immediato all'utente. In particolare, molte collezioni per il *text mining* sono messe a disposizione rendendo ampio il ventaglio degli strumenti a vantaggio del ricercatore che affronti un'analisi come quella oggetto di questa tesi. Il grande numero di lavori di ricerca e di manuali di *data mining* che utilizzano R come standard rende chiaramente più semplice per chiunque si approcci a questo tipo di analisi venire guidato

nella applicazione di un modello su un caso reale. R è inoltre liberamente accessibile sotto la GNU General Public License e, pur semplicemente utilizzabile con l'interfaccia *command-line* tipica, in questo lavoro di ricerca adotteremo il più comune ambiente disponibile per interfacciarsi al software: RStudio.

Capitolo 3

LA SENTIMENT ANALYSIS SUL CASO STARBUCKS

3.1 Il caso Starbucks e #PalmeMilano

3.1.1 Profilo del brand Starbucks

Starbucks è una multinazionale statunitense di caffetterie. Fondata il 30 marzo 1971 a Seattle, è con il suo amministratore delegato Howard Shultz che ha trovato il successo, diventato oggi mondiale. Con un fatturato di 15,6 miliardi [52] e una presenza diffusa in oltre 70 paesi [52], il brand Starbucks è oggi uno dei più famosi nel settore della ristorazione.
Starbucks Corporation è più precisamente un rivenditore di caffè, prima che una catena di caffetterie. La compagnia compra e quindi tosta il caffè che poi vende nei suoi bar, assieme a te, bevande e assortimenti di pasticceria e di snack salati.

Il target e il posizionamento Il segmento di consumatori principale per Starbucks è costituito da adulti tra i 25 e i 40 anni di età con un reddito

medio-elevato; seguono i giovani tra i 18 e i 24 anni che appartengono a famiglie ricche [46]. In generale è dai consumatori della Generazione Y (nati tra il 1977 e il 2000) che deriva il maggior profitto per Starbucks [37]. I clienti tendono ad essere quindi adolescenti e giovani adulti che vivono in città e possiedono un reddito, personale o di famiglia, relativamente alto. Starbucks ha incentrato la sua *value proposition* sulla visione del caffè come culto, dal complesso e raffinato valore esperienziale, per il quale i clienti sono disposti a spendere di più. L'azienda è stata infatti capace di affermarsi su scala mondiale presentandosi come "primo fornitore del migliore caffè del mondo"[37] curando in particolare il servizio al consumatore, più personalizzato rispetto a quello delle catene fast-food tradizionali, e il design dei locali in cui l'esperienza del caffè viene consumata e che è parte integrante dell'immagine del brand [18]. La qualità del servizio offerto e dell'ambiente degli *store* Starbucks, unite alla proposta di un assortimento di caffè, bevande e prodotti snack distintivi, costituiscono la *unique selling proposition* che posiziona il brand nella fascia medio-alta del settore. Il vantaggio competitivo ottenuto da Starbucks soprattutto per la capacità di trasformare la consumazione del caffè in un momento ad alto valore esperienzale si è tramutata, per certi versi e in alcuni Paesi più che in altri [95] nella rappresentazione del caffè di Starbucks come *status symbol*, anche in ragione del prezzo a cui tutti i prodotti del suo assortimento vengono venduti.

La comunicazione Parte della strategia di comunicazione di Starbucks si basa sulla *customer loyalty*, intesa come la capacità di creare una relazione con il consumatore continuativa e duratura - una logica, questa, che risiede nel concetto di *LifeTime Value* del cliente [11]. L'esperienza che viene creata nel negozio dovrebbe essere cercata nuovamente dal consumatore, che desi-

dera replicarla con frequenza. In quest'ottica si inserisce anche la costruzione di una vera e propria *community* intorno al brand Starbucks, i cui membri condividono l'amore per il caffè e l'essere giovani e benestanti. Sul concetto di *community*, le strategie di comunicazione del marchio hanno fatto leva specialmente negli ultimi anni sfruttando in particolare le piattaforme di *social networking* e la presenza in più canali digitali [95].

Le strategie di crescita Oltre al rafforzamento del *loyalty program* di Starbucks, le strategie di crescita includono la diversificazione dei prodotti offerti, l'aumento delle occasioni di consumo e l'espansione in nuovi Paesi [62]. Proprio a partire da quest'ultimo punto si può iniziare a considerare l'analisi del caso di studio che viene presentato in questo lavoro di tesi, ovvero l'apertura del primo *store* Starbucks in Italia e la campagna di marketing associata a questo evento.

3.1.2 Starbucks in Italia

L'apertura di Starbucks in Italia non è un'apertura qualunque. Le motivazioni di questa unicità sono diverse e afferiscono non solo al prodotto venduto da Starbucks e al valore storicamente peculiare che gli è attributo in Italia (il caffè), ma anche alla storia della compagnia stessa, che nasce – secondo le narrazioni del suo fondatore e attuale CEO, Howard Schultz – proprio a Milano. Durante un viaggio in Italia nei primi anni '80, Schultz avrebbe per la prima volta avuto esperienza del caffè come momento di pausa e di socialità nella città di Milano: questo avrebbe poi ispirato il suo progetto imprenditoriale che lo ha portato ad acquisire la già esistente compagnia Starbucks trasformandola nel brand che conosciamo oggi [8].

Nonostante l'espansione a livello mondiale, l'apertura di Starbucks in Italia non è ancora avvenuta. I legami commerciali sono stati fino ad ora limitati alla partnership con l'HMSHost, divisione americana del gruppo italiano Autogrill, concessionario del marchio per la rete commerciale su strada nel Nord America, che è in piedi dal 1991[88]. La stessa azienda ha sempre affermato di non avere progetti di espansione nel nostro Paese, motivando questa decisione sia a) con la diffusione capillare di bar in cui è possibile consumare prodotti di caffetteria a prezzi decisamente ridotti rispetto a quelli offerti da Starbucks, sia b) con l'ostacolo culturale rappresentato dalla tradizione del caffè come simbolo italiano, difficilmente paragonabile a quello offerto dal potenziale *competitor* statunitense [89].

Il rapporto dell'Italia con Starbucks è quindi peculiare rispetto a quello di altri Paesi con la multinazionale del caffè. L'esperienza che di Starbucks hanno potuto fare gli italiani fino ad oggi è stata limitata ai soli viaggi nei Paesi esteri in cui la compagnia è già presente. L'arrivo su suolo nazionale italiano è argomento di cui il brand ha iniziato a parlare solo in tempi relativamente recenti. A febbraio 2016, Schultz anticipò questa decisione in un post sul blog aziendale, dichiarando che si sarebbe avvicinato al mercato italiano "con umiltà e rispetto" e rendendo così più credibile parlare di una apertura imminente. La dichiarazione di Schultz è stata immediatamente seguita dal dibattito sulle conseguenze della presenza della multinazionale del caffè in quella che è considerata la "patria" dell'espresso; lo stesso dibattito ha diviso l'opinione pubblica in merito all'accoglienza che la popolazione italiana avrebbe riservato al marchio [89].

Le più recenti informazioni sull'apertura risalgono a martedì 28 febbraio, quando Starbucks ha presentato ufficialmente il progetto per aprire il suo primo negozio in Italia, in piazza Cordusio a Milano. Il negozio sarà a Palazzo

Broggi, già sede della Borsa di Milano e delle Poste Italiane. Non si tratterà di un normale bar di Starbucks, ha anticipato il CEO della compagnia, ma di una *roastery*, una torrefazione particolarmente grande e arredata "in modo più elegante del solito (...) il cliente vedrà tubi che attraverseranno i soffitti nei quali passano i grani. Potrà comprare le miscele e i nostri prodotti legati al marchio. Poi ci sarà la tecnologia: Wi-Fi super veloce, musica con i partner di Spotify, servizi di pagamento *fintech* (...) ci saranno cinque nuovi caffè realizzati con tecnologie ideate da noi, oltre al tradizionale espresso". Starbucks collaborerà con la catena di panetterie di Milano "Princi" rivendendo nella torrefazione i prodotti da forno. A collaborare sempre con la multinazionale è anche il gruppo imprenditoriale italiano Percassi, che aprirà alcuni altri Starbucks a Milano nel 2018. A questo proposito, Schultz ha dichiarato che "ciascuno sarà progettato accuratamente e curato per rispettare la comunità locale e l'unicità del contesto milanese. Adottare un approccio rispettoso e misurato nell'apertura dei negozi è al centro della strategia di Starbucks a Milano".

La notizia dell'apertura di Starbucks è però stata preceduta da due settimane di scalpore attorno al brand, grazie all'operazione di comunicazione che ha anticipato l'annuncio e che si lega in particolar modo alla città di Milano. Proprio questa operazione commerciale offre il punto di partenza per l'indagine presentata in questo lavoro di ricerca.

3.1.3 Le palme in Piazza del Duomo e la campagna di Starbucks

Il 15 febbraio in Piazza del Duomo a Milano è iniziata ufficialmente la campagna di comunicazione che culminerà con l'apertura del primo Starbucks in

Italia. L'esordio è avvenuto con la sistemazione del boschetto nella piazza principale del capoluogo lombardo, sponsorizzata da Starbucks con la piantumazione di palme, banani e begonie. Starbucks ha infatti vinto il bando di sponsorizzazione indetto da Palazzo Marino [24] ed ha quindi iniziato a realizzare le nuove aiuole che, secondo il progetto, dovrebbero rimanere in Piazza del Duomo per almeno tre anni.
La notizia è stata in particolare diffusa da un post su Instagram da parte del sindaco di Milano, Giuseppe Sala:

> *"Milano si risveglia con palme e banani in piazza Duomo. Come nella tradizione ottocentesca. Buona o cattiva idea? Certo che Milano osa eh..."* [24].

Lo stesso Sala ha poi aggiunto:

> *"Da cittadino sospendo anch'io il giudizio: vediamo quando sarà finito il lavoro. Tendenzialmente non mi dispiace, però, voglio vedere bene, quando tutto sarà finito... Il riferimento storico c'è. Richiama l'Ottocento e la Sovrintendenza è stata positiva."* [24]

L'affermazione del sindaco di Milano, con il riferimento all'Ottocento che la presenza delle palme in Piazza del Duomo potrebbe richiamare, si spiega alla luce della polemica divampata proprio attorno a questa iniziativa. Nei giorni successivi all'inizio della campagna, gli utenti dei social network hanno iniziato a fotografare e commentare la presenza delle palme nella Piazza del Duomo. L'*hashtag* #palme è in breve diventato *Trending Topic* su Twitter. La diffusione della notizia è stata rafforzata anche dalle prese di posizione di più parti politiche in relazione all'evento. Molti politici del centrodestra

hanno criticato la scelta, associandola ad una rinuncia ai valori tradizionali della città e ad una "africanizzazione" dell'Italia, vedendo le palme e i banani come concessione culturale alla provenienza degli immigrati presenti a Milano.

> *"#Palme e banani in piazza Duomo? Follia. Mancano sabbia e cammelli, e i clandestini si sentiranno a casa. #motosega #starbucksgohome"* [67]

> *"Milano si sta trasformando in una piccola Africa, aprendo le porte a immigrati e clandestini, e quindi vuole anche mettere palme e banani in piazza del Duomo. Non si è mai vista una grande cattedrale europea con piante di banane di fronte"* [67]

Oltre alla polemica di indirizzo politico, la discussione sull'arrivo delle palme in Piazza del Duomo ha incluso commenti di vario carattere: da quelli relativi al gusto estetico della scelta a quelli di carattere ambientalista; associazioni sull'esoticità delle palme sono state fatte richiamando anche la California o i paesi equatoriali e, più in generale, il sole e il mare. Altri commenti sono seguiti alla stessa polemica politica, che ha generato un ulteriore ramo di discussione spostando l'attenzione non tanto sulla iniziativa di Starbucks, quanto sui contenuti della polemica stessa. La discussione si è ulteriormente accesa in seguito alle dimostrazioni in piazza compiute da LegaNord e CasaPound; la protesta è culminata nell'atto di vandalismo della notte del 18 febbraio, durante il quale ignoti hanno dato fuoco ad alcune delle palme piantate nelle nuove aiuole [28].
La discussione sulla campagna di Starbucks è diventata particolarmente intensa anche in seguito alla presentazione del progetto di apertura della torrefazione, quando le intenzioni della multinazionale si sono fatte più chia-

re. Lo stesso CEO ha commentato quanto accaduto nei giorni precedenti all'annuncio:

> *"Pensavamo di offrire qualcosa di bello alla città. Ma ogni mercato può presentare temi diversi. In questo caso Starbucks è finita dentro un problema di tipo politico. Mi dicono però che i milanesi all'inizio criticano ma poi si affezionano (...) Era un'idea bella realizzata da un noto architetto, Marco Bay. Ci era piaciuto molto. Starbucks è lo sponsor e ha investito circa 200mila euro. Comunque, abbiamo grande rispetto per il Paese del caffè dove ho imparato molto."* [78]

3.2 L'applicazione del *sentiment* sulle palme

3.2.1 Orientamento dell'indagine

Il contesto descritto nei paragrafi precedenti è stato scelto ai fini dell'applicazione della *sentiment analysis* su un caso reale per diversi motivi. In primo luogo, la peculiarità di Starbucks in relazione al contesto Italia e la considerazione contrastante che l'opinione del consumatore italiano può avere nei confronti del brand, di cui si è accennato più sopra, rendono l'analisi particolarmente interessante dal punto di vista di marketing e di comunicazione d'impresa. Vedremo in particolare come l'analisi costringerà il ricercatore a considerare non solo il brand Starbucks, ma anche il brand Milano, depositario di un'immagine e di un'identità che vengono necessariamente toccati e condizionati dalla campagna messa in atto dalla multinazionale statunitese. Dall'altro lato, poi, la ricchezza del dibattito online divampato in relazione a questa campagna, caratterizzato da implicazioni di tipo politico e culturale e

non squisitamente circoscritto al brand Starbucks, ne fa un argomento ancor più interessante se interpretato come sfida che i piani diversi su cui il dibattito si muove pongono per la *sentiment analysis* e ancor più, come vedremo, per la scelta delle dimensioni dell'indagine – scelta che ricade, nell'ottica di una classificazione puramente supervisionata, sul ricercatore *in primis*. Infine, l'attualità degli eventi di cui si sta scrivendo rende l'indagine sostanzialmente aperta: considerando che il lancio del primo Starbucks non avverrà prima della fine del 2017 e che la posa delle palme è, presumibilmente, solo la prima di una serie di iniziative di marketing del marchio in Italia, l'analisi potrebbe prestarsi a ulteriori sviluppi ed evoluzioni, incorporando una più estesa ricerca sul brand Starbucks e sul *sentiment* degli italiani prima e dopo l'apertura.

3.2.2 Domanda e metodo di ricerca

Alla luce delle considerazioni di cui sopra, la posa delle palme in Duomo coinvolge almeno tre livelli di discussione, o di *sentiment*:

- Il *sentiment* degli italiani rispetto alla iniziativa di posa delle palme
- Il *sentiment* degli italiani rispetto al brand Starbucks
- Il *sentiment* degli italiani rispetto alla polemica seguente alla posa, alle proteste politiche ed agli atti di vandalismo contro le piante

Rispetto a questi tre livelli di discussione, la ricerca è partita volutamente da una domanda che indirizzasse il primo *sentiment*, in quanto oggetto principale della ricerca di marketing:

Come viene percepita dagli italiani l'iniziativa di posa delle palme?

A questo punto è utile definire meglio i confini in cui la ricerca di *sentiment analysis* online si muove. Sarà interessante vedere come la domanda di cui sopra verrà "ridimensionata" per rendere conto proprio di quei confini (e inevitabilmente dei limiti) descritti di seguito.

In primo luogo, quando parliamo degli italiani ci stiamo di fatto riferendo *non solo* ai soli italiani che fanno parte di quel 52% che accede mensilmente ad Internet, bensì ai soli italiani di quel 52% che accede mensilmente a Internet e che possiede un account Twitter. La sottolineatura è importante non soltanto per dare un'idea della significatività del campione considerato (sono 6.4 milioni [68] gli utenti attualmente registrati su Twitter, un numero che si riduce ulteriormente se si considerano i soli utenti attivi), ma anche in relazione all'esclusione, in sede d'indagine, degli altri spazi online in cui la discussione si è evoluta, non meno importanti di Twitter: Facebook ed Instagram *in primis*, seguiti dai blog e dai siti di news che hanno divulgato le notizie relative alla campagna.

In secondo luogo, la fase di raccolta dei dati e quindi di costruzione del dataset su cui svolgere l'indagine è stata effettuata sulla base di parole chiave relative al fenomeno, con cui è stata scandagliata la piattaforma di *social networking* considerata. La fase di raccolta dei dati, preliminare a qualsiasi tipo di indagine, e il modo in cui questa viene decisa e gestita, non è da sottovalutare rispetto alla sua significatività nel determinare l'esito dell'indagine stessa. È chiaro che una collezione di dati non esaustiva o limitata rischia di tagliare fuori aspetti significativi del fenomeno considerato. Dall'altra parte, il pericolo di allargare troppo il raggio d'indagine, includendo nel dataset una numerosità di osservazioni non pertinenti, è senza dubbio da tenere in considerazione. Nello specifico di questa ricerca, si è cercato di adottare un approccio il più inclusivo possibile rispetto alle parole chiave ritenute rilevan-

ti per l'indagine e di sfruttare invece i limiti dettati dall'orizzonte temporale considerato (01 gennaio 2017 – 28 febbraio 2017) per eliminare le osservazioni chiaramente *offtopic*. È infatti facilmente intuibile che l'*hashtag* #palme usato su Twitter da account di utenti italiani ad Agosto 2016 avrebbe incluso nel dataset una numerosità di *tweet* non pertinenti con l'indagine in oggetto.

Un altro "limite naturale" che si è sfruttato nella composizione del dataset è quello dovuto alla lingua italiana. Il carattere globale del brand Starbucks, la sua predisposizione all'utilizzo delle piattaforme social per dialogare con le *community* di consumatori e infine l'assenza di un suo profilo Twitter esclusivamente italiano avrebbero reso difficile al collettore discernere tra i *tweet* destinati alla campagna Starbucks in Italia e quelli invece rivolti dalla universalità del popolo web al marchio ogni giorno. L'impostazione della ricerca sui soli *tweet* scritti in lingua italiana, quindi, ha automaticamente risolto buona parte del rischio di inclusione di contenuti non pertinenti, eccezione fatta per quelli contenenti "espressioni universali" (si pensi a "LOL", "Frappuccino <3", e così via) e quelli costituiti dal solo *hashtag* corredato da un'immagine (si pensi a chi usa postare la sola foto del proprio caffè con *hasthtag* #starbucks).

Un ulteriore *caveat* è quello relativo appunto alle parole chiave utilizzate: queste sono state scelte effettuando una preliminare indagine sul social network per individuare le parole utilizzate in modo ricorrente nella discussione sul fenomeno delle palme a Milano. Oltre agli *hashtag* comuni (#palme e #palmeMilano soprattutto, assieme ai loro corrispettivi senza *hashtag*), così come a quelli meno comuni (#tropical, #duomodimilano), è doveroso sottolineare che non tutti i *tweet* legati in qualche modo al fenomeno indagato sono rintracciabili nella fase di raccolta dei dati. Si pensi ad esempio ai commenti che gli utenti esprimono nelle conversazioni online sottintendendo un

riferimento senza che questo venga espressamente nominato: una foto della piazza di Milano con le palme sullo sfondo, scrivendo *"Che bellezza"*, non contiene alcun elemento che permetta al collettore di risalire al *tweet* e includerlo nel dataset – tuttavia questo rappresenta un'espressione di *sentiment* assolutamente pertinente per l'indagine.

Si fornisce quindi di seguito la lista delle *keyword* e degli *hashtag* che sono stati inclusi nella fase di raccolta:

#palme	**banani**	**#piazzaDuomo**
palme	**milano**	**@beppesala**
#palms	**#Milano**	**starbucks**
#palmeMilano	**#duomomilano**	**#starbucks**
#milanpalms	**#duomodimilano**	**#starbucksgohome**
#tropical	**#duomo**	
#banani	**duomo**	

Nel capitolo successivo vedremo una più puntuale descrizione del dataset e quindi della fase di preparazione del dato, propedeutica all'analisi di *sentiment* vera e propria che sarà effettuata, come già anticipato, con RStudio.

3.2.3 Descrizione del dataset

Il dataset su cui è stata effettuata l'analisi comprende 24.715 osservazioni[1] relative al periodo 1 gennaio 2017 – 28 febbraio 2017, con una diversificazio-

[1]Si ringrazia *Voices From The Blog* per la raccolta dei dati, che è stata effettuata utilizzando loro provider.

ne della ricerca spiegata di seguito. Il fenomeno della polemica delle palme è concentrato solo nel periodo 15 febbraio – 28 febbraio, ovvero da quando la campagna di apertura di Starbucks in Italia è partita. Solo in quelle due settimane, quindi, la ricerca ha incluso anche le parole chiave e gli *hashtag* relativi alle palme. Dall'1 gennaio al 28 febbraio invece, la ricerca è stata focalizzata solo sulle chiavi relative a Starbucks: questo ai fini di verificare il *sentiment* degli italiani su Twitter verso la multinazionale prima che la vicenda delle palme lo influenzasse. In tutto i *tweet* raccolti nei due mesi su Starbucks soltanto è di 10.485; ammontano invece a 14.230 quelli che riguardano la vicenda delle palme. In sostanza quindi la fase di data collection ha portato ad ottenere due dataset diversi: 1) Dataset Starbucks (1 gennaio – 28 febbraio). 2) Dataset Palme (15 febbraio – 28 febbraio). La informazioni che si possono ricavare grazie alle API di Twitter non si limitano al solo testo dei *tweet*. La funzione qui considerata è quello dello *Streaming API*, utilizzata dal collettore messo a disposizione da Voices Analytics per questa ricerca. Le Streaming API forniscono tre livelli di recupero dei *tweet*: *POST statuses / filter, GET statuses / sample e GET / firehose.*
La prima in particolare è in grado di restituire un insieme di *tweet* filtrati per determinati parametri:

- ***follow***: fornisce la possibilità di indicare di quali utenti recuperare i relativi messaggi. I messaggi recuperati saranno quelli creati dall'utente, quelli ricondivisi dall'utente, quelli che sono il *retweet* di un suo messaggio e quelli in risposta ad un *tweet* dell'utente

- ***track***: 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

- ***delimited***: 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.

I testi vengono recuperati sotto forma di file *JSON* (JavaScript Object Notation), contenente diverse informazioni, o *meta-tag*, che vanno dall'id dell'utente che ha generato il *tweet* alla lingua al geotag:

- ***id***: valore che identifica in modo univoco il *tweet*
- ***text***: testo completo del messaggio che, com'è noto, non può superare i 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
- ***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"
- ***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.

Grazie a questa panoramica dei dati contenuti in un singolo *tweet*, è facile immaginare altri possibili sviluppi della ricerca del *sentiment*: la localizzazione geografica dei *tweet* o una clusterizzazione degli stessi sulla base del sesso o dell'età anagrafica, risalendo con l'id alle informazioni sull'utente che lo stesso abbia reso pubbliche. Ai fini di questa analisi, tuttavia, l'indagine è stata limitata ad uno solo dei campi di informazione ottenuti in fase di raccolta dei dati: quello testuale, ovvero il messaggio invato da ogni utente che abbia utilizzato una delle parole chiave rilevanti, per esempio:

≪Ma se in Perù e in Azerbaijan c'è lo Starbucks, perché in Italia ancora no? Sinceramente non lo capisco #starbucksitalia #semprepeggio≫

≪@niky_flower "quello di Starbucks!"≫

≪Teringin Starbucks.. teringin my Caramel Macchiato huhuhuhu≫

≪Caramel Macchiato yasssssss (@Starbucks) #Yelp #Yelfie `https://t.co/ss30lYgQfG` `https://t.co/akNpSQlvnF`≫

≪Ultimo frappucino da starbuck @StarbucksUK `https://t.co/wXtVplXAzv`≫

La fase finale di raccolta ha quindi ordinato ogni *tweet* nei file che costituiscono i due diversi dataset specificati sopra. Il file relativo alla campagna di Starbucks con le palme in Duomo sarà la base dell'analisi di seguito proposta.

3.3 Applicazione della classificazione supervisionata

3.3.1 *Pre-processing*

Nella decisione di applicare la *sentiment analysis* al caso "Palme", il primo ostacolo è determinato dalla lingua dei testi che vengono esaminati: trattan-

dosi infatti di un dataset di *tweet* in italiano, non è stato possibile sfruttare le numerose risorse facilmente reperibili in rete per le applicazioni di *sentiment analysis* in inglese. Pur non basandosi, a differenza della classificazione supervisionata basata sui dizionari di parole "etichettate", anche per la costruzione di modelli di *machine learning* ci si può avvalere di strumenti di supporto; questi entrano in gioco in tutte le fasi di preparazione del modello rendendo più veloce il lavoro del ricercatore. Vedremo come la poca disponibilità di risorse in lingua italiana si sia ripresentata come problematica in diversi momenti dell'analisi, il primo dei quali costituito dalla necessità di far apprendere i modelli su un insieme di *tweet* già classificati sulla base di sentimento positivo, neutro e negativo. La classificazione avviene manualmente ed è solitamente affidata a più di un ricercatore, per poi essere riesaminata in modo da far aumentare l'oggettività e risolvere in modo imparziale l'assegnazione dei testi alle classi per i casi controversi. La fase di classificazione manuale diventa quindi particolarmente onerosa in termini di tempo e di risorse da impiegarvi. Non disponendo di un dataset riguardante la tematica in oggetto che fosse già stato etichettato, si è utilizzato il *dataset Sentipolc* [90] messo a disposizione in rete dal Politecnico di Torino. Il dataset consiste di 2000 *tweet* in italiano per cui è stata realizzata una classificazione manuale su più parametri. Per lo scopo di questa tesi le uniche tre classi che sono state mantenute sono quelle standard della *sentiment analysis*, ovvero polarità positiva, polarità negativa e neutralità, quest'ultima intesa sia come mancanza di un sentimento prevalente, sia come assenza di soggettività.
È stato quindi ottenuto un file unico a due colonne: la prima ad identificare la classe, la seconda come testo del *tweet*. Le categorie delle classificazioni sono esemplificate di seguito:

CLASS	TEXT
0	Tra 5 minuti presentazione piano scuola del governo #Renzi. #passodopopasso #labuonascuola Stay tuned
1	@matteorenzi: Alle 10 appuntamento su `http://t.co/YphnXknDML` #italiariparte #labuonascuola" #Grandinsegnanti ... #Buonlavoro"
-1	#labuonascuola gli #evangelisti #digitali non devono essere già dentro la scuola, esempi

Prima di dare i testi in pasto all'algoritmo, è necessario sottoporre il testo ad una fase di *pre-processing* che consente di ridurlo a dato quantitativo. Si tratta di una fase di analisi prettamente lessicale, che ripulisce il testo sottoponendolo ad una normalizzazione e ad una semplificazione, da una parte rimuovendo quanto considerato inutile o non significativo, dall'altra standardizzandolo e uniformandolo. La gran parte degli strumenti e delle risorse realizzati finora per la *sentiment analysis*, come già anticipato, sono in lingua inglese: questo vale anche per gli algoritmi che scansionano il testo in fase di *pre-processing*. Mentre alcuni passaggi sono indipendenti dalla lingua in cui il testo è scritto, come ad esempio la trasformazione *lowercase* (al fine di considerare equivalente una stessa lettera scritta in maiuscolo o in minuscolo), altri interventi di trasformazione dipendono strettamente dal dizionario linguistico: è il caso delle *stop word*, cioè di quelle parole "prive di significato" che però permettono di costruire le frasi, come gli articoli o le preposizioni. Per la nostra analisi, è stato sfruttato uno dei pochi pacchetti in italiano, TextWiller[91], ed è stato integrato con il più complesso pacchetto Snowball di R [14].

Nonostante l'analizzatore lessicale di TextWiller standardizzi il processo includendo in un unico comando tutte le fasi necessarie a preparare i testi per l'analisi, l'approccio che è stato utilizzato nei confronti del *pre-processing* è stato quello di valutare ad ogni stadio di trasformazione il risultato raggiunto e di raffinarlo ulteriormente includendo ulteriori passaggi dove necessario.
Il primo stadio è quello di uniformazione del testo convertendolo a lettere minuscole. Il secondo è quello di rimozione degli URL, ovvero degli indirizzi web inclusi come link in numerosi *tweet*: dapprima individuandoli grazie al comando apposito di TextWiller, poi rimuovendoli con un comando del pacchetto tm [33]. Prima della rimozione della punteggiatura, sono state individuate le *emoticon* in quanto fondamentali per comprendere la polarità. L'analizzatore "normalizzaemote" è in grado di riconoscerle nel testo e tradurle in una parola standard, EMOTEGOOD per quelle positive, EMOTEBAD per quelle che esprimono tristezza o rabbia. Solo in seguito si può procedere con la rimozione di punti e altri simboli, esclusi gli *hashtag*: fondamentale a questo proposito sottolineare che l'eliminazione del simbolo # lascerebbe una parola "pura", che l'analizzatore considererebbe come parola equivalente alle altre. Per come è consuetudine usare gli *hashtag*, questo comporterebbe che in alcuni casi la parola privata di # sia in realtà una concatenazione di parole (si pensi ad esempio a #palmeMilano).
Dopo aver eliminato gli spazi bianchi in eccesso, si procede con la rimozione delle *stop word*. Oltre a quelle comuni come gli articoli e le preposizioni di cui si accennava prima, è importante notare che ulteriori altre *stop word* potrebbero essere incluse a seconda del contesto e della tipologia di testo. La creazione di *stop list* specifiche e quindi la loro rimozione unitaria in fase di *pre-processing* permetterebbe di dare in pasto all'algoritmo di classificazione un testo ridotto all'essenziale, facendo diminuire ulteriormente l'onere com-

putazionale. Un esempio di ulteriore *stop word* frequente che si è deciso di eliminare è "RT", tipica appunto dei testi provenienti da Twitter e indicante il fatto che il *tweet* includa un *"retweet"*, inteso come replica ad un messaggio precedente di un altro utente. Vediamo di seguito un esempio:

«#labuonascuol docent potra dimostr val @miursocial»

3.3.2 *Stemming*

A questo stadio i *tweet* sono stati trasformati in liste di parole. L'obiettivo ora è quello di ridurre ulteriormente la complessità, riducendo le parole al loro tema, inteso come la parte della parola che rimane dopo essere stata separata dalla desinenza. Il tema non necessariamente coincide con la radice morfologica, ma consente ugualmente di mappare le parole correlate considerandole equivalenti: "andare", "andai", "andò" saranno tutte trattate come tema "and", il quale non corrisponde alla radice di quelle parole.
La creazione di un algoritmo di *stemming* è stata propedeutica a rendere sempre più sofisticate le interrogazioni dei motori di ricerca ed è in generale considerata un grosso problema dell'informatica e del *Natural Language Processing* [15]. L'obiettivo dello *stemming* è sostanzialmente quello di riconoscere con la maggiore accuratezza e precisione possibili la correlazione tra termini a cui attribuiamo un significato comune, sfruttando questa correlazione per sostituire tutti i termini correlati con il tema corrispondente. Vediamo l'esempio delle parole "gatto", "gattone", "gattino", "gattaccio". Esse saranno tutte trattate come "gatto", grazie alla loro riduzione operata dallo *stemmer*. È bene però sottolineare che, soprattutto nella lingua italiana, la desinenza di una parola può essere rilevante ai fini della comprensione dell'orientamento di un testo: è evidente che "gattaccio" non ha lo stesso valore emotivo di "gattino". Guardando al contesto del nostro dataset, la

frase “La scuoletta di Renzi è pronta a partire” diventerebbe equivalente a “La scuola di Renzi è pronta a partire”, facendo perdere all’algoritmo un significativo contributo alla polarità del *tweet*. Uno *stemmer* sufficientemente sofisticato dovrebbe essere in grado di tenere conto di queste differenze rispetto ad esempio alla lingua inglese. Quello utilizzato per questa analisi è quello di Porter, contenuto nel pacchetto SnowballC [14] e considerato standard. Tra i suoi limiti, oltre a quello poc’anzi evidenziato, c’è quello del multilinguismo, laddove cioè in alcuni *tweet* siano presenti parole in lingua inglese, che vengono però ridotte in lingua italiana. Da notare inoltre che anche le parole con *hashtag* sono sottoposte allo stesso procedimento:

≪@matteorenz 10 appunt #italiaripart #labuonascuol #grandinsegn #buonlavor≫

3.3.3 Creazione della matrice di termini

La nostra collezione di *tweet* classificati è stata trasformata secondo il *bag of words model* [106]: ogni *tweet* è un documento, inteso come insieme di parole in cui l’ordine di apparizione non è rilevante. Ognuno di questi *tweet* viene trasformato in un vettore le cui componenti sono le parole elaborate secondo lo *stemming*, cui verrà associato dall’algoritmo il “peso” in termini di sua rilevanza nell’influenzare il sentimento negativo o positivo del testo. In altre parole, per ogni tema verrà calcolata dal classificatore la probabilità che il testo che lo contiene si attesti in uno dei due poli, o in nessuno nel caso della neutralità. Il primo passo è quello di creare la *Document Term Matrix* (in seguito indicata come DTM), ovvero una matrice con tante righe quanti sono i *tweet* del nostro dataset e tante colonne quanti sono gli *stem* dell’intero documento. Ogni elemento della matrice sarà quindi dato dal

numero di volte che lo *stem* corrispondente ad una certa colonna si trova in quel *tweet*.

<<DocumentTermMatrix (documents: 2000, terms: 5025)>>	
Non-/sparse entries:	17108/10032892
Sparsity:	100%
Maximal term length:	91
Weighting:	term frequency (tf)

Il passaggio successivo è quello di impostare la matrice in modo che i valori dei suoi elementi siano unicamente binari: non siamo interessati a sapere quante volte uno *stem* appare in un certo *tweet*, ma solo se c'è o non c'è. Fatto questo passaggio, abbiamo quello che serve al classificatore per eseguire l'analisi: la matrice dei *tweet* e la corrispondente colonna a valori -1, 0, +1, ovvero il vettore delle classi assegnate ai testi, che verrà identificata come "outcome".

3.3.4 Esecuzione del classificatore *Naive Bayes*

Un classificatore bayesiano è un classificatore basato sull'applicazione del teorema di Bayes. Esso esegue una classificazione di tipo statistico basata sul calcolo della probabilità delle cause: avvenuto un certo evento si determina la probabilità di quale causa lo abbia scatenato [83]. Il teorema si fonda sulle tre leggi fondamentali della probabilità: teorema della probabilità condizionata, teorema della probabilità composta, teorema della probabilità assoluta. Da queste tre leggi deriva il teorema di Bayes, qui enunciato nella sua forma più semplice:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

La formula si legge in questo modo: dati due eventi A e B, qual è la probabilità che si verifichi l'evento A dato che si è verificato B? Questa è pari alla probabilità che si verifichi B dato che si è verificato A, per la probabilità che si verifichi A, diviso per la probabilità di B.
Nel caso della *sentiment analysis*, il teorema di Bayes è alla base del classificatore *Naive Bayes*, uno dei più comuni classificatori probabilistici utilizzati nell'ambito del *text mining*. Il classificatore calcola la probabilità a posteriori di una classe basata sulla distribuzione delle parole nel documento, secondo il già citato modello *bag of words*: si ignora l'ordine delle parole e quindi la posizione delle parole nel documento non è rilevante; sfruttando il teorema di Bayes, calcola la probabilità con la quale una data parola appartenga ad una particolare etichetta, in base alle sue caratteristiche

$$P(label|feature) = \frac{P(feature|label)P(label)}{P(feature)}$$

$P(label)$ è la probabilità a priori di un'etichetta o la probabilità con cui un messaggio abbia quell'etichetta, $P(feature|label)$ è la probabilità che una *feature* sia presente dato che il messaggio è stato classificato con quell'etichetta, mentre $P(feature)$ è la probabilità a priori di un insieme di *feature* di apparire [48]. Ad ogni parola del *corpus* in esame, quindi, il classificatore impara ad assegnare una probabilità di positività, negatività o polarità sulla base della classificazione già fatta e quindi applicherà quella regola ai futuri testi, non già classificati, che gli saranno sottoposti. Nella classificazione del testo, il classificatore bayesiano impiegato generalmente è *naif*, ossia basato su un modello di probabilità che ipotizza come indipendenti le *feature* (le variabili predittive): in altri termini, il classificatore assume che la presenza o l'assenza di una particolare *feature* (parola) in un documento testuale non sia correlata alla presenza o all'assenza delle altre parole che compongono il

testo. Questa assunzione, come sottolineato da Heimann e Danemann [48] difficilmente può essere sostenibile, meno ancora nell'analisi testuale, dove la compresenza di determinate parole può essere rivelatrice nel comprendere il senso e il sentimento di una frase. Si pensi ai modi di dire, ai proverbi e alle "frasi fatte": se il classificatore tenesse conto della compresenza delle parole "lupo" e "pelo" nella stessa frase, potrebbe immaginare probabile che nella frase compaia anche "vizio" e tendenzialmente assegnare alla compresenza di queste tre parole una maggiore probabilità nel determinare un orientamento negativo della frase che le contiene. Allo stesso modo, ignorare le correlazioni tra *feature* consente di fare due cose che sarebbero altrimenti problematiche [48]:

1. includere un grande numero di caratteristiche. Nell'analisi testuale, le parole "uniche" hanno spesso caratteristiche predittive e i documenti spesso contengono migliaia di parole uniche. Altri modelli non sono in grado di processare con la stessa agilità un tale numero di predittori.

2. ottenere risultati solidi in modo veloce. Il classificatore *Naive* è molto efficiente e permette di essere allenato su un *training set* anche particolarmente grande.

Fatte queste premesse, passiamo alla parte pratica. Per l'applicazione del *Naive Bayes*, è stato scelto di utilizzare il pacchetto e1071 [66]. Il comando riceve in entrata la matrice e la colonna "outcome". Il classificatore opera sull'intero dataset, senza prevedere una divisione in *training* e *test*. Questo perché non ci interessa valutare la performance del nostro classificatore, ma istruire nel modo migliore possibile l'algoritmo per poi utilizzarlo sull'oggetto della nostra analisi: il *sentiment* della campagna di Starbucks. Si può quindi

realizzare una tabella di contingenza per confrontare l'*outcome* effettivo (la classe assegnata ai *tweet*) con la classe predetta invece dal modello.

	Effettivi		
Predetti	-1	0	1
-1	30,2%	0,7%	1,55%
0	6,5%	46,8%	11,15%
1	0%	0%	3,1%

La matrice di confusione, detta anche tabella di errata classificazione, restituisce una rappresentazione dell'accuratezza di classificazione statistica. Ogni entrata della matrice rappresenta su 2000 messaggi, la percentuale di testi che sono stati classificati in una certa classe, data dalle righe della matrice, sapendo che appartengono alla classe data dalle colonne della matrice. I testi che sono stati classificati correttamente sono quindi indicati sulla diagonale principale: per questo è immediato osservare dalla matrice se il modello ha commesso degli errori e in che proporzione. Grazie a questa matrice è quindi possibile osservare subito se c'è "confusione" nella classificazione di diverse classi e ottenere al contemempo le statistiche relative alle performance. Il risultato si attesta ad un 80% di accuratezza, con un errore concentrato in prevalenza nella interpretazione del *sentiment* positivo rispetto a quello neutro: il modello ha classificato correttamente solo 62 *tweet*, mentre ha assegnato classe neutra a 223 *tweet* che erano invece da etichettare come positivi.

3.3.5 Applicazione del modello *Naive Bayes*

Abbiamo allenato il nostro modello sul dataset dei 2000 *tweet* già classificati: questo significa che il classificatore ha appreso delle regole per assegnare un

sentimento a dei testi sulla base degli *stem* che lo compongono ed è quindi pronto ad applicare quelle stesse regole a qualsiasi altro dataset gli venga sottoposto. Chiaramente anche per il dataset "palme" è necessario procedere con le stesse operazioni che sono state effettuate sul dataset precedente: il *pre-processing*, lo *stemming* e quindi la trasformazione in matrice di parole. L'unica differenza rispetto allo script precedente è che in questo caso lavoriamo da un file con la sola colonna "text", in quanto i nostri *tweet* non sono classificati.

TEXT
Ecco le prime palme piantate nelle aiuole di Piazza Duomo sponsorizzate da Starbucks `https://t.co/uNWvpDNe4p`
Ancora un poco ed il Duomo diverrà una moschea in tanto piantano le palme poi si vedrà.Povera Italia
Sono arrivate le palme in Duomo ?? #palmtrees #milano #igersmilano @Duomo - Milano City `https://t.co/YH2Yr7kcCF`

Il nostro corpus viene processato con gli stessi passaggi del precedente. A questo punto è pronto per essere trasformato in *Document Term Matrix*.

```
<<DocumentTermMatrix (documents: 14230, terms: 10398)>>
Non-/sparse entries:    135251/147828289
Sparsity:               100%
Maximal term length:    38
Weighting:              term frequency (tf)
```

La DTM contiene ben 10.398 *feature* (gli *stem* complessivi del dataset), ma non tutti ovviamente sono utili per la classificazione. Con l'obiettivo di ridur-

re l'onere computazionale e di evitare di incorrere in un errore di memoria, si è deciso di ridurre il numero di *feature* ignorando quelle che appaiono in meno di 20 *tweet*. Lo script che segue identifica le frequenze dei termini e restringe la DTM includendo solo le *feature* che superano la soglia di frequenza minima stabilita. Vedremo che queste scenderanno a 1191. Prima di procedere con la classificazione, ci interessa vedere quali sono le parole più frequenti nella nostra matrice. Prepariamo la trasposta della matrice, trasformandola in una *Term Document Matrix*:

PAROLA	FREQUENZA
palm	14932
duom	6268
mil	5361
piazz	4964
banan	2752
#mil	1221
bruc	1137
dop	1072
salvin	1059
piant	951

PAROLA	FREQUENZA
mett	724
arriv	699
ital	643
sol	625
domen	497
far	487
starbucks	438
cos	398
africanizz	397
milanes	393

I primi posti in classifica sono occupati, prevedibilmente, dagli temi delle parole chiave che sono state utilizzate per effettuare la raccolta dei *tweet*. È invece interessante, ai fini della nostra indagine, notare come "bruc" e "salvin", stemmi che rimandano evidentemente al rogo delle palme e alla polemica portata avanti dal segretario della Lega Nord, siano di gran lunga più citati rispetto a "starbucks". Il dato è interessante perché sembra confermare due delle ipotesi relative alla analisi sul caso di marketing in questione,

ovvero a) il fatto che Starbucks non sia stato individuato subito dall'opinione pubblica come mandatario della posa delle palme e che la critica sia stata inizialmente rivolta solo al comune di Milano b) il fatto che la campagna di comunicazione del brand sia stata particolarmente influenzata dalla discussione politica, al punto da rendere il brand stesso meno rilevante tra gli attori dello scenario.

Fatte queste considerazioni, possiamo passare all'applicazione del modello. Il classificatore è lo stesso di prima, questa volta la funzione riceve semplicemente un'altra matrice. Procediamo quindi applicando la funzione di predizione, senza bisogno di richiamare il classificatore, che era stato preparato precedentemente:

Sentiment	Frequenza	Percentuale
-1	4818	33.85805
0	7951	55.87491
1	1461	10.26704

Secondo il nostro modello, il *sentiment* prevalente è neutro, con 7951 *tweet*, seguito da quello negativo (4818) e da quello positivo, molto ridotto (1461). Abbiamo realizzato un classificatore *Naive Bayes* e lo abbiamo allenato su un dataset di 2000 *tweet* precedentemente etichettati, per poi applicarlo ai 14230 *tweet* di rilievo per il nostro problema. Il risultato ci dice che il sentimento prevalente verso l'iniziativa di marketing realizzata da Starbucks in Piazza del Duomo sia sostanziamente neutro, con una percentuale comunque rilevante di sentimento negativo (34%). Non avremmo modo di sapere se la predizione sia corretta, se non classificando manualmente i 14230 *tweet* del nostro dataset. In questo paragrafo sfruttiamo allora le DTM già realizzate e le utilizziamo per creare e implementare un nuovo modello, con un altro

classificatore: quello basato sull'algoritmo *Support Vector Machine*. Questo ci permetterà di confrontare i risultati ottenuti, verificando se con un altro tipo di classificazione otteniamo di nuovo una prevalenza di *sentiment* neutro o se i risultati, invece, variano in modo significativo rendendo l'analisi meno conclusiva.

3.3.6 Classificazione supervisionata usando algoritmi di *Support Vector Machine*

La classificazione supervisionata con algoritmo *Support Vector Machine* consiste nell'applicazione di una funzione matematica che mappa i parametri in entrata (le *feature*) in un elemento del dominio delle classi (gli *output*). A differenza del *Naive Bayes*, dove la classificazione avviene tramite la valutazione delle probabilità degli attributi rispetto alle varie classi, il *Support Vector Machine* risponde meno ad un approccio probabilistico e più analitico. I classificatori di tipo *Support Vector Machines* (SVM) funzionano determinando i separatori lineari che sono in grado di dividere tra loro nel modo migliore le diverse classi: il modello trova infatti le migliori linee di separazione che definiscono gli iperpiani in cui dividere lo spazio di ricerca, capaci di rappresentare le diverse etichette della classificazione, mentre le linee di separazione sono ottenute massimizzando la distanza dei punti più vicini delle diverse classi.

Prendendo il caso di una classificazione binaria, quando si ha a disposizione un *training set* di dati e viene creata una funzione che riesca a classificare quei dati al meglio, l'algoritmo permette di trovare i parametri caratteristici del miglior separatore delle due classi. Teoricamente esistono infiniti iperpiani che permettono di separare perfettamente i campioni in due insiemi

distinti, ma per una classificazione corretta è necessario determinare i parametri caratteristici (w, b) dell'iperpiano che separa i dati al meglio.

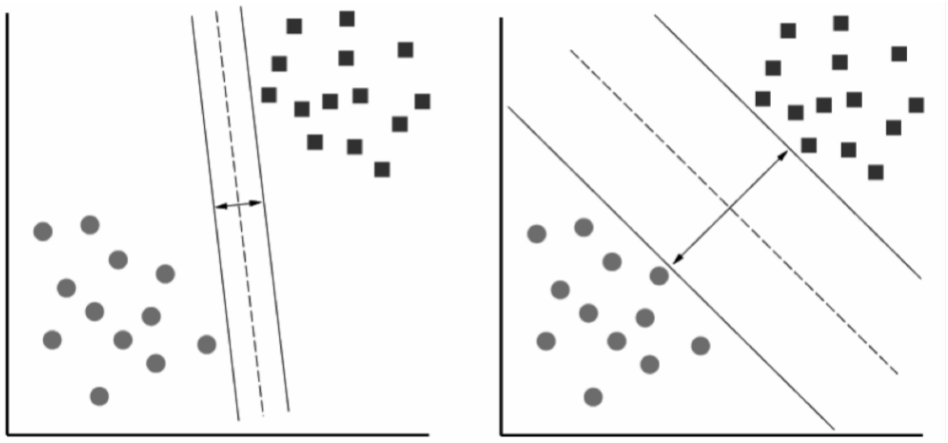

Figura 3.1: Schema di esempio di due classificazioni con SVM

In figura si vedono due classi di dati separate da due iperpiani diversi. Entrambe le soluzioni funzionano, separando il dataset in due insiemi, ma per capire qual è il migliore si può notare che mantenendo più ampio il corridoio tra le due classi si riduce di molto il rischio di *overfitting*. Di conseguenza, la classificazione di dati non appartenenti al *training set* risulterà più precisa. L'iperpiano B è, quindi, quello che offre la miglior separazione tra le due classi, poiché la distanza normale di ogni punto è la maggiore. L'algoritmo che le *Support Vector Machines* utilizzano per trovare i due parametri del separatore ottimo consiste proprio nel rendere massimo il margine appena definito, per fare in modo che la distanza tra i membri delle due classi risulti la più ampia possibile. I dati ottenuti dai testi sarebbero i più adatti per le classificazioni tramite *Support Vector Machines*, per via della natura sparsa del testo: poche delle *feature* presenti sono irrilevanti, ma tutte tendono ad essere correlate le une con le altre e generalmente organizzate

per essere all'interno di categorie linearmente separabili [2]. In letteratura [74], il classificatore *Support Vector Machine* è stato dimostrato come molto efficace nella categorizzazione testuale tradizionale, in molti casi ottenendo risultati molto migliori rispetto al *Naive Bayes*. Per lo scopo di questa tesi, non ci soffermeremo a confrontare nel dettaglio i diversi metodi: come già fatto per *Naive Bayes*, ci limiteremo a darne una descrizione e a svolgere la fase applicativa.

Di nuovo è necessario allenare il classificatore sul nostro dataset già etichettato, riutilizzando quindi la DTM precedente. Per la creazione del modello, è stato scelto di utilizzare il pacchetto RTextTools [54].
Confrontiamo i risultati del classificatore con quelli effettivi, in modo da valutare l'accuratezza:

	Effettivi		
Predetti	-1	0	1
-1	26,3%	4,45%	1,85%
0	9,7%	42%	5,95%
1	0,7%	1,05%	8%

Notiamo che l'accuratezza sul set di *training* è più bassa rispetto a quella ottenuta con il classificatore precedente: circa 76%. Confrontiamo le performance in una tabella:

SVM				**Naive Bayes**			
	-1	0	1		-1	0	1
-1	30,2%	0,7%	1,55%	-1	26,3%	4,45%	1,85%
0	6,5%	46,8%	11,15%	0	9,7%	42%	5,95%
1	0%	0%	3,1%	1	0,7%	1,05%	8%

La performance del *Naive Bayes* è chiaramente superiore, nonostante la forte imprecisione nella classificazione dei *tweet* con sentimento positivo. Per entrambi comunque il sentimento prevalente è di nuovo quello neutro, seguito da quello negativo. A questo punto possiamo effettuare la classificazione sui *tweet* relativi alle palme di Starbucks e verificare se questo modello conferma o meno i risultati precedenti.
Il classificatore con *Support Vector Machine* non smentisce il risultato precedente, ma restituisce una ratio simile a quella del *Naive Bayes*. Da notare che la percentuale assegnata al *sentiment* positivo è ancora più ridotta, mentre è sensibilmente maggiore quella del *sentiment* neutro rispetto a quanto interpretato dal classificatore *Naive Bayes*.

Nella prossima ed ultima parte di questo lavoro, amplieremo l'indagine utilizzando un modello di *sentiment* più sofisticato di quelli visti finora e spostando il focus dalla *sentiment analysis* all'*opinion mining*. Vedremo come strumenti più elaborati condurranno a risultati in parte diversi da quelli ottenuti con i modelli implementati autonomamente, ma soprattutto permetteranno di arricchire l'indagine con *insight* di più ampio spettro, mantenendo l'impegno in termini di tempo relativamente contenuto.

3.3.7 Dalla *sentiment analysis* all'*opinion mining*

Nelle sezioni precedenti abbiamo illustrato un esempio di *sentiment analysis* utilizzando un approccio *machine learning* classico, seguendo passo passo le fasi di preparazione del dato, di allenamento dell'algoritmo e quindi di produzione dell'analisi del *sentiment* nelle tre categorie: positivo, negativo, neutro.
L'obiettivo era quello di toccare con mano la *sentiment analysis* dal punto di vista pratico: solo attraverso il confronto con le problematiche insite nel

linguaggio umano, con le difficoltà e i limiti della raccolta dei dati e dell'impostazione dei paletti di ricerca, e quindi solo con l'effettivo approccio agli strumenti a disposizione per effettuare una prima analisi di *sentiment*, è infatti possibile capire davvero la sfida e l'opportunità che questa costituisce per il marketing. La convinzione che ha mosso questo tipo di analisi è proprio quella che una, pur non esuastiva, conoscenza dei modelli e degli strumenti che sono alla base della *sentiment analysis* (così come di qualsiasi altra applicazione di *data mining*) costituisca un importante valore aggiunto per qualsiasi tipo di ricerca di marketing che abbia l'ambizione di confrontarsi con il mondo dei dati *social* e delle opportunità offerte dal web.

Nei prossimi paragrafi invece, il raggio dell'indagine sarà allargato sfruttando strumenti già predisposti per la ricerca. L'obiettivo sarà quello di mostrare tutta la potenzialità della *sentiment analysis*, al punto da dimostrare come lo stesso termine si possa ritenere per certi versi riduttivo. Si ripartirà dal fenomeno delle palme e di Starbucks, ritenuto esempio – come vedremo tra poco – emblematico per le possibili applicazioni che un'analisi del sentimento può generare. In questa parte della tesi, la conoscenza dei modelli statistici dietro agli strumenti rimane utile, ma non verrà approfondita ulteriomente. Verrà invece lasciato spazio alle considerazioni più propriamente legate all'ambito del marketing e dell'uso della *sentiment analysis* applicato ai contesti della comunicazione aziendale e, nello specifico, del monitoraggio della *brand image* e della *brand reputation*. Verranno fornite una descrizione del modello su cui è basato lo strumento utilizzato per effettuare l'analisi e i riferimenti in letteratura utili ad approfondire la comparazione di questo modello con gli altri propri della *sentiment analysis*. Si passerà poi alla descrizione ed interpretazione dei risultati, che terrà più propriamente conto anche delle implicazioni e dei possibili sviluppi che l'analisi del *sentiment*

comporta.

3.3.8 iSA e Voices Analytics

Come anticipato, si riconsiderano la reputazione di Starbucks e il fenomeno di *tweeting* sulla campagna delle palme in Piazza del Duomo. Vengono quindi impiegati entrambi i dataset. Per questo tipo di analisi ci avvaliamo della piattaforma Voices Analytics, messa a disposizione da *Voices From The Blogs*. La piattaforma sfrutta la metodologia di Ceron, Iacus e Curini, offrendo la giustificazione scientifica del modello su cui la piattaforma è costruita. Alla base di Voices Analytics c'è iSA, *Integrated Sentiment Analysis* [21], ovvero un algoritmo che riprende il metodo di classificazione supervisionata Hopkins-King e ne propone una versione specificamente disegnata per lavorare con i social network e con il contesto del web, caratterizzato da abbondanza di rumore rispetto alla quantità di informazione rilevante che si può estrarre. L'algoritmo si basa sulla classificazione aggregata [21]: a differenza di quella individuale, qualora si stia analizzando un numero elevato di testi e si vogliano conoscere le proporzioni nelle varie classi si utilizza una classificazione aggregata. Per passare dalla classificazione individuale a quella aggregata si utilizza solitamente il metodo definito *"Classify-and-Count"* [20], ovvero si fa una classificazione individuale e si conta in seguito il numero di testi assegnati a ciascuna categoria. Il problema nell'utilizzo di un processo di questo tipo è che operando una classificazione individuale e poi contando il numero di testi assegnati a ciascuna categoria si amplificano gli errori di misclassificazione rispetto alla classificazione aggregata [20].
Abbiamo visto la classificazione individuale con i modelli *Naive Bayes* e *Support Vector Machine*, in cui la proporzione dei *tweet* appartenenti alle tre classi è stata valutata andando a classificare il sentimento di ogni singolo

testo. Il modello su cui è fondato iSA, invece, è basato sul metodo aggregato che fu formulato da Hopkins e King [21]. Secondo il metodo aggregato si stima direttamente la distribuzione aggregata del sentimento, usando i testi del *training set* e, come per quanto visto precedentemente, prevede una fase di codifica manuale, il *tagging*. La novità di questo approccio è che l'errore di misclassificazione decresce rispetto all'aggregazione dei risultati ottenuti da classificazione individuale, evitando il passaggio intermedio della classificazione individuale.

Indichiamo con $P(S)$ la distribuzione degli *stem* dell'intero insieme di dati (*training set* e *test set*). Poiché vale

$$P(S = s) = \sum_{j=1}^{K} P(S = s|D = D_j) \times P(D_j)$$

al variare di s in S, la distribuzione degli *stem* può essere scomposta in

$$P(S) = P(S|D)P(D)$$

dove $P(S)$ è il vettore che contiene le probabilità dei vari *stem* e $P(D)$ il vettore delle probabilità delle categorie.

La matrice $P(S|D)$ contiene le probabilità che una particolare sequenza di *stem* compaia all'interno dei testi che sono classificati secondo una particolare categoria D_j. Questa matrice $P(S|D)$ ha dimensione $2M \times k$.

Essendo nota la distribuzione $P(S|D)$ solo per i testi del *training set*, cioè quelli effettivamente codificati, si deve fare l'ipotesi che le parole utilizzate nel *training set* (TR) per esprimere D_j siano le stesse usate da tutti i testi del *corpus*, ovvero che:

$$P_{TR}(S|D) = P(S|D)$$

La precedente equazione si può scrivere quindi come:

$$P(S) = P_{TR}(S|D) \times P(D)$$

Da cui si ricava:

$$P(D) = P_{TR}(S|D)^{-1} \times P(S)$$

Il vantaggio di questo approccio è che non viene utilizzata la classificazione individuale e successivamente aggregata, ma viene stimata direttamente la distribuzione aggregata $P(D)$. Questo dovrebbe evitare la "amplificazione dell'errore" che si ottiene con il metodo *"Classify-and-Count"* [21]. L'aspetto fondamentale di cui tenere conto per il buon funzionamento di questo metodo, però, è che le categorie scelte siano esaustive: diventa fondamentale l'introduzione della categoria *Offtopics*, contenente tutti quei testi che non trattano l'argomento in esame. Hopkins e King prevedono come unica ipotesi alla base del loro metodo la rappresentatività linguistica dei testi del *training set* rispetto a tutto il *corpus* di testi, ovvero che sia verificata l'assunzione $P_{TR}(S_j|D) = P(S_j|D)$ [21]. Questo avviene quando i temi (gli *stem*) del *training set* sono presenti in numero sufficiente da poter individuare e caratterizzare ogni categoria: quando si etichettano manualmente i testi del *training set* si deve essere certi che questi siano ben rappresentativi del linguaggio usato in tutto il *corpus* di testi. Questo significa che nel *training set* deve essere presente un numero cospicuo degli *stem*, anche se non è necessario che in ciascuna categoria ci sia lo stesso numero di testi: ad esempio, se si ha a che fare con categorie che, per loro natura, sono poco numerose si procede alla ricerca di appositi testi da inserire nel *training set* in modo da arricchire la variabilità di testi in quella categoria e, qualora questo non fosse possibile, si può modificare la catalogazione evitando così la presenza di categorie poco

numerose – un passaggio, questo, che vedremo proprio in fase applicativa. La fase di etichettatura manuale diventa perciò particolarmente delicata e richiede la miglior accuratezza possibile. Sulla base della supposizione che $P_{TR}(S_j|D) = P(S_j|D)$, ovvero sulla base della rappresentatività dei testi del *training set* rispetto all'intero dataset, i risultati ottenuti saranno fortemente dipendenti dalle scelte eseguite in questa fase. Il processo di assegnazione dei testi a ciascuna categoria permetterà poi al classificatore di imparare le regole decisionali che lo portano a stimare le percentuali richieste dal set di testi etichettati manualmente. Chi effettua l'etichettatura manuale deve quindi avere ben presente quale sia il problema e in che modo assegnare correttamente i *tweet* alle categorie, quindi aver compreso esattamente il significato dell'etichetta rappresentata da una determinata categoria. Di nuovo, si vedrà meglio in fase applicativa cosa questo comporti. A questo punto è utile introdurre l'evoluzione che si intende effettuare con questa seconda analisi, permessa dall'utilizzo della piattaforma Voices Analytics, e quindi il concetto di *opinion mining* rispetto a quello visto finora di *sentiment analysis*.

3.3.9 L'evoluzione del *sentiment*

Nell'analisi finora effettuata, abbiamo trattato il *sentiment* come due facce di un problema di classificazione del testo, corrispondente al chiedersi se il *tweet* esprimesse un sentimento positivo, negativo o neutro. È facile immaginare che una simile classificazione presenti dei limiti legati non solo, a monte, alla soggettività spesso presente nella determinazione di ciò che è sentimento positivo o negativo, ma anche alla riduttività di un'analisi di marketing che si proponga solo di estrarre un'informazione di polarità (+ o -, buono o cattivo, mi piace/non mi piace) e non di indagare le motivazioni legate a quella polarità, le suggestioni che la motivano, le opinioni che la costituiscono

o influenzano.

Il problema non è di poco conto ed è anzi quanto mai rilevante, non solo per i brand e le aziende che utilizzano i social network per comunicare con le loro audience, ma anche per i social network stessi, costretti dagli stessi utenti che li utilizzano a tenere conto di simili tematiche. Se è infatti vero, come abbiamo provato a spiegare nella prima parte di questa tesi, che le piattaforme di interazione sociale come Facebook e Twitter sono sempre più un veicolo importante di comunicazione ed espressione del sé, la capacità di queste piattaforme di trasmettere in modo efficace e completo il messaggio di chi le utilizza è un problema che riguarda *in primis* quelle piattaforme. Che i gestori delle piattaforme si siano posti questo tema e si stiano interrogando sulle possibili soluzioni è un dato di fatto, di cui un caso emblematico è fornito proprio da Facebook. Se fino al 2016 l'utente di Facebook poteva esprimere la sua opinione in modo veloce con il tasto "Mi piace", oltre che con il meno immediato commento, dal 2016 Facebook ha iniziato ad introdurre le cosiddette *"Reactions"*: un multitasto che si ottiene tenendo premuto il tasto "Mi piace" e che permette di scegliere, oltre al pollice alto, altre cinque reazioni possibili rappresentate da emoji: *"Love", "Haha", "Wow", "Sigh" o "Grr"*. Le *Reactions* sono una novità che indirizza, anche se non elimina, il dibattito relativo all'assenza di un tasto "Non mi piace" in Facebook: dibattito innescato dagli stessi utenti del social.

> *"Abbiamo condotto a livello internazionale focus group e interviste per capire quali tipi di reazioni le persone vorrebbero esprimere maggiormente. Abbiamo anche approfondito le modalità con cui le persone già oggi stanno esprimendo le proprie reazioni alle storie condivise su Facebook attraverso commenti, adesivi ed emoticon. Abbiamo inoltre testato Reactions in alcuni mercati,*

> *nel corso dell'ultimo anno, e ricevuto feedback positivi ... Continueremo ad approfondire e ad ascoltare i feedback delle persone per assicurare un'esperienza utile e piacevole in tutto il mondo ... Ci interessa di più fornire alle persone modi diversi e articolati di esprimere le proprie sensazioni".* [25]

Il caso di Facebook e delle *Reactions* è un utile indicatore della necessità delle piattaforme di *social networking* di confrontarsi con quelle che si possono definire le sfaccettature del sentimento, impossibili da sintetizzare in due poli contrapposti. La stessa difficoltà di esprimere il concetto di *dislike* (alternativamente traducibile con "non mi piace" o "mi dispiace") fornisce un ulteriore esempio della capacità di un modello "polare" di catturare effettivamente il sentimento di un testo.

> *"Non volevamo creare semplicemente un tasto 'Non mi piace' perché non vogliamo che Facebook diventi un forum dove le persone votano positivamente o negativamente i post. [...] Quello che le persone vogliono è la possibilità di esprimere complicità e compassione. Non tutti i momenti sono buoni, no? E se stai condividendo qualcosa di triste – che si tratta di eventi attualità come la crisi dei migranti o di qualcosa che riguarda un lutto in famiglia – le persone potrebbero non sentirsi a loro agio nel mettere un 'Mi piace'. Ma i tuoi amici e i tuoi familiari vogliono comunque essere in grado di dirti che ti capiscono e che ti sono vicini."* [76]

Il dibattito è aperto e verte ancora sulle potenzialità, i limiti e le opportunità di crescita che i social network vedranno nel futuro. E apre la strada alla riflessione su cui si concentra questa parte del lavoro di ricerca, affrontato con la piattaforma Voices Analytics: quello, cioè, dell'evoluzione della

sentiment analysis come *opinion mining*, ovvero come contestualizzazione della polarità del sentimento all'interno di un insieme di livelli da cui si può analizzare il risultato di una azione di comunicazione come quella intrapresa da Starbucks e rappresentata dalla posa delle palme in Duomo.

Uno degli aspetti più delicati di qualsiasi analisi di *text mining* è quello rappresentato dal contesto. Gli algoritmi di classificazione basati sul *machine learning* esaminano gli elementi già classificati e costruiscono un modello statistico che, apprendendo da quanto già classificato, permette di classificare nuovi testi basandosi sulla presenza o assenza delle *feature* identificate durante il *training*. Un esempio della differenza della capacità di comprendere il contesto tra un umano e una macchina è appunto illustrato da Farina [32] facendo un paragone con le immagini. Nel caso per esempio del riconoscimento dei volti, un umano riconosce un volto perché analizza nello stesso momento tutti gli elementi di un'immagine, ricavandone la situazione rappresentata, il luogo, il tempo atmosferico, le età, i generi e le etnie in modo da comprendere il contesto in modo molto profondo e immediato. L'esempio di Farina è quello della foto di una persona che spegne delle candele su una torta: un umano che veda questa foto intuisce che si tratta di un compleanno e che quella persona è il festeggiato, oltre agli innumerevoli altri indizi che può ricavare. Allo stesso modo, sentendo qualcuno parlare nella propria lingua, l'umano comprende ed è anche in grado di intuire i termini usati qualora il suono sia distorto o la comunicazione disturbata, mentre "i programmi attualmente si limitano a contestualizzare i termini usando delle statistiche sui termini adiacenti in una certa lingua. (...) In sostanza, il *Machine Learning* mira a ottenere un'approssimazione dei risultati di un processo cognitivo senza nemmeno tentare di replicare il processo cognitivo stesso, approccio che finora ha dato scarsi risultati anche per via della enorme differenza tra le

capacità di calcolo di una macchina e quelle di un cervello animale"[32].

La contestualizzazione nella determinazione della positività o negatività dei termini usati per esprimere il *sentiment* non è un problema del solo *machine learning*, come ben descritto in tanta parte della letteratura. La cosiddetta *domain dependency*, ovvero la rilevanza del contesto rispetto al vocabolario utilizzato, è talmente fondamentale nell'attribuire correttamente la polarità (cioè classificare un testo come positivo o negativo, o neutro) che la stessa espressione può essere catalogata in modo esattamente opposto a seconda di dove venga espressa. Si pensi alla frase "Leggi il libro": nel contesto delle recensioni letterarie, esprime presumibilmente un commento positivo verso il libro, ma non si potrebbe dire lo stesso se si stesse parlando della recensione di un film [73].
L'importanza del contesto sottolinea ancora di più il limite di un approccio esclusivamente basato sulla polarità del sentimento e pone di fronte al ricercatore la necessità di allargare il campo d'indagine, andando a considerare il contesto e le motivazioni del sentimento non solo come "cause" e "contorno", ma come caratteristiche determinanti per comprenderlo e per classificarlo correttamente. Alla luce di questa considerazione, il muoversi della ricerca verso quella che viene chiamata *opinion mining* somiglia più non tanto ad una implementazione della *sentiment analysis*, quanto ad una sua naturale evoluzione.

La metodologia proposta da Ceron, Curini e Iacus [21] va proprio in questa direzione, abbinando l'analisi del *sentiment* ad una classificazione che include anche altri livelli e arricchisce, completandola, la determinazione di positività o negatività. La piattaforma offre infatti la possibilità di includere nell'indagine diverse prospettive da cui osservare il fenomeno: non solo quindi assegnare ad un *tweet* l'etichetta di "*sentiment* positivo", "*sentiment*

negativo", "*sentiment* neutro", ma anche delle ulteriori *feature*, trattate al pari del sentimento come classi di appartenenza. A differenza della determinazione del sentimento, in cui la soggettività del ricercatore è limitata alla fase del di etichettatura, nell'ambito dell'*opinion mining* la personalizzazione dell'indagine, e l'importanza del ruolo svolto dal codificatore, avviene ancora prima della etichettatura, a monte: quando, cioè, vengono scelte le caratteristiche che verranno prese in considerazione.

Nella fase preliminare dell'indagine, che consiste nell'inquadramento del problema (in questo caso, la campagna delle palme in Piazza del Duomo), il ricercatore è chiamato a scegliere in modo del tutto arbitrario le caratteristiche da tenere in considerazione per la classificazione dei testi. Ogni problema di *opinion mining*, infatti, comporterà una selezione di variabili di cui si vuole tenere conto che diventeranno etichette del testo, aggiungendosi al dato sulla sua polarità, che non è costituito a priori ma che cambia in base all'esigenza della ricerca e, soprattutto, alla capacità del ricercatore di rilevare correttamente i più livelli in cui l'opinione della popolazione osservata può muoversi. Ancora più, quindi, il modello iSA rende evidente il vantaggio e la responsabilità della classificazione di tipo supervisionato, in cui l'intervento umano gioca un ruolo fondamentale non sono nella fase di *tagging* e di *training*, ovvero di codifica di parte del dataset sulla base del sentimento e di conseguente apprendimento del modello su quanto codificato, ma anche nella fase di *pre-tagging*, cioè di scelta delle variabili su cui il modello deve classificare. In tutta la parte di preparazione della macchina, quindi, l'automazione è drasticamente ridotta, così come è significativamente ridotta la standardizzazione dell'approccio rispetto al problema. Recuperando quanto detto nella seconda parte di questa tesi a proposito della ricerca qualitativa rispetto a quella quantitativa nelle ricerche di marketing, appare chiaro come

questo tipo di analisi, rispetto alla pura *sentiment analysis*, sia in grado di fare un passo in più verso i vantaggi della ricerca qualitativa. Come avviene nell'ambito di un *focus group*, invece di omologare le domande su uno schema definito a priori (positivo/negativo/neutro), si amplia l'osservazione sulla base delle domande e degli aspetti di rilievo che possono emergere dalla ricerca, personalizzando di volta in volta l'indagine per tenere conto di quegli elementi che hanno contribuito a produrre la polarità. In questo senso anche il concetto di "neutralità", che è pure un possibile risultato dell'analisi dell'espressione soggettiva, diventa meno blando perché contestualizzato e arricchito dalle altre dimensioni.

3.3.10 La scelta delle dimensioni

Il fenomeno che si va ad indagare, ovvero la posa delle palme in Piazza del Duomo, coinvolge più di un attore e, vedremo, più livelli di analisi. Il marchio principale oggetto dell'analisi è chiaramente Starbucks, assieme al suo ingresso nel mercato italiano, di cui le palme e l'annuncio dell'apertura a Milano costituiscono il primissimo approccio di una campagna di *awareness* del pubblico italiano rispetto al brand, già ampiamente conosciuto ed affermato all'estero.

Per la determinazione delle variabili di cui tenere conto, il ricercatore è chiamato ad eseguire una prima analisi di ciò che il web "ha detto" da quando la posa delle palme è avvenuta. Dal momento che si è deciso di concentrare l'indagine soltanto su Twitter, si è proceduto alla lettura sommaria dei *tweet* per individuare gli argomenti su cui è virata la discussione. Questa lettura era stata già parzialmente svolta nella scelta delle parole chiave necessarie alla raccolta dei dati; in questa fase, la lettura viene approfondita andando a cercare non solo il "cosa" (palme, Milano, Starbucks), ma anche tutto quello

che è ruotato attorno al dibattito, quali altri attori sono stati coinvolti, in che modo la popolazione analizzata ha interagito con la notizia e quali aspetti sono stati citati nei *tweet*.
Questa indagine preliminare ha portato a individuare le seguenti componenti della "opinione" sulle palme:

```
OPINIONE POSITIVA:
Ispirano mare - spiaggia - calore
Le palme a Milano c'erano anche nel passato
Le palme ci sono anche in altre parti d'Italia
Più verde
Sono belle
Buona operazione di marketing

OPINIONE NEGATIVA: Africanizzazione
Contrarietà Starbucks
Contrasto estetico
Spesa inutile
Incoerenza
Perdita identità nazionale
Rappresentano negativamente brand Milano
```

In base alla lettura del dataset, alcuni aspetti citati in più di un *tweet* hanno catturato l'attenzione del ricercatore. All'indagine sono state quindi aggiunte componenti non tanto legate alle motivazioni dell'opinione, quanto alle suggestioni e alle associazioni riferite all'immagine delle palme così come percepita dalla popolazione, non necessariamente caratterizzate da un *sentiment*

positivo o negativo:

IMMAGINE PALME
Americano
Comico
Esotico
Furba operazione commerciale
Innovativo
Occidentali's karma
Provocatorio
Verde

Lo stesso è avvenuto anche in relazione all'altro "attore" dell'indagine, che è stato ritenuto parimenti rilevante: la città di Milano.

IMMAGINE MILANO
Associazioni con spiaggia e mare
Gusto retrò
Milano audace
Milano globale
Milano innovativa
Smog

È qui utile sottolineare che l'approccio della ricerca è propriamente quello del ricercatore di marketing. Considerata la natura dell'indagine e, soprattutto, il background della storia del marchio Starbucks intrecciato a quello del ca-

poluogo lombardo, si è ritenuto valesse la pena considerare la città di Milano non solo come teatro dell'operazione di comunicazione, ma anche come brand a sé, depositario di una identità precisa che la comunicazione di Starbucks influenza a sua volta. Tanta parte dei commenti relativi all'operazione di comunicazione di Starbucks, e quindi del relativo sentimento che ha prodotto, è stata determinata (nel bene e nel male) dall'associazione con la città e con la piazza in cui l'operazione è avvenuta.

Un'ultima considerazione riguarda infine il legame che immancabilmente la campagna ha avuto con la discussione sul piano politico. Di questo legame si è già tenuto conto nelle motivazioni del *sentiment* negativo ("africanizzazione", "perdita di identità nazionale"..) e pure nelle motivazioni del *sentiment* positivo ("Le palme a Milano c'erano anche nel passato", "Le palme ci sono anche in altre parti d'Italia"..). Tuttavia, il dibattito ha finito per includere anche *meta*-considerazioni sul dibattito stesso, cioè sulla polemica che le palme hanno generato e sulle effettive azioni, anche particolarmente forti, in cui la polemica si è concretizzata. Per questo motivo, un'ulteriore determinante del sentimento dei *tweet* è stata fatta risalire al focus sulla polemica in sé, declinata in queste categorie:

```
FOCUS SULLA POLEMICA
Chi critica è fascista/razzista
Chi critica è incoerente
Giusta polemica
Giusta polemica ma contrarietà al falò
Polemica sterile e inutile
```

Un'ultima classificazione di cui tenere conto è quella rappresentata dai mes-

saggi della categoria Offtopic, ovvero da quei *tweet* che sono stati inclusi dal collettore nella fase di raccolta a causa della presenza delle parole chiave considerate nell'indagine. Anche questi vengono classificati in fase di etichettatura e quindi non considerati nell'analisi del *sentiment*. Possiamo quindi riassumere tutte le componenti dell'indagine svolta, d'ora in poi definite "dimensioni" della ricerca che, unite al sentimento, forniranno un quadro più complesso e articolato dell'esito dell'operazione di marketing di Starbucks:

```
DIMENSIONI:
-IMMAGINE MILANO
-IMMAGINE PALME
-FOCUS SULLA POLEMICA
-SENTIMENT CAMPAGNA(positivo; negativo; neutro)
-MOTIVAZIONI SENTIMENT POSITIVO
-MOTIVAZIONI SENTIMENT NEGATIVO
```

A questo punto è utile riprendere la distinzione tra i due dataset, poi utilizzati come unico dataset nel corso dell'indagine operata nella parte prima. Ricordiamo che la raccolta dei dati copre il periodo 1 gennaio – 1 marzo, mentre la campagna delle palme si è concentrata nelle due settimane 15 febbraio – 1 marzo. Con la piattaforma di Voices Analytics, si è deciso allora di operare l'analisi del *sentiment* distinguendola nei due dataset, in modo da ottenere due diverse *sentiment analysis*:

1) La *sentiment analysis* relativa al solo brand Starbucks, nel periodo 1 gennaio – 1 marzo.

2) La *sentiment analysis* relativa alla sola campagna palme, nel periodo 15 febbraio – 1 marzo.

La decisione è stata motivata dalla volontà di confrontare il sentimento verso il brand Starbucks rispetto al sentimento verso l'iniziativa complessiva di marketing. Questo permetterà di osservare in che modo la campagna abbia impattato sul brand "principale" dell'indagine, verificando anche le fluttuazioni di tale sentimento giorno per giorno.
Una volta deciso il *framework*, è possibile passare alla fase di codifica vera e propria, che avviene agilmente con l'interfaccia di Voices Analytics.

3.3.11 L'attribuzione di parole chiave (*tagging*)

La codifica manuale diventa particolarmente delicata in sede di *opinion mining*, ovvero nel momento in cui ci si confronta con diverse dimensioni. Come anticipato nelle considerazioni preliminari, la determinazione di ciò che è *sentiment* positivo e ciò che è *sentiment* negativo può rivelarsi a volte difficile anche per il codificatore umano. Nel caso di più dimensioni d'analisi, interpretare correttamente il testo assegnandogli le categorie più adatte a rappresentarne il significato vuol dire incorrere non solo nei problemi già noti dell'ambiguità del testo (le figure retoriche dell'ironia e della preterizione, ad esempio), ma anche il rischio proprio del ricercatore di forzare una classificazione, inserendo il *tweet* in una categoria che non lo rappresenta completamente al fine di non "perdere" la quantità di informazione contenuta nel *tweet*. Su quest'ultimo aspetto, vale la pena menzionare una delle potenzialità di Voices Analytics ovvero la possibilità di includere, durante la fase di etichettatura, nuove categorie mano a mano che la lettura dei *tweet* ne fa riscontrare la necessità. Dall'altra parte invece, è possibile a valle dell'indagine, quindi al termine della fase di codifica, decidere di incorporare una o più categorie in un'altra, ritenendo una delle categorie non sufficientemente

"rappresentativa".
Un esempio può illustrare meglio in che modo effettuare la fase di *coding*:

≪Nebbiolina ed umido a Milano. Le palme di Starbucks si domandano: ma che ce sto a fa'? ≫

In questo caso, sarebbe difficile decidere in che modo classificare il sentimento rispetto a Starbucks. Il *tweet* però fornisce informazioni rispetto a:

```
IMMAGINE MILANO:
-associazioni con spiaggia e mare
IMMAGINE PALME:
-esotico
SENTIMENT VERSO LA CAMPAGNA
-negativo
MOTIVAZIONI SENTIMENT NEGATIVO
-incoerenza
```

In questo caso invece il *tweet* non è particolarmente significativo rispetto alla campagna, ma lo è rispetto al focus della polemica:

≪Vai a vedere che gli stessi a cui non piacciono palme e banani in Piazza Duomo a Milano sono quelli che andranno da #Starbucks per il caffè≫

```
FOCUS:
-chi critica è incoerente
```

Alcuni *tweet* hanno fatto riferimento in modo neutro all'immagine commer-

ciale dell'iniziativa:

≪Comunque geniali quelli di Starbucks: avessero messo delle piante qualunque non se li sarebbe filati nessuno. E, invece, le palme....≫

In questo caso l'interpretazione del codificatore è importante: il *tweet* potrebbe essere etichettato come *sentiment* neutro e fatto cadere sotto il cappello della categoria "immagine palme" come "trovata commerciale"; in questo caso invece si è deciso che il "geniali" come attributo rappresentasse comunque un valore positivo rispetto alla campagna:

```
SENTIMENT POSITIVO
MOTIVAZIONI SENTIMENT POSITIVO
buona operazione di marketing
```

È da sottolineare che la fase di *coding*, proprio per l'arbitrarietà di alcune decisioni come quella appena illustrata, viene generalmente supervisionata da un altro codificatore che abbia la stessa conoscenza del campo di indagine e che quindi possa assicurarsi della bontà della classificazione, intervenendo per minimizzare l'eventuale errore umano (una codifica involontariamente errata a causa di distrazione o inceppamento tecnico) o ancora mettere in discussione una scelta di classificazione ritenuta non idonea. Il problema non è comunque risolto completamente, come dimostrano gli studi effettuati sulle discordanze di codifica che si sono effettutati paragonando le classificazioni manuali di codificatori diversi [51]. Un'ultima considerazione riguarda l'impossibilità di assegnare più di una motivazione di *sentiment* negativo o positivo allo stesso testo: delle categorie di motivazione a disposizione, il co-

dificatore è costretto a sceglierne solo una. Nei casi, quindi, in cui il *tweet* si prestasse a essere interpretato con più di una motivazione di *sentiment*, si è scelta quella ritenuta più rappresentativa o inclusiva.

Fatte queste precisazioni, è possibile passare alla fase relativa ai risultati ottenuti. Ci limitiamo soltanto a segnalare la differenza di iSA e quindi dell'uso della piattaforma Voices Analytics rispetto all'analisi condotta precedentemente per quanto riguarda l'aspetto dell'ampiezza del *training set*. Secondo il modello Hopkins-King, su cui è stato sviluppato iSA, il numero minimo di testi da codificare per l'apprendimento dell'algoritmo non può essere calcolato a priori: ciò che conta è invece avere sufficienti codifiche di ogni categoria considerata [51]. Negli approcci tradizionali esistono delle formule che, sulla base del numero di categorie e dell'ampiezza del *corpus*, ci permettono di determinare il valore di numerosità del *training set* rispetto al test set come misura percentuale, dipendente quindi dall'ampiezza della popolazione d'osservazione. Nel caso del modello Hopkins-King, invece, non si ha una numerosità a priori di osservazioni da etichettare, bensì un numero individuale di codifiche per ogni categoria: la codifica manuale dovrebbe perciò proseguire fino a che non si raggiunga un numero sufficientemente elevato di osservazioni catalogate per quella categoria. Non si conosce un numero di codifiche di testi ottimale: empiricamente, si ritiene che secondo il modello Hopkins-King un numero tra le 30 e le 50 codifiche sia sufficiente. La quantità di lavoro di codifica manuale richiesta per avere dei risultati accurati dipende quindi:

1) dall'effettiva rappresentatività delle categorie: si ricorda a questo proposito quanto accennato precedentemente sulla decisione eventuale, in fase di *tagging*, ovvero di incorporare due categorie in una, includen-

done una troppo specifica (e quindi rara, poco rappresentativa) in una più generale.

2) dalla quantità di osservazioni *offtopic*, che rendono quindi non significativa la singola codifica di questa categoria e chiedono al codificatore di procedere con la successiva. La quantità di osservazioni *offtopic* può essere anche molto elevata, particolarmente nell'ambito dei social network, in cui la raccolta dei dati è affidata alle parole chiave e agli *hashtag*.

Nel nostro caso, la codifica manuale ha portato ad analizzare un totale di 292 post, 11 dei quali rilevati come *offtopic* e quindi scartati dall'analisi. È opportuno segnalare che per una delle categorie inizialmente previste per la dimensione "Motivazioni di *sentiment* positivo", ovvero la categoria "Rappresentano positivamente il brand Starbucks", non si sono trovate evidenze sufficienti; la categoria è stata fatta convogliare in "Buona operazione di marketing", che era invece già prevista.

3.3.12 I risultati

La piattaforma Voices Analytics processa i testi seguendo sostanzialmente le stesse regole che abbiamo visto nella parte terza: il dataset viene sottoposto alla fase di *pre-processing* che rimuove le *stop word*, i "RE" del *retweet*, la punteggiatura, gli spazi bianchi, l'eventuale codice HTML, i link, producendo una *bag of words* composto di temi. Il resto della procedura include l'apprendimento del modello e la conseguente classificazione dei dati del test set sulla scorta di quanto codificato nel *training set*, fino alla produzione del report di *sentiment analysis* (e di *opinion mining*) che segue.

Tabella 3.1

SENTIMENT CAMPAGNA PALME	%
Negativo	61,8
Positivo	30,9
Neutro	7,3

Tabella 3.2

MOTIVAZIONE SENTIMENT NEGATIVO	%
Contrasto estetico	26,3
Rappresentano negativamente brand Milano	21,2
Contrarietà Starbucks	19,5
Incoerenza	17,3
Africanizzazione	12,7
Perdità identità nazionale	1,8
Idea pessima	1,2

Tabella 3.3

MOTIVAZIONE SENTIMENT POSITIVO	%
Ispirano mare - spiaggia - calore	28,8
Sono belle	28,4
Le palme a Milano c'erano anche nel passato	13,8
Più verde	13,5
Le palme ci sono anche in altre parti d'Italia	10,8
Buona operazione di marketing	4,6

La prima tabella raffigura la misura del *sentiment* verso la campagna delle palme come dato medio rispetto al periodo 15 febbraio – 1 marzo. Quasi il

62% del *sentiment* complessivo è negativo. Risulta invece molto contenuto il sentimento neutro: secondo questo modello l'utente medio, quando ha parlato delle palme, lo ha fatto per esprimere un parere. I risultati sembrano a prima vista diversi da quelli ottenuti con le due classificazioni precedenti, che vedevano un sentimento prevalentemente neutrale e, in secondo luogo, negativo. Tuttavia, sono da rilevare due fondamentali differenze rispetto alla *sentiment analysis* effettuata con *Naive Bayes* e poi con *Support Vector Machine*:

1) Nelle prime due classificazioni, è stata utilizzata per il training una collezione di *tweet* relativi ad argomenti vari e non specifici rispetto al caso in oggetto; al contrario, il training è stato effettuato su un sottoinsieme del dataset complessivo.

2) La poca incidenza del valore neutro si può spiegare anche con la decisione, in fase di etichettatura, di classificare come neutri solo i *tweet* di testate giornalistiche e blog di news di vario genere che si limitassero a veicolare l'informazione della posa delle palme; nei casi invece in cui la testata riportasse citazioni di terzi, si è deciso di classificare il *tweet* come effettiva espressione di un parere, anche se "per interposta persona".

Nonostante queste due rilevanti differenze, anche la classificazione con *Naive Bayes* e *Support Vector Machine* ha dato un risultato tendenzialmente negativo. La classificazione con iSA non stravolge in modo sostanziale l'analisi.

1. Motivazioni *sentiment* negativo Fatte queste considerazioni preliminari relative al *sentiment*, l'interesse si sposta sull'analisi dell'opinione,

intesa come pensiero soggettivo che giustifica un determinato sentimento. La tabella 3.2 illustra le componenti del *sentiment* negativo, in ordine di dimensione: al primo posto tra le opinioni dei detrattori della campagna c'è il contrasto estetico che le palme creano rispetto all'ambiente di piazza Duomo, seguito a stretto giro dalla sua opinione complementare, ovvero il fatto che le palme "rappresentano negativamente il brand Milano". Il fatto che i due primi posti delle opinioni negative siano legati a Milano ed agli effetti che la posa delle palme possano avere sulla città conferma quanto intuito già in sede di analisi preliminare: non è il solo brand Starbucks ad essere attore del fenomeno, ma anche il brand Milano come identità portatrice di un particolare mix comunicativo che entra nell'equazione relativa alla bontà o meno dell'iniziativa commerciale.
Nelle motivazioni negative, il secondo dato di particolare interesse è il terzo posto in merito alla "contrarietà a Starbucks": una categoria che il codificatore ha incluso ritenendo che parte della polemica non fosse motivata tanto dalla posa delle piante, quanto dalla generale avversione all'ingresso della multinazionale americana in suolo italiano. Di tutte le motivazioni negative, questa è sicuramente quella che può essere di maggiore interesse per il brand: quasi il 20% dei commenti su Twitter nel periodo considerato riflette l'opinione negativa dell'italiano verso Starbucks. Vedremo poi nel confronto con il *sentiment* verso il brand quanto questo dato sia effettivamente rappresentativo.

2. Motivazioni *sentiment* positivo Il 28,8 e il 28,4 dei *tweet* che esprimono un parere favorevole lo giustificano ritenendo le palme un simbolo di calore, mare, positività, o quantomeno ritenendo che siano "belle", negando il contrasto estetico che è invece nelle motivazioni del *sentiment* negativo.

Al terzo posto vediamo la motivazione "le palme a Milano c'erano anche nel passato": è uno dei temi emersi a inizio del dibattito, quello in risposta alla polemica contro le palme che sarebbe stata giustificata da una incoerenza della pianta rispetto al contesto territoriale. Di questo tema si è tenuto conto anche in relazione alla "Immagine Milano" e in particolare alla categoria del "Gusto retrò", riscontrata nei *tweet* che hanno associato le palme all'Ottocento recuperando anche vecchie foto della città.

Infine è di assoluto rilievo il 13,5% occupato dal "Più verde" come motivazione favorevole alla posa delle palme. In questa categoria rientrano sia le osservazioni degli utenti sul fatto che la città sia "grigia", "inquinata", e che giustificano quindi il loro apprezzamento delle piante come occasione per Milano di diventare più green, sia i commenti relativi più genericamente ad uno spirito ambientalista, non necessariamente legato quindi ad una esigenza contigente della città.

3. L'immagine delle palme

IMMAGINE PALME	%
Esotico	25,5
Provocatorio	17,6
Occidentali's karma	17,4
Americano	15,2
Furba - operazione commerciale	13,0
Innovativo	6,1
Verde	3,3
Comico	1,8

Veniamo alle osservazioni relative alle altre due dimensioni di interesse per l'*opinion mining*: quelle in cui il modello ha catturato suggestioni e associazioni rispetto alla immagine degli attori considerati, non necessariamente legata in modo univoco ad un *sentiment* o ad un altro. Per essere ancora più chiari su questo punto, ricorriamo ad un esempio: la categoria "smog" sulla dimensione "immagine Milano" può essersi ritrovata sia in un *tweet* di *sentiment* negativo (*i.e.* "Milano è inquinata, non saranno due palme a risolvere la situazione"), sia in uno di *sentiment* positivo (*i.e.* "Milano è inquinata, ben vengano le iniziative green!").

La maggioranza dei *tweet* menzionanti le palme suggerisce che la popolazione le associa prevalentemente con l'idea di "esotico", nel senso lato di estraneo rispetto al contesto italiano. Si sottolinea di nuovo come questo non sia necessariamente associato ad un parere negativo. L'esotismo è stato richiamato con le associazioni al continente africano e ai paesi del Medioriente, ma anche con Miami, Los Angeles e gli Stati Uniti, di cui si è tenuto conto anche nella categoria a sé "Americano". A differenza delle motivazioni

di *sentiment*, infatti, si ricorda che per le categorie relative alla dimensione "Immagine" e alla dimensione "Focus" la scelta non era alternativa, ed era invece possibile selezionare più di una *feature*. Una caratteristica di interesse, meritevole di menzione speciale, è quella di "Occidentali's Karma". La scelta di inserirla nelle categorie date in analisi al modello è stata dovuta alla constatazione, in fase di ricerca preliminare e di lettura sommaria del dataset, che la concomitante diffusione della canzone vincitrice del Festival di Sanremo (tenutosi nel periodo 7-11 febbraio) aveva creato un'associazione tra il testo e il messaggio della canzone. La suggestione è stata accolta dal codificatore come tentativo (l'aspettativa era di non arrivare ad un numero sufficiente di classificazioni in fase di *tagging* e di dover quindi accorpare la categoria con un'altra) e si è rivelata invece più significativa di quanto previsto. Di nuovo si sottolinea la forza del modello Hopkins-King, che ha rovesciato il paradigma su cui si sono basati i modelli di classificazione supervisionata più noti: mentre in precedenza infatti l'analisi dalla scelta delle parole da cercare nei testi del dataset avveniva prima e a prescindere dai testi del *corpus*, il modello in questo caso è capace di catturare anche le parole "suggerite" dalla lettura dei testi, con un dizionario che sia proprio del contesto e del fenomeno che altrimenti il codificatore avrebbe difficilmente immaginato di considerare.

Rimane da segnalare il 13% relativo alla "Furba operazione commerciale", categoria che associa alle palme esclusivamente una campagna mediatica, vedendole come simbolo di una operazione di comunicazione "dichiarata". Il dato è di maggiore interesse se confrontato con quello della categoria "verde": appena la metà. Tra chi vede nelle palme una trovata di marketing e chi ci vede un simbolo ambientalista, vincono di oltre il doppio i primi.

4. L'immagine di Milano

IMMAGINE MILANO	%
Associazioni con spiaggia e mare	39,1
Milano globale	19,7
Gusto retrò	18,3
Smog	13,2
Milano audace	7,5
Milano innovativa	2,2

Consideriamo adesso la dimensione "Immagine Milano". Al primo posto c'è la categoria che associa l'introduzione delle palme in Duomo alla spiaggia e al mare, vedendo quindi un'evoluzione dell'immagine di Milano verso un contesto "marittimo" o vacanziero. Si evidenzia quello che potrebbe essere un possibile sviluppo dell'indagine, ovvero un più approfondito monitoraggio della *brand image* di Milano a molti mesi dalla posa delle palme, andando a comprendere se questa associazione rilevata nelle due settimane di osservazioni perduri nel tempo: in altre parole, se la posa delle palme in Duomo contribuisca nel lungo periodo a modificare la percezione di Milano come città "urbana", "grigia", "lavoratrice", "frenetica".

Da tenere in considerazione il secondo posto su Milano globale: il capoluogo meneghino vanta già una predisposizione cosmopolita che gli è stata riconosciuta ulteriormente in seguito ad Expo2015. Sarebbe quindi interessante, in ottica di futuri sviluppi della ricerca, capire se questo carattere *global* sia più apprezzato o denigrato e quanto l'introduzione delle palme e di eventuali altri elementi di arredo urbano "del mondo", o comunque ritenuti poco rappresentativi dell'italianità, possa essere accolta di buon grado o no dalla cittadinanza.

Infine, poco sotto l'attitudine *global* è da considerare il "gusto retrò", segnale che il tema della presenza delle palme in Duomo nel 1800 ha avuto un certo seguito ed è stato ripreso nel corso del dibattito.

5. Il focus sulla polemica

FOCUS SU POLEMICA	%
Polemica sterile e inutile	39,3
Giusta polemica	36,6
Giusta polemica ma no a falò	12,9
Chi critica è incoerente	9,9
Chi critica è fascista/razzista	1,3

La polemica sulla discussione stessa è un'ulteriore dimensione dell'indagine di *opinion mining*, derivata dal tema principale: raccoglie le impressioni della popolazione osservata sugli sviluppi del dibattito, incluso l'atto di vandalismo con cui ignoti hanno dato fuoco nella notte alle palme del Duomo. Il focus è interessante perché, pur spostando l'attenzione dalla campagna marketing alla polemica vera e propria, costituisce un indicatore dell'"umore del web" e in qualche misura può essere interpretata come misura dell'intensità del *sentiment*: quanto negativa è l'opinione verso la posa delle palme? Se chi ha espresso un parere ha anche aggiunto un commento in relazione agli atti vandalici o alle prese di posizione politiche, questa dimensione dell'indagine permette di catturare quel dato e restituire una contestualizzazione più precisa del sentimento. Quasi il 40% dei *tweet* ha ritenuto la polemica "sterile e inutile": significa che, indipendentemente dalla propria considerazione personale sulla campagna, che può essere positiva o negativa, l'utente ha ritenuto che se ne stesse facendo una questione più grossa di quanto meritevole.

Dall'altra parte, un consistente 37% ha invece espresso che la polemica fosse motivata – un numero che aumenterebbe ulteriormente se si includessero i *tweet* polemici *tout cour*, quelli cioè che hanno fomentato la polemica senza rivendicare espressamente il diritto a farlo. Il 13% dei *tweet* ha alzato la voce contro il falò, prendendo una posizione netta rispetto all'atto vandalico, ma rivendicando comunque che la polemica fosse opportuna: anche questo dato è interessante, perché, forse controintuitivamente, rafforza il fatto che le critiche verso la posa delle piante fossero comunque legittime. Unito al precedente 37%, il dato complessivo in relazione al *sentiment* negativo si può leggere come indicatore di una negatività intensa verso le palme: una lettura, questa, confermata dai toni generalmente molto accesi che sono stati riportati dall'opinione pubblica durante tutto il periodo considerato e manifestati anche fisicamente, con le occupazioni in piazza Duomo e con lo stesso atto vandalico.

6. L'evoluzione del *sentiment* Arrivati a questo punto dell'indagine, torniamo a Starbucks: il monitoraggio del sentimento verso il marchio ha volutamente coperto uno spettro più ampio rispetto a quello della campagna delle palme. Il dataset relativo al periodo 1 gennaio – 1 marzo è stato oggetto di una parallela indagine di *sentiment*, non quindi riferita soltanto alla campagna delle palme, ma comprendente anche il mese e mezzo precedente. Questo dato è utile per realizzare un confronto rispetto al sentimento generato dalla campagna e può rappresentare, per l'azienda, uno strumento di *benchmark* della sua brand image con cui analizzare in che modo iniziative di comunicazione e di marketing, così come eventi esterni che lo riguardano, influiscano sull'impressione soggettiva della popolazione.

SENTIMENT STARBUCKS GENNAIO-FEBBRAIO	%
Negativo	48,5
Neutro	25,9
Positivo	25,7

Il sentimento medio del periodo 1 gennaio – 1 marzo verso il brand è in misura significativa negativo. A contribuire a questo dato, come vediamo nella tabella sotto, sono le rilevazioni anche precedenti alla campagna delle palme, di cui proviamo a rendere conto nell'analisi successiva:

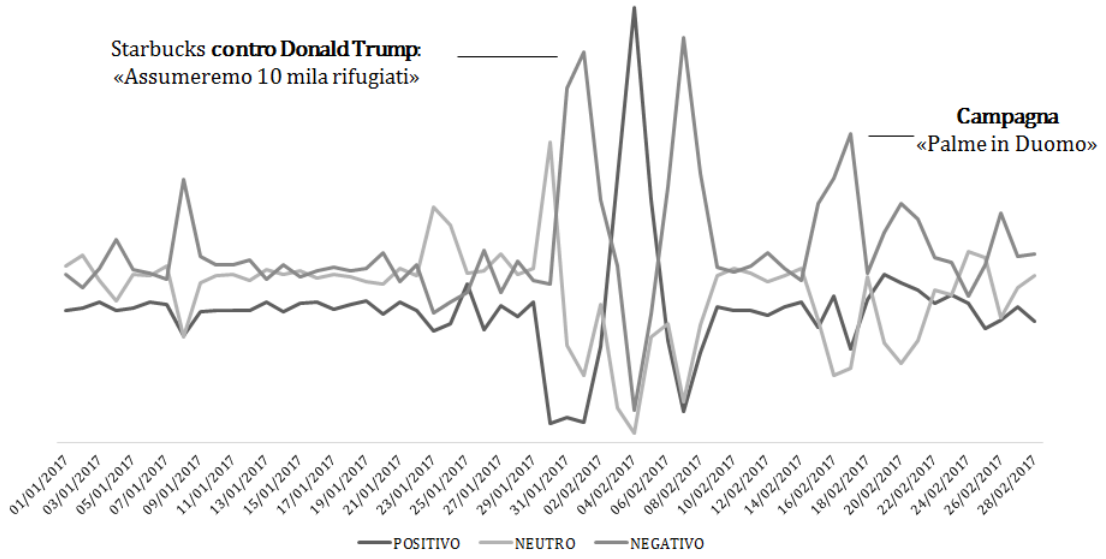

Figura 3.2: Andamento del sentiment verso Starbucks dall'1 gennaio al 28 febbraio 2017

Il grafico mostra l'andamento dei 3 *sentiment* nel periodo considerato. Nell'insieme di osservazioni giornaliere, si è cercato di individuare in quali momenti del periodo considerato si ottenessero picchi di *sentiment*, sia in positivo che in negativo: anche a colpo d'occhio, è chiaro che quelli negativi sono in netta prevalenza.
Una prima considerazione riguarda la frequenza di picchi negativi dal 15 febbraio in poi, giorno in cui la posa delle palme è avvenuta: sui due mesi considerati, i picchi negativi di sentimento verso Starbucks sono concentrati proprio nelle ultime due settimane. Il picco più alto è il 17 febbraio, con 65% di *sentiment* negativo: è in linea con l'evoluzione del dibattito, dal momento che l'attore protagonista (Starbucks appunto) non era stato immediatamente individuato dall'opinione pubblica, inizialmente concentrata più sul Comune di Milano e in parte condizionata dalla credenza errata che la posa delle palme fosse un'iniziativa dell'amministrazione.
Un altro dato interessante è che da quando la posa è avvenuta non si sono più verificati picchi positivi di *sentiment* per Starbucks: anche questo è sostanzialmente in linea con quanto ci si aspettava.

Il periodo precedente alla campagna offre a sua volta alcuni spunti interessanti. Pur non avendo monitorato gli eventi e i fenomeni precedenti alla campagna e correlati al brand, si è provato a cercare evidenza dei picchi di *sentiment* negativo e dei picchi di *sentiment* positivo là dove le rilevazioni sono state effettuate: su Twitter. Si è quindi cercato per le date corrispondenti ai picchi di individuare le possibili discussioni che giustificassero quel sentimento, aiutando la ricerca con la lettura delle notizie relative a Starbucks che fossero state riprese in quelle date dai giornali e dai siti web italiani. Per il primo picco negativo, l'8 gennaio 2017, si è inviduata la notizia ripresa da alcune testate italiane [87] in merito alla competizione di Starbucks con Mac

Donald's. Il 31 gennaio, il *sentiment* negativo raggiunge il 75%, seguito dal 1 febbraio con l'82% e un 50% il 2 febbraio: la notizia di quei giorni è la polemica del colosso americano contro Donald Trump, in cui Starbucks annuncia di voler assumere 10mila rifugiati [94]. Subito dopo il trend diventa positivo, arrivando al picco del 91% il 4 febbraio, in concomitanza con la notizia diffusa da più testate [92] sulla volontà di Starbucks di aprire 350 assunzioni anche a Milano. Si potrebbe continuare; ciò che interessa in questa sede è però dare conto del *sentiment* verso il brand in relazione alla campagna iniziata il 15 febbraio, paragonandolo ai periodi precedenti (e, in ottica di sviluppo della ricerca, a quelli successivi). Sono principalmente due i dati che sono stati ritenuti maggiormente interessanti osservando l'andamento del *sentiment* nel periodo considerato:

La volubilità del *sentiment* Per volubilità si intende la capacità dell'umore degli utenti Twitter di spostarsi significativamente, raggiungendo anche picchi molto alti, in modo repentino e rispondendo immediatamente a notizie e fenomeni avvenuti nello stesso momento. Non è una novità che i social network più di altri spazi di comunicazione siano capaci di far circolare le informazioni, aprire i dibattiti, accendere le polemiche, portare l'attenzione su un tema (il *Trending Topic* che detta spesso ciò di cui si parlerà sui canali più tradizionali); dall'altra parte è però interessante notare come sia assolutamente repentino anche il cambiamento dell'umore medio, esattamente da un giorno all'altro. A questo proposito riprendiamo l'avvertimento di Matthew Russell [103]: come tutte le opinioni, il sentimento è intrinsecamente soggettivo da persona a persona, e può essere persino completamente irrazionale. Per questo motivo diventa fondamentale scavare in un ampio e rilevante bacino di dati quando si prova a misurare il sentimento: nessun dato in particolare

è necessariamente rilevante, ma è l'aggregato che conta. Il sentimento di un individuo verso un brand o un prodotto può essere influenzato da uno o più cause indirette; "qualcuno può avere una brutta giornata e fare un *tweet* e un commento negativo su qualcosa di cui altrimenti avrebbe avuto una opinione piuttostro neutrale in merito. Con un campione sufficientemente ampio di dati, i valori anomali sono diluti nel dato aggregato" [103].
Quanto sottolineato da Russell si potrebbe estendere. Non solo il sentimento di un individuo può essere influenzato da fattori esterni che lo portano ad esternare, per la natura di immediatezza comunicativa del social network, il suo *sentiment* in modo non mediato da una riflessione e sedimentato nel tempo: lo può essere anche il sentimento di tutta la popolazione del web.

L'impatto delle palme sul *sentiment* di Starbucks È da notare il relativamente poco significativo impatto che l'arrivo delle palme ha avuto sul periodo considerato rispetto al brand Starbucks. I picchi negativi più alti si sono avuti ben prima rispetto alla campagna, mentre il *sentiment* negativo in seguito alla posa delle palme è sì stato registrato dal sentimento del brand, ma in misura sostanzialmente minore rispetto ad altre notizie. Questo aspetto è interessante, ma non esaustivo. Ad una prima interpretazione, infatti, si potrebbe essere portati a pensare che la posa delle palme non abbia influito particolarmente sulla *brand image* complessiva di Starbucks vista, generalizzando, dall'Italia. Questa conclusione però sarebbe riduttiva. Dal momento che il sentimento molto probabilmente cambia nel tempo, in base all'umore di una persona, agli eventi e così via, è solitamente importante guardare ai dati anche dal punto di vista temporale [103]. In primo luogo, come già spiegato, il *sentiment* non si può considerare un parametro sufficiente a rendere conto della reputazione di un marchio, né in generale di un qualsiasi

fenomeno. Questo è dovuto sia alla natura volatile del sentimento stesso (si rimanda al punto 1, ma anche al grafico che rappresenta l'andamento variabile del *sentiment*) sia alla più generale impulsività che caratterizza il veicolo di comunicazione social. Il Real Time Web [27], espressione utilizzata per indicare la tendenza tipica delle piattaforme sociali di creare e sviluppare condivisione di contenuti in tempo reale, favorisce una comunicazione istantanea e, per certi versi, istintiva: si pensi all'utilizzo del *direct marketing* e del *location based advertising* per favorire l'acquisto d'impulso, sfruttando l'immediatezza del messaggio comunicato e della reazione dell'utente [99] o ancora alla "caduta delle barriere" [19]. Di questa istintività si deve tenere conto nel momento in cui si traggono inferenze dai contenuti sviluppati dal web, a maggior ragione se le inferenze sono supportate soltanto dall'analisi delle polarità assunte dal *sentiment*.

In secondo luogo, poi, va ricordato l'orizzonte temporale in cui l'indagine è stata effettuata. Il dataset relativo alla campagna è stato costruito in due settimane di osservazioni immediatamente successive alla posa delle palme e contemporaneamente al divampare e poi all'affievolirsi del dibattito. Questa raccolta, benché ritenuta rappresentativa dell'opinione della popolazione rispetto a Starbucks e alla campagna, andrebbe estesa anche ai mesi successivi alla posa e possibilmente ad un orizzonte temporale ben più lungo, che includa anche, ad esempio, l'apertura e l'effettivo inizio delle attività commerciali di Starbucks a Milano.

Conclusioni

Come anticipato nei capitoli introduttivi, la scelta di analizzare l'operazione di comunicazione di Starbucks e della posa delle palme a Milano è stata dovuta alla volontà di proporre un caso di studio che rivelasse il potenziale della *sentiment analysis* e, allo stesso tempo, costringesse a fare i conti con la sovrapposizione dei livelli di comunicazione che si verifica quando più attori vengono coinvolti nell'operazione stessa. In questo caso, mentre è stato individuato come attore principale il colosso americano in qualità di marchio che agisce da comunicatore in ottica promozionale e commerciale, dall'altra parte è stata riconosciuta la città di Milano come portatrice a sua volta di un'immagine e di una identità messe in gioco nello scenario della campagna di comunicazione, non solo quindi come luogo degli eventi. La stessa discussione ci sarebbe stata se la posa delle palme non fosse avvenuta in piazza Duomo, ma in piazza del Plesbiscito a Napoli?

Il ruolo di Milano e del suo brand ci è parso fondamentale per contestualizzare e comprendere i risultati dell'*opinion mining* e per dare realmente conto degli effetti della comunicazione di Starbucks e delle reazioni della popolazione di Twitter. Contemporaneamente all'inclusione di Milano come "brand da osservare", si è avvertito come rilevante anche il ruolo della comunicazione politica. Tra le prime domande che ci si è posti: cosa succede al modello di *sentiment analysis* quando una operazione di comunicazione

nata a scopo puramente commerciale sfocia nella politica? Quando il modello analizza il *sentiment* della campagna, sta ancora misurando il *sentiment* verso il brand/prodotto, o finisce per misurare il *sentiment* verso un simbolo strumentalizzato?

Partendo da queste domande, la decisione di osservare il fenomeno da due punti di vista paralleli (quello del sentimento complessivo di Starbucks nei due mesi e quello del sentimento specifico della campagna nelle due settimane) è sembrata ancora più opportuna. Allo stesso tempo, la possibilità offerta da Voices Analytics di includere nell'indagine più dimensioni ha permesso di toccare anche la discussione politica, dando così la dovuta rilevanza al suo ruolo all'interno del dibattito. A questo proposito, uno dei dati del report effettuato dal modello ci è sembrato indicativo: le due categorie di motivazione del *sentiment* negativo su cui si è più concentrata l'opinione pubblica non occupano i primi posti per percentuale di grandezza. L'accusa di "africanizzazione" è del 12,7% dei tweet a *sentiment* negativo, la "Perdita di identità nazionale", concettualmente più fine e quindi più rara come espressione nei *tweet* (ma sostanzialmente incorporabile nella prima) solo all'1,8%. Le motivazioni del "No" sono state in percentuale maggiore legate al gusto estetico e alle conseguenze sull'immagine di Milano, ponendo invece in secondo piano la tematica "politica" che è stata oggetto della protesta in piazza e dell'atto vandalico. Di converso, lo stesso punto di vista che ha motivato i pareri negativi è quello che ha motivato i positivi: ai primi posti per un *sentiment* favorevole c'è il fatto che le palme "ricordino il mare, la spiaggia" e "siano belle". Ne risente anche l'immagine del brand Milano, per il 40% associato proprio ad un *mood* "marittimo" e vacanziero. Semplificando, si potrebbe arrivare a concludere che la dialettica tra i pro-palme e i contro-palme si riduca ad un banale *de gustibus* di natura estetica: a chi piacciono e a chi

no. Una lettura, questa, che può rappresentare un'utile semplificazione per quanto riguarda il sentimento, ma che taglia fuori tutte le altre dimensioni e le loro categorie. Il caso in oggetto è in questo senso emblematico perché obbliga il ricercatore, motivato dal monitoraggio della *brand image* e della *brand reputation*, anche a fare i conti con il resto dei piani di comunicazione interessati dal marchio o dalla campagna. La numerosità dei piani coinvolti rende ancora più fondamentale, soprattutto quando il modello di classificazione è supervisionato, avere bene in mente cosa si sta cercando – in altre parole, fare le domande giuste.

APPENDICE

APPENDICE A

Listati di R

A.1 Applicazione su dataset “Training”

```
# carico i pacchetti richiesti
>library(readxl)
>library(TextWiller)
>library(tm)
>training <- read_excel("C:/Users/consuelo.angioni/
   Desktop/training.xlsx")

# creo il dataset
>outcome <- training$CLASS
>outcome <- as.factor(outcome)
>tweets <- training$TEXT

# utilizzo normalizzaTesti per rimuovere maiuscole,
   emoticon, punteggiatura...
```

```
>tweets<-normalizzaTesti(tweets, tolower = TRUE,
    normalizzahtml = TRUE, normalizzacaratteri = TRUE,
    normalizzaemote = TRUE, normalizzapunteggiatura =
    TRUE, normalizzaslang = TRUE, fixed = TRUE, perl =
    TRUE, preprocessingEncoding = TRUE, encoding = "UTF
    -8", sub = "", contaStringhe = c("\\?", "\\!", "@",
    "#", "(|euro)","(\\$|dollar)", "SUPPRESSEDTEXT"),
    suppressInvalidTexts = TRUE, verbatim = FALSE,
    remove = TRUE)

# converto nel formato richiesto per utilizzare
    PlainTextDocument
# e rimuovere url e parole tipiche di Twitter
>tweets <- as.data.frame(tweets)
>matrice <- cbind(outcome, tweets)
>tweets <- Corpus(VectorSource(tweets))
>corpus <- Corpus(VectorSource(matrice$tweets))
>corpus <- tm_map(corpus, PlainTextDocument)
>corpus <- tm_map(corpus, stripWhitespace)
>corpus <- tm_map(corpus, removeWords, c("wwwurlwww"))

>corpus <- tm_map(tweets, removeWords, c("rt"))

# rimuovo le stopwords
> corpus <- tm_map(corpus, removeWords, stopwords('
    italian'))
```

```
# creo lo stemming del documento
> corpus <- tm_map(corpus, stemDocument, "italian")

# creo la DTM
> dtm <- DocumentTermMatrix(corpus)

# creo la funzione per classificare e la applico
> convert_count <- function(x) {
+     y <- ifelse(x > 0, 1,0)
+     y <- factor(y, levels=c(0,1), labels=c("No", "Yes
   "))
+     y
+     }
> dataset <- apply(dtm, 2, convert_count)

# importo il pacchetto necessario per usare Naive Bayes
# e lo applico ottenendo in "pred" il risultato della
# classificazione
>require(e1071)
>classifier <- naiveBayes(dataset, outcome, laplace =
   1)
>pred <- predict(classifier, newdata=dataset)
```

A.2 Applicazione Naive Bayes su Dataset "Palme"

```
# leggo il file .xlsx da cui ottengo i dati e lo
```

```
    converto
# in un dataframe
>palme <- read_excel("C:/Users/consuelo.angioni/Desktop
    /palme.xlsx")
>palme<-as.data.frame(palme)
>tweetpalme<-palme$text

# normalizzo, rimuovo le stopwords e applico lo
    stemming
>tweetpalme<-normalizzaTesti(tweetpalme, tolower = TRUE
    , normalizzahtml = TRUE, normalizzacaratteri = TRUE,
     normalizzaemote = TRUE, normalizzapunteggiatura =
    TRUE, normalizzaslang = TRUE, fixed = TRUE, perl =
    TRUE, preprocessingEncoding = TRUE, encoding = "UTF
    -8", sub = "", contaStringhe = c("\\?", "\\!", "@",
    "#", "(|euro)",

                                                "(\\$|dollar)
    ", "SUPPRESSEDTEXT"), suppressInvalidTexts = TRUE,
    verbatim = FALSE, remove = TRUE)
>corpuspalme <- Corpus(VectorSource(tweetpalme))
>corpuspalme <- tm_map(corpuspalme, PlainTextDocument)
>corpuspalme <- tm_map(corpuspalme, stripWhitespace)
>corpuspalme <- tm_map(corpuspalme, removeWords, c("
```

```
    wwwurlwww"))
>corpuspalme <- tm_map(corpuspalme, removeWords, c("rt
    "))
>corpuspalme <- tm_map(corpuspalme, removeWords,
    stopwords('italian'))
>corpuspalme <- tm_map(corpuspalme, stemDocument, "
    italian")

# ottengo la DTM
> dtmpalme <- DocumentTermMatrix(corpuspalme)

# salvo gli stem più frequenti
>fivefreq <- findFreqTerms(dtmpalme, 20)

# riduco la DTM
>dtmpalme.reduced <- DocumentTermMatrix(corpuspalme,
    control=list(dictionary = fivefreq))
>datasetpalme <- apply(dtmpalme.reduced, 2,
    convert_count)
> m = t(dtmpalme.reduced)
> m <-as.matrix(m)
> v <- sort(rowSums(m),decreasing=TRUE)
> d <- data.frame(word = names(v),freq=v)

# applico Naive Bayes
> pred <- predict(classifier, newdata=datasetpalme)
```

A.3 Applicazione Support Vector Machine su Dataset "Palme"

```
# carico il pacchetto
>require(RTextTools)

# creo un container su cui applicare SVM
>container <- create_container(as.matrix(dtm), outcome,
     trainSize = 1:2000, testSize <- 1:2000, virgin=
    FALSE)

# istruisco il modello
>classifier <- train_model(container, algorithm = c("
    SVM"))
> table(as.numeric(as.factor(outcome)), results[,"
    SVM_LABEL"])



>pred<-as.factor(results$SVM_LABEL)
>conf.mat <- confusionMatrix(pred, outcome)
```

Bibliografia

[1] Agarwal A., Xie B., Vovsha I., Rambow O., Passonneau R. (2011). Sentiment analysis of Twitter data. In *Proceedings of the Workshop on Language in Social Media*, LSM 2011, pp. 30–38

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

[3] Akkaya, C., Wiebe, J., Mihalcea, R. (2009). Subjectivity word sense disambiguation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Vol. 1. Association for Computational Linguistics, pp. 190–199

[4] Amati, G., Bianchi, M., Marcone, G. (2014). Sentiment estimation on twitter. In *Proceedings of the 5th Italian Information Retrieval Workshop* (IIR'14), Vol. 1127. CEUR Workshop Proceedings, pp. 39-50.

[5] Anderson, E. (1998). Customer satisfaction and word-of-mouth, *Journal of Service Research*, Vol. 1 No. 1, pp. 5-17.

[6] Andreevskaia, A., Bergler, S. (2006). Sentiment Tagging of Adjectives at the Meaning Level. In *Advances in Artificial Intelligence, Canadian Society for Computational Studies of Intelligence*, pp. 336-346

Bibliografia

[7] Andrew Ng Courses. Url: http://www.andrewng.org/courses/

[8] Autogrill sigla accordo strategico con Starbucks (2011). Url: http://www.autogrill.com/it/comunicati-stampa/autogrill-sigla-accordo-strategico-con-starbucks

[9] Basile P., Novielli N. (2014). Predicting tweet sentiment polarity combining micro-blogging, lexicon and semantic features. In *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & the Fourth International Workshop EVALITA 2014*, pp. 58-63

[10] Bates, M. (1995). Models of natural language understanding. In *Proceedings of the National Academy of Sciences of the United States of America, Vol. 92*, pp. 9977–9982.

[11] Berger, P. D., Nasr, N. I. (1998). Customer lifetime value: Marketing models and applications. In *Journal of Interactive Marketing*, Vol.5, pp.1-17

[12] Bettetini, M. (2004) *Il maestro e la parola. Il maestro, la dialettica, la retorica, la grammatica.* Bompiani, Milano

[13] Blinov, P.D., Klekovkina, M.V., Kotelnikov, E.V., Pestov, O.A. (2013): Research of lexical approach and machine learning methods for sentiment analysis. Url: www.dialog-21.ru/media/1226/blinovpd.pdf

[14] Bouchet-Valat, M. (2014). SnowballC: Snowball stemmers based on the C libstemmer UTF-8 library. R package version 0.5.1. https://CRAN.R-project.org/package=SnowballC

[15] Branca, M. M. (2013/2014). *Strategie di Sentiment Analysis: confronti e nuove proposte*. Tesi di Laurea Magistrale, Università degli Studi di Padova

[16] Cacco, F. (2012). *Metodo Hopkins-King per la Sentiment Analysis: una valutazione basata sui tweets della campagna elettorale*. Tesi di Laurea Magistrale, Università degli Studi di Padova

[17] Calzolari N., Magnini B., Soria C., Speranza M. (2009). La lingua italiana nell'era digitale. META-NET Collana Libri bianchi, Springer, Milano

[18] Cappellari, R. (2007). *Le strategie aziendali*. McGraw-Hill, Milano

[19] Carapellese, C. (2013). *Impatto dei social media sul cervello. Opinioni e scenari della trasformazione delle facoltà mentali*. Tesi di Laurea Magistrale, Politecnico di Milano

[20] Ceron, A., Curini, L., and Iacus, S. M. (2014). *Social media and sentiment analysis. L'evoluzione dei fenomeni sociali attraverso la rete*. Springer, Milano

[21] Ceron, A., Curini, L., and Iacus, S. M. (2016). iSA: A fast, scalable and accurate algorithm for sentiment analysis of social media content. In *Information Sciences*, pp.105-124.

[22] Chomsky, N. (1958). *Linguistics, logic, psychology, and computers*. In Carr, J. (1958) *Computer programming and artificial intelligence*. Ann Harbor, Michigan *CD* Chomsky, N. (1998). *Linguaggio e problemi della conoscenza*. Il Mulino, Bologna

[23] Coleman, J. (2012). The Life of Slang. URL: http://www.independent.co.uk/arts-entertainment/books/reviews/the-life-of-slang-by-julie-coleman-7554016.html

[24] Cosa vuole fare Starbucks a Milano (2017). Url: http://www.ilpost.it/2017/02/28/starbucks-milano/5etwf (ultima visita il 25 maggio 2017)

[25] Cosimi, S. (2016) Facebook Reactions, da oggi il "mi piace" non è più solo. Url: http://www.repubblica.it/tecnologia/social-network/2016/02/24/news/facebook_reactions_da_oggi_il_mi_piace_non_e_piu_solo-134124337/ (ultima visita il 25 maggio 2017)

[26] D'Adda, C. (2010). *Social media intelligence: L'analisi della influence nel microblogging*. Tesi di Laurea Magistrale, Politecnico di Milano

[27] De Felice, L. (2011). *Marketing Conversazionale*. Il Sole 24 Ore, Milano

[28] De Riccardis, S. (2017) Palme bruciate a Milano, c'è l'identikit del vandalo. Url: http://milano.repubblica.it/cronaca/2017/02/20/news/milano_palme_bruciate_identikit_vandalo-158760929/ (ultima visita il 25 maggio 2017)

[29] Deshmukh S.N., Shirbhate A. G. (2016). Feature extraction for sentiment classification on Twitter data. In *International Journal of Science and Research (IJSR)*, Vol. 5, pp: 6633-6639

[30] Digital in 2017 Report. Url: http://wearesocial.com/it/blog/2017/01/digital-in-2017-in-italia-e-nel-mondo (ultima visita il 25 maggio 2017)

[31] Esuli A., Sebastiani F. (2006). SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, pp. 417-422

[32] Farina, J. (2013). *Estrazione, sentiment analysis e rappresentazione di grandi quantità di messaggi pubblici tramite le tecnologie Big Data*. Tesi di Laurea Magistrale, Politecnico di Milano

[33] Feinerer, I, Hornik, K. (2015). tm: Text Mining Package. R package version 0.6-2. https://CRAN.R-project.org/package=tm

[34] Fellbaum, C. (1998). *Wordnet: an electronic lexical database*. MIT Press, Boston

[35] Fielding, N. G., Lee, R. M., Blank, G. (2008). *The SAGE handbook of online research methods*. Sage, London

[36] Forman, G. (2003), An Extensive Empirical Study of Feature Selection Metrics for Text Classification. In "Journal of Machine Learning Research", pp. 1289-1305

[37] Fromm, J. (2014) Why Starbucks is Still Number One With Millennials. Millennial Marketing, url: `http://www.millennialmarketing.com/2014/02/why-starbucks-is-still-number-one-with-millennials/`

[38] Gagliardi, C. (2014). *Origini delle teorie sociali sulla comunicazione. Fondamenti, capisaldi classici, protagonisti.* LAS, Roma

[39] Gama, J., de Carvalho, A.C. (2009). Machine Learning. In *M. Khosrow-Pour Encyclopedia of Information Science and Technology*,pp. 2462-2468

[40] Genovese, A. (2011). *Le Nuove Regole del Marketing*. In "Social Media Marketing. Manuale di Comunicazione Aziendale 2.0" (2011) a cura di G. Di Fraia, Hoepli, Milano

[41] Gentry, J. (2015). twitteR: R Based Twitter Client. R package version 1.1.9. https://CRAN.R-project.org/package=twitteR

[42] Given, L. M. (2008). *The Sage Encyclopedia of Qualitative Research Methods*. Sage Publications, Los Angeles

[43] Granovetter, M.S. (1973). The Strength of Weak Ties. In *American Journal of Sociology*, Vol. 78, pp. 1360–1380

[44] Grimmer J., Stuart B. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. In *Political Analysis*, pp: 267-297

[45] Hadot, P. (2007). *Wittgenstein e i limiti del linguaggio*. Bollati Boringhieri, Torino

[46] Haskova, K. (2015). Starbucks Marketing analysis, CRIS Bulletin. Url: `https://doi.org/10.1515/cris-2015-0002`

[47] He, W., Zha, S., Li, L. (2013). Social media competitive analysis and text mining: A case study in pizza industry. In *International Journal of Information Management*,Vol.33, pp: 464–472

[48] Heimann, R., Danneman, N. (2014). *Social Media Mining with R*. Packt Publishing, Birmingham

[49] Heisenberg, W. K. (1961). *Fisica e filosofia*. Il Saggiatore, Milano

[50] Honkela, T., Korhonen, J., Lagus, K., Saarinen, E. (2013) Five-dimensional sentiment analysis of corpora, documents and words. In Villmann T., Schleif F. M., Kaden M., Lange M. (2013) "Advances in Self-Organizing Maps and Learning Vector Quantization. Advances in Intelligent Systems and Computing", Vol. 295. Springer, Cham

[51] Hopkins, D., King, G. (2010). A method of automated nonparametric content analysis for social science. In *American Journal of Political Science*, Vol. 5, pp. 229-247

[52] Introduction to the semantic web. Url: `http://www.cambridgesemantics.com/semantic-university/introduction-semantic-web` (ultima visita il 25 maggio 2017)

[53] Jurafksky, D., Martin, J. (2009). *Speech and natural language processing: A introduction to natural language processing, computational linguistics, and speech recognition.* Prentice Hall, New Jersey

[54] Jurka, T., Collingwood, L., Boydstun, A., Grossman, E., van Atteveldt (2012). *RTextTools: Automatic text classification via supervised learning.* Url: `https://faculty.washington.edu/jwilker/tft/Rtexttools_use.pdf`

[55] Kaplan, A. M., Haenlein, M. (2010). Users of the World, Unite! The Challenges and Opportunities of Social Media. In Business Horizons, pp. 59–68

[56] Katz, E. (1957). The two-step flow of communication: An up-to-date report on an hypothesis. In *The Public Opinion Quarterly*, Vol.21, pp. 61-78

[57] Katz, J. M. (2009). *Defining influence as a strategic marketing metric*. Forrester Research Inc, Boston

[58] Kotler, P., Armstrong, G. (2010). *Principles of Marketing*. Prentice Hall, Boston

[59] Levine, R., Locke, C. , Searls, D., Weinberger, D. (2000). *The Cluetrain Manifesto: The End of Business as Usual*. Perseus Books, New York

[60] Liu, B. (2010). *Sentiment Analysis and Subjectivity. Handbook of Natural Language Processing*. Chapman and Hall, Florida

[61] Lolli, G. (2017). *Ambuigità. Un viaggio fra letteratura e matematica*. Il Mulino, Bologna

[62] Maguire, J. S., Hu, D. (2013). Not a simple coffee shop: local, global and glocal dimensions of the consumption of Starbucks in China. In *Journal for the Study of Race, Nation and Culture*, Vol. 19, pp. 670-684

[63] McQuail D., Windahl S. (1993). *Communication Models for the Study of Mass Communications*. Longman, London

[64] McWorther, J. (2013). Txtng is killing language. JK!!! . Url: `https://www.ted.com/talks/john_mcwhorter_txtng_is_killing_language_jk`

[65] Melloncelli, D. (2012/2013). *Sentiment analysis in Twitter*. Tesi di laurea, Università di Bologna

[66] Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F. (2017). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.6-8. https://CRAN.R-project.org/package=e1071

[67] Milano, cambia look il verde piazza Duomo. Arrivano palme e banani (2017). Url:
http://www.affaritaliani.it/milano/milano-cambia-look-il-verde-piazza-duomo-con-palme-banani-460073.html

[68] Mosca, G. (2016) Social media in Italia: crollano Google+ e Twitter, esplode Snapchat. Url:
https://www.wired.it/internet/social-network/2016/04/04/social-media-italia-crollo-twitter-esplode-snapchat/ (ultima visita il 25 maggio 2017)

[69] Mudinas, A., Zhang, D., Levene, M. (2012). Combining lexicon and learning based approaches for concept-level sentiment analysis. In *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining*, pp. 1-8

[70] Niola, M. (2014). *Hashtag. Cronache da un paese connesso*. Bompiani, Milano

[71] Paccagnella, L. (2010). *Sociologia della comunicazione*. Il Mulino, Bologna

[72] Pang B., Lee L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pp. 115-124

[73] Pang B., Lee. L. (2008). Opinion mining and sentiment analysis. In *Foundations and Trends in Information Retrieval*, Vol.2, No 1-2, pp. 1–135

[74] Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of EMNLP*, pp. 79–86

[75] Pastore, A., Vernuccio, M. (2008). *Impresa e Comunicazione. Principi e strumenti per il Management*. Apogeo, Milano

[76] Perché Facebook non ha un tasto "Non mi piace" , Url: `http://www.ilpost.it/2014/12/12/tasto-non-mi-piace-facebook/` (ultima visita il 25 maggio 2017)

[77] Pinker, S. (2007). *The Stuff of Thought: Language As a Window Into Human Nature*, Penguin Group, New York

[78] Polizzi, D. (2017) Starbucks, a Milano il più grande negozio d'Europa. Url: `http://milano.corriere.it/notizie/cronaca/17_febbraio_27/starbucks-milano-piu-grande-negozio-d-europa-palme-piazza-duomo-piaceranno-186e4d20-fd30-11e6-8717-6cdb036394a5.shtml` (ultima visita il 25 maggio 2017)

[79] R. Url: `https://www.r-project.org/` (ultima visita il 25 maggio 2017)

[80] Rambocas, M. Gama J., (2013). Marketing research: The role of sentiment analysis. Url: `http://wps.fep.up.pt/wplist.php`

[81] Ravindran, S. K., Garg, V. (2015). *Mastering Social Media Mining with R*. Packt Publishing, Birmingham

[82] Rogelberg, S.G. & Stanton, J.M. (2007). Understanding and Dealing with Organizational Survey Nonresponse. In *Organizational Research Methods*, Vol.10, pp. 195-209

[83] Russell, S., Norvig, P. (2003). *Artificial Intelligence: A Modern Approach*. Prentice Hall, New Jersey

[84] Russo, M., (2014) *Social Media e Sentiment Analysis*. Springer, Milano

[85] Saif, H., He, Y., Alani, H. (2012). Semantic Sentiment Analysis of Twitter. In Cudré-Mauroux P. *The Semantic Web – ISWC 2012. Lecture Notes in Computer Science*, Vol. 7649. Springer, Berlin

[86] Satta, M. (2015). *Nuove strategie per la sentiment analysis applicata ai social network*. Tesi di Laurea Magistrale, Politecnico di Milano

[87] Scarci, E. (2017) Sfida americana nella ristorazione: Starbucks vuole scalzare McDonald's. Url: `http://emanuelescarci.blog.ilsole24ore.com/2017/01/08/sfida-americana-nella-ristorazione-starbucks-vuole-scalzare-mcdonalds/` (ultima visita il 25 maggio 2017)

[88] Schultz, H. (2012). *Onward: How Starbucks Fought for Its Life without Losing Its Soul*. Hoepli, Milano

[89] Scocco, D. (2007). Why Starbucks is not present in Italy? Url: `http://innovationzen.com/blog/2007/01/15/why-starbucks-is-not-present-in-italy/` (ultima visita il 25 maggio 2017)

[90] Sentipolc. Url: `http://www.di.unito.it/~tutreeb/sentipolc-evalita16/` (ultima visita il 25 maggio 2017)

[91] Solari, D., Sciandra, A., Rinaldo, M., Redaelli, M., Finos, L. (2016). TextWiller: Collection of functions for text mining, specially devoted to the italian language. R package version 2.0

[92] Starbucks annuncia 350 assunzioni. Url: http://www.ecodibergamo.it/stories/bergamo-citta/starbucks-milano-350-assunzionitanta-bergamo-alla-presentazione_1226538_11/ (ultima visita il 25 maggio 2017)

[93] Starbucks Financial Data, Google Finance. Url: https://www.google.com/finance?cid=655693 (ultima visita il 25 maggio 2017)

[94] Starbucks sfida Donald Trump «Assumeremo 10mila rifugiati». Url: http://www.ecodibergamo.it/stories/Cronaca/starbucks-sfida-donald-trumpassumeremo-10mila-rifugiati_1222026_11/ (ultima visita il 25 maggio 2017)

[95] Strickland, T. (1999) Strategic management. Url: http://mhhe.com/business/management/thompson/11e/case/starbucks-2.html

[96] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M. (2011). Lexicon-based methods for sentiment analysis. In *Computational Linguistics*, Vol.37, Issue 2, pp. 267-307

[97] Thompson, J.B. (1998). *Mezzi di comunicazione e modernità*. Il Mulino, Bologna

[98] Turney, P. D. (2002), Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the*

40th Annual Meeting of the Association for Computational Linguistics, (2002), Philadelphia, Pennsylvania, pp. 417-424

[99] Unni, R., Harmon, R. (2007) Percieved effectivenessof push vs. pull mobile locatin based advertising. In *Journal of Interactive Advertising*, Vol.7, Issue 2, pp. 28-40

[100] Vescovi, T. (2007). *Il marketing e la rete*. Il Sole 24 ore, Milano

[101] Vohra, S.M., Teraiya, J.B. (2013). A comparative study of Sentiment Analysis techniques. In *Journal Of Information, Knowledge And Research In Computer Engineering*, Vol.2, pp. 317-322

[102] Web 2.0, `http://www.treccani.it/enciclopedia/web-2-0/` (ultima visita il maggio 25, 2017)

[103] Webb, J. (2011) With sentiment analysis, context always matters. Url: `http://radar.oreilly.com/2011/03/sentiment-analysis-context.html` (ultima visita il 25 maggio 2017)

[104] Wittgenstein, L., (1990). *Tractatus logico-philosophicus e Quaderni* 1914-1916, Einaudi, Milano

[105] Xiaowen, D. X., Bing L. B., P. S. P. Yu (2008). A holistic lexicon-based approach to opinion mining. In *Proceedings of the Conference on Web Search and Web Data Mining*,New York, pp. 231–240

[106] Zhang, Y., Jin, R., Zhou, ZH. (2010) Understanding bag-of-words model: a statistical framework. In *International Journal of Machine Learning and Cybernetics*, Vol. 1, Issue 1, pp. 43–52

[107] Zhao Y., Dong S., Li L. (2014). Sentiment Analysis on news comments based on supervised learning method. In *International Journal of Multimedia and Ubiquitous Engineering*, Vol.9, pp. 333-346

Ringraziamenti

Chi mi conosce sa quanto sofferta e attesa questa ultima pagina sia stata per me. Lo sa soprattutto perché con ogni probabilità ha dovuto patire un po' della mia ansia: in misure, occasioni e modi diversi, ne hanno fatto parte tutti. Il primo ringraziamento è, di default, per la pazienza.
Ringrazio il professor Carlo Gaetan, per la flessibilità e la disponibilità senza le quali sarebbe stato difficile conciliare la realizzazione di questa tesi con le esigenze della mia vita professionale. Ringrazio il professor Andrea Stocchetti, che mi ha permesso di chiudere l'indice mettendoci tutto quello che volevo ci fosse. Un grazie sentitissimo va a tutto il team di *Voices From The Blogs* e in particolare a Fabio e al professor Andrea Ceron, che mi hanno seguito nella parte di sviluppo del progetto. E a proposito di Voci, dalla carta stampata però, grazie alla redazione de *La Voce di Rovigo*: di fronte a un foglio, non c'è volta in cui quanto imparato come giornalista non torni utile. Il grazie va allora anche a *This MARKETERs Life*, il progetto editorale con cui ho declinato la scrittura sugli argomenti che adesso sono il mio lavoro. Grazie al dottor Alberto Becattelli, che mi ha dato l'occasione di cimentarmi subito nel marketing, ad un tempo in cui iniziavo appena a sognare gli uffici di Madison Avenue. Grazie a Lucia che mi ha portato in quelli di via Quadrio, con tutte le belle coincidenze che ci legano – una delle quali è la presenza del suo cognome in bibliografia. Grazie ad Alice, collega affezionata, supporter e complice. Grazie a Met, che ha reso il sole disponibile nella mia amata Casa Macello. A Chiara, Anna, Beatrice e Enrica, che non hanno mai smesso di essere le mie compagne di classe. A Mara e Arianna che sono le compagne di merende. A Juliette che lo è stata di molte avventure. Grazie – sapevate che sarebbe arrivato – all'India – eccolo – e a tutti quanti ho lasciato in vari e bellissimi angoli di mondo, assieme a sparsi pezzettini della mia anima. Grazie ad Alberto, che uno di quei pezzettini me lo ha voluto far riavere. Ad Arianna, la mia migliore amica; a Mattia, il mio migliore amico. Grazie alla mia grande famiglia, di cui sento da sempre fortissimi la stima e l'affetto. A Consy, Nonno Gianni e Nonna Rina, che di quella famiglia fanno ancora parte. Grazie a Max, per tutti gli ovvi motivi, che includono il fatto di esserci, e sempre sopra il dovuto. Grazie a papà e mamma, mentori ed eroi. A Davide, idolo senza pari e formidabile aiutante. Grazie a Dan, mio sentimento e mia analisi.