

**La struttura informativa delle recensioni online
Un sistema di indicatori diagnostici e il ruolo
dell'intelligenza artificiale**

di Enrico Bonetti, Josè Compagnone

IPE Working Paper

N. 34

September 17, 2026

ISSN 2284-1229

<https://doi.org/10.5281/zenodo.22813882>

La struttura informativa delle recensioni online. Un sistema di indicatori diagnostici e il ruolo dell'intelligenza artificiale

Enrico Bonetti¹, Josè Compagnone²

Abstract

L'analisi delle recensioni online viene spesso presentata come un problema di scala, che i modelli linguistici di grandi dimensioni consentirebbero finalmente di superare grazie alla capacità di elaborare corpus troppo ampi per l'analisi manuale. Questo lavoro propone una prospettiva diversa, secondo la quale il principale vincolo non è soltanto computazionale, ma concettuale, perché il valore informativo dell'analisi dipende anzitutto dalla scelta di che cosa osservare e misurare.

Il contributo propone un sistema di otto indicatori elementari, calcolabili su informazioni direttamente osservabili e senza richiedere, nella prima fase, interpretazione semantica del testo. Gli indicatori sono organizzati in tre gruppi: come valutano i clienti, come si esprimono e come risponde l'impresa. La loro applicazione è illustrata su due dataset volutamente differenti: 2.142 recensioni italiane relative a undici brand in cinque settori e 701.528 recensioni statunitensi della categoria beauty. L'analisi empirica ha finalità prevalentemente dimostrativa ed esplorativa: mostra come misure semplici possano rendere visibili configurazioni che la sola valutazione media o indicatori aggregati tendono a occultare. In particolare, l'asimmetria di risposta (cioè la differenza tra il tasso di risposta alle valutazioni negative e quello alle positive) rende osservabili diverse modalità di allocazione dell'attenzione dell'impresa e, nel dataset analizzato, varia sensibilmente anche all'interno

¹ Università degli studi della Campania "Luigi Vanvitelli" – enrico.bonetti@unicampania.it

² Università degli studi di Napoli Federico II – jose.compagnone@unina.it

dello stesso settore. Inoltre, il lavoro individua tre ruoli dell'intelligenza artificiale nel processo: validazione dell'integrità del dataset, lettura orientata dagli indicatori e supporto all'esecuzione operativa. In tutti e tre i casi, il modello resta strumento di supporto, mentre la definizione delle misure e la loro interpretazione rimangono responsabilità del ricercatore.

Parole chiave: recensioni online; passaparola digitale; indicatori diagnostici; risposte aziendali; marketing relazionale; intelligenza artificiale; ricerche di marketing.

Online review analysis is often framed as a problem of scale, one that large language models would finally make it possible to overcome thanks to their ability to process corpora too large for manual analysis. This paper advances a different perspective: the main constraint is not only computational but conceptual, because the informational value of the analysis depends first and foremost on deciding what should be observed and measured.

The paper proposes a system of eight elementary indicators, calculated from directly observable information and requiring no semantic interpretation of the text in the initial stage. The indicators are organized into three groups: how customers evaluate, how they express themselves, and how firms respond. Their application is illustrated using two deliberately different datasets: 2,142 Italian reviews covering eleven brands across five sectors and 701,528 U.S. reviews in the beauty category. The empirical analysis is primarily demonstrative and exploratory: it shows how simple measures can reveal patterns that average ratings or aggregate indicators tend to conceal. In particular, response asymmetry—the difference between the response rate to negative ratings and the response rate to positive ratings—makes different patterns of managerial attention visible and, in the dataset analyzed, varies substantially even within the same sector. The paper also identifies three roles for artificial intelligence in the analytical process: validating dataset integrity, enabling indicator-guided reading, and supporting operational execution. In all three cases, the model remains a support tool, while the definition of the measures and the interpretation of their implications remain the responsibility of the researcher.

Keywords: online reviews; electronic word of mouth; diagnostic indicators; managerial responses; relationship marketing; artificial intelligence; marketing research.

1. Introduzione

Le recensioni online costituiscono da tempo una fonte rilevante per comprendere il comportamento dei consumatori, la reputazione delle imprese e le dinamiche del word of mouth elettronico (Dellarocas, 2003; Godes e Mayzlin, 2004; Chevalier e Mayzlin, 2006). Negli ultimi anni, tuttavia, il dibattito sulla loro analisi si è progressivamente spostato sulle opportunità offerte dall'intelligenza artificiale generativa e, in particolare, dai modelli linguistici di grandi dimensioni (LLM). La capacità degli LLM di elaborare grandi quantità di testo ha infatti ampliato in misura considerevole le possibilità della ricerca di marketing e dell'analisi dei dati non strutturati (Sarstedt et al., 2024; Arora, Chakraborty e Nishimura, 2025; Grewal et al., 2025). Corpus di migliaia o centinaia di migliaia di recensioni, impossibili da leggere manualmente, possono oggi essere analizzati in tempi e con costi prima impensabili.

La maggiore capacità di elaborazione, tuttavia, non coincide necessariamente con una maggiore capacità di produrre conoscenza utile alle decisioni. È questa l'osservazione da cui prende avvio il presente lavoro. Se applicata indistintamente a un insieme molto ampio di recensioni, l'analisi automatica tende innanzitutto a restituire ciò che nel dataset ricorre con maggiore frequenza: i temi dominanti, gli aspetti più apprezzati o criticati, le valutazioni prevalenti. Si tratta di informazioni certamente utili, ma spesso già note, almeno nelle loro linee essenziali, a chi gestisce la marca. Passare da mille a un milione di recensioni analizzate può quindi rendere più precisa la rappresentazione di fenomeni già evidenti senza necessariamente far emergere quelli meno visibili e, proprio per questo, potenzialmente più rilevanti per la decisione.

Il punto, dunque, non è soltanto leggere di più, ma sapere che cosa cercare. Prima della capacità di elaborare i dati viene la capacità di formulare domande alle quali quei dati possano fornire risposte utili. Da questa prospettiva, il principale limite dell'analisi delle recensioni non è necessariamente computazionale, ma può essere concettuale e riguarda la scelta delle dimensioni da osservare e degli indicatori attraverso i quali renderle misurabili. Un indicatore semplice, costruito per rispondere a una domanda manageriale precisa, può in alcuni casi fornire insight molto più utili di quelli derivanti da un'analisi molto sofisticata applicata all'intero corpus senza una preventiva selezione di ciò che si intende osservare.

Il presente lavoro propone, in questa prospettiva, un sistema di otto indici elementari, costruiti a partire da informazioni strutturate normalmente disponibili nelle piattaforme di recensione: valutazione numerica, data di pubblicazione, presenza o assenza di una risposta dell'impresa ed eventuale momento in cui tale risposta viene fornita. Gli indici non richiedono, per essere calcolati, alcuna interpretazione semantica del testo. Alcuni descrivono la struttura delle valutazioni espresse dai clienti, evidenziando fenomeni che la sola media tende a comprimere; altri osservano invece il comportamento dell'impresa nei confronti di chi recensisce, permettendo di cogliere aspetti della relazione che il semplice tasso complessivo di risposta non rende visibili. L'obiettivo non è contrapporre questi indicatori all'analisi testuale né sostenere che misure semplici siano, in assoluto, superiori a tecniche più sofisticate. La proposta è piuttosto quella di utilizzare gli indicatori per stabilire dove guardare e che cosa cercare, riservando successivamente l'analisi semantica ai fenomeni e alle porzioni del corpus per i quali essa può generare un reale valore informativo.

L'analisi empirica condotta nel presente lavoro svolge, coerentemente con questa impostazione, una funzione prevalentemente dimostrativa ed esplorativa. Non è costruita per produrre stime rappresentative dell'universo delle recensioni online né per verificare relazioni causali tra le variabili considerate. Il suo scopo principale è mostrare come gli otto indici possano essere concretamente applicati a dati reali e quale capacità abbiano di mettere in evidenza configurazioni e comportamenti che gli indicatori aggregati normalmente utilizzati possono lasciare in secondo piano. In altri termini, i dati vengono utilizzati non tanto per dimostrare che determinate configurazioni caratterizzino in generale un settore o un mercato, quanto per verificare se il sistema di misurazione proposto sia in grado di rendere osservabili differenze manageriali altrimenti poco evidenti.

A questo scopo vengono utilizzati due dataset volutamente molto differenti. Il primo comprende 2.142 recensioni pubblicate su una piattaforma generalista italiana e relative a undici brand appartenenti a cinque settori: servizi bancari e di pagamento, distribuzione beauty, gioielleria online, moda online e distribuzione culturale. Il secondo comprende 701.528 recensioni di prodotti della categoria beauty pubblicate su una piattaforma di commercio elettronico statunitense nell'arco di vent'anni. La distanza tra i due dataset — per lingua, mercato, piattaforma, ordine di grandezza e natura dell'atto comunicativo — è intenzionale. Essa consente di mostrare l'applicazione degli stessi strumenti in contesti molto diversi e, al tempo stesso, di osservare quali indicatori mantengano un comportamento

interpretabile al variare del contesto e quali, invece, richiedano criteri di lettura specifici. In quest'ultimo caso, la dipendenza dal contesto non costituisce necessariamente un limite dello strumento, ma un'informazione utile per definirne correttamente le condizioni di impiego.

Il lavoro è articolato come segue. Il paragrafo 2 presenta il quadro teorico di riferimento, richiamando in particolare la letteratura sul word of mouth elettronico, sulla struttura delle valutazioni online e sul recupero del servizio. Il paragrafo 3 introduce il sistema degli indici, partendo dai limiti della valutazione media, tuttora dominante nella pratica manageriale ma incapace di rappresentare pienamente la struttura delle esperienze che la determinano. Il paragrafo 4 descrive i dataset e le procedure di preparazione dei dati, con particolare attenzione alle condizioni di validità poste dall'estrazione automatica. Il paragrafo 5 presenta i risultati dell'applicazione empirica. Il paragrafo 6 discute il ruolo che l'intelligenza artificiale può assumere nelle diverse fasi del processo analitico. Il paragrafo 7 traduce i risultati in implicazioni per le decisioni di marketing strategico e operativo. Il paragrafo 8, infine, presenta i principali limiti del lavoro e le possibili direttrici di ricerca futura.

2. Quadro di riferimento

La letteratura sulle recensioni online si sviluppa inizialmente nell'ambito degli studi sul word of mouth elettronico, concentrandosi soprattutto sul valore informativo delle valutazioni espresse dai consumatori e sui loro effetti sul mercato. I primi contributi mostrano che le recensioni influenzano in misura significativa le vendite e le scelte dei consumatori (Godes e Mayzlin, 2004; Chevalier e Mayzlin, 2006), mentre Dellarocas (2003) interpreta i sistemi di reputazione digitale come vere e proprie istituzioni economiche. Hennig-Thurau et al. (2004) spostano l'attenzione anche sulle motivazioni che inducono il consumatore a pubblicare una recensione, mettendo in evidenza come essa non costituisca soltanto un segnale informativo rivolto al mercato, ma anche un atto sociale attraverso il quale il cliente rende pubblica la propria esperienza con l'impresa.

Un secondo insieme di studi mostra che l'informazione contenuta nelle recensioni non è adeguatamente rappresentata dalla sola valutazione media. Sun (2012) evidenzia che la varianza delle valutazioni produce effetti sulle vendite indipendenti dalla media e che tali effetti dipendono dalle caratteristiche del prodotto; Hu, Pavlou e Zhang (2006) documentano

la frequente bimodalità delle distribuzioni, riconducendola anche a meccanismi di autoselezione nella decisione di recensire; Moe e Trusov (2011) mostrano come le interazioni sociali tra recensori contribuiscano alla formazione di tali distribuzioni nel tempo. Nel loro insieme, questi contributi suggeriscono che la forma della distribuzione (consenso, dispersione, polarizzazione) contenga informazioni che la media tende a comprimere. È su questa base che si collocano gli indicatori di struttura valutativa utilizzati nel presente lavoro: non come nuove misure statistiche, ma come strumenti per tradurre proprietà già note delle distribuzioni in segnali più immediatamente utilizzabili nella diagnosi manageriale.

Anche le caratteristiche della singola recensione possono fornire informazioni ulteriori rispetto al rating. Mudambi e Schuff (2010) mostrano, ad esempio, che la relazione tra estremità della valutazione e utilità percepita non è lineare e varia in funzione della natura del prodotto, mentre Filieri (2015) evidenzia il ruolo della qualità informativa e della credibilità della recensione. Questa letteratura è rilevante per il presente lavoro soprattutto perché richiama l'attenzione sulla dipendenza dal contesto delle caratteristiche testuali, ma lunghezza, articolazione e contenuto non assumono necessariamente lo stesso significato in piattaforme, mercati e categorie differenti. Ne discende l'opportunità di evitare soglie o regole interpretative trasferite automaticamente da un contesto all'altro.

Il passaggio più rilevante ai fini della tesi qui proposta avviene quando l'attenzione si sposta dal comportamento di chi recensisce a quello dell'impresa che decide se e come rispondere. La letteratura sul service failure e sul service recovery aveva già mostrato che la valutazione dell'esperienza dipende non soltanto dalla prestazione originaria, ma anche dal modo in cui l'impresa gestisce l'eccezione e il reclamo (Bitner, Booms e Tetreault, 1990; Smith, Bolton e Wagner, 1999; Homburg e Fürst, 2005). Con la diffusione delle piattaforme digitali, la risposta al cliente diventa però pubblica, e quindi non costituisce più soltanto un intervento di recupero rivolto al singolo consumatore, ma anche un comportamento osservabile dagli altri clienti e dal mercato.

Da questa prospettiva si sviluppa la letteratura sulle managerial responses e sulla webcare. In tal senso, Van Noort e Willemsen (2012) analizzano gli effetti delle risposte aziendali nei contesti digitali, mentre Ma, Sun e Kekre (2015) mostrano che l'intervento pubblico dell'impresa può produrre effetti differenti in funzione delle caratteristiche del messaggio del cliente. Proserpio e Zervas (2017) documentano inoltre come la scelta di rispondere alle recensioni incida sulla successiva reputazione online dell'impresa. Gli studi

successivi approfondiscono la natura selettiva della risposta: Wang et al. (2020) analizzano a quali recensioni convenga destinare risorse limitate; Lopes et al. (2023), attraverso una revisione sistematica della letteratura sulla webcare, mostrano l'importanza della valenza della recensione, della tempestività e della personalizzazione della risposta; Wu, Ye e Chu (2026) distinguono esplicitamente gli effetti dei tassi di risposta alle recensioni positive e negative. La domanda non è quindi più soltanto se l'impresa risponda, ma a chi scelga di rispondere, con quale tempestività e secondo quale logica.

È precisamente in questo passaggio che la gestione delle recensioni assume una dimensione relazionale. La teoria del relationship marketing attribuisce a fiducia, impegno e reciprocità un ruolo centrale nella costruzione di rapporti durevoli tra impresa e cliente (Morgan e Hunt, 1994). In modo complementare, la nozione di contratto psicologico, sviluppata originariamente in ambito organizzativo (Rousseau, 1989) e successivamente estesa alla relazione impresa-consumatore (Hannah et al., 2016), consente di interpretare alcune manifestazioni di insoddisfazione non soltanto come giudizi negativi sulla singola prestazione, ma come reazioni alla percezione che una promessa o un'obbligazione implicita della relazione sia stata disattesa. In questa prospettiva, anche il comportamento dell'impresa successivo alla recensione — rispondere o tacere, intervenire rapidamente o tardivamente, privilegiare chi protesta o riconoscere anche chi esprime apprezzamento — diventa un segnale osservabile del modo in cui l'impresa gestisce la relazione.

Un ulteriore problema riguarda la qualità dei dati sui quali tali misure vengono calcolate. La letteratura sull'integrità dei dataset di recensioni ha sviluppato sia approcci fondati su segnali comportamentali (Jindal e Liu, 2008), sia metodi basati sull'analisi linguistica del testo (Ott et al., 2011). La questione assume oggi nuova rilevanza con la disponibilità dei modelli linguistici di grandi dimensioni. La ricerca più recente sottolinea infatti le potenzialità degli LLM come strumenti di supporto alla ricerca di marketing e, al tempo stesso, l'opportunità di integrarli con il giudizio e il controllo del ricercatore anziché considerarli sostituti autonomi del processo analitico (Arora, Chakraborty e Nishimura, 2025). Per l'analisi delle recensioni ciò apre uno spazio importante per l'intelligenza artificiale, ma pone anche il problema di distinguere le attività nelle quali il modello può assistere l'analista da quelle nelle quali la misura deve restare direttamente verificabile e riproducibile.

Il presente lavoro si colloca all'intersezione di questi contributi, ma ne sposta il punto di osservazione. La letteratura ha ampiamente studiato gli effetti delle recensioni e delle risposte

aziendali: sulle vendite, sulla reputazione, sulle valutazioni successive e sul comportamento dei consumatori. Molto meno centrale è stata la costruzione di un sistema parsimonioso di indicatori finalizzato non tanto a spiegare gli esiti di mercato, quanto a diagnosticare la configurazione osservabile della relazione tra impresa e recensori a partire da dati pubblicamente disponibili. La tesi proposta è che una parte significativa di questa diagnosi possa essere condotta prima di qualunque interpretazione semantica, attraverso un numero limitato di misure semplici e riproducibili.

Il contributo è dunque prevalentemente di natura metodologica e manageriale. Gli indicatori proposti non intendono sostituire i costrutti e le tecniche sviluppati dalla letteratura precedente, ma organizzarne alcune intuizioni in un sistema di lettura orientato alla decisione: distinguere consenso e polarizzazione, osservare se l'impresa risponde e a chi risponde, misurarne la tempestività e individuare le porzioni del corpus nelle quali un'analisi più approfondita può produrre maggiore informazione. In questa architettura anche il ruolo dell'intelligenza artificiale risulta meglio definito: non come fonte autonoma della misura, ma come strumento utilizzabile per verificarne le condizioni di validità, approfondire i fenomeni segnalati dagli indicatori e supportare l'esecuzione operativa delle decisioni che ne derivano. In questa prospettiva, la crescente capacità di elaborazione offerta dall'intelligenza artificiale non riduce il ruolo del ricercatore, ma ne ridefinisce e, per certi versi, ne rafforza la centralità. Se la macchina può assumere una quota crescente del lavoro di lettura, classificazione ed esecuzione, resta infatti all'analista il compito non delegabile di formulare le domande, scegliere le dimensioni rilevanti, costruire le misure e interpretarne le implicazioni. Quanto più aumenta la capacità di elaborare i dati, tanto più il valore dell'analisi dipende dalla capacità di decidere che cosa valga la pena misurare. È in questo spazio che si colloca la proposta sviluppata nel presente lavoro.

3. Il sistema di indici

3.1. La media non basta

La valutazione media è l'indicatore attorno al quale è organizzata gran parte dell'esperienza pubblica delle recensioni online: è il numero mostrato accanto al nome della marca o del prodotto, quello utilizzato per costruire classifiche e confronti e, spesso, quello che entra nei cruscotti direzionali. La ragione della sua diffusione è evidente, in quanto riesce a concentrare in un unico valore un insieme anche molto ampio di giudizi ed è immediatamente comprensibile. Proprio questa capacità di sintesi costituisce, tuttavia, anche il suo principale limite.

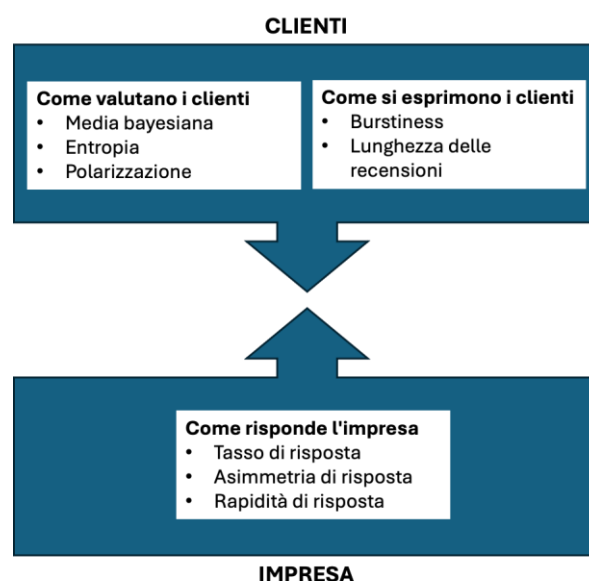
Due distribuzioni profondamente diverse possono infatti produrre la stessa media. Quattro valutazioni da cinque stelle e una da una stella restituiscono una media di 4,2; lo stesso valore si ottiene con quattro valutazioni da quattro stelle e una da cinque. Dal punto di vista manageriale, le due situazioni raccontano storie differenti. Nel primo caso il servizio soddisfa pienamente la maggioranza dei clienti ma produce, per una minoranza, un'esperienza fortemente negativa: il problema consiste allora nell'individuare le condizioni nelle quali si verifica il fallimento e intervenire sulle eccezioni. Nel secondo caso le esperienze sono invece omogeneamente positive e non emerge un analogo segnale di criticità. La media è la stessa, ma le decisioni che ne dovrebbero derivare sono diverse.

Il problema non è quindi che la media sia un indicatore errato, ma che risponde a una domanda soltanto: qual è il livello medio delle valutazioni? Non dice quanto i clienti concordino tra loro, se i giudizi siano concentrati o polarizzati, né quale struttura si nasconda dietro lo stesso valore sintetico. Per rispondere a queste domande è necessario osservare anche la distribuzione delle valutazioni.

A questa perdita di informazione si aggiunge un secondo problema. Quando il numero di recensioni è ridotto, poche osservazioni possono modificare sensibilmente la media e rendere poco affidabile il confronto tra marche o prodotti caratterizzati da numerosità molto diverse. Anche in questo caso non si tratta di rinunciare alla media, ma di correggerla e affiancarla ad altre misure capaci di descrivere ciò che essa, per costruzione, non può mostrare.

Esiste poi una dimensione che la distribuzione delle valutazioni, da sola, non può cogliere. Una recensione non è soltanto un giudizio espresso dal cliente, ma una volta pubblicata, può diventare anche l'inizio di un'interazione pubblica con l'impresa. L'azienda può rispondere o non rispondere, può farlo rapidamente o dopo molti giorni, può concentrare la propria attenzione sui clienti insoddisfatti oppure rivolgersi anche a quelli soddisfatti. Questi comportamenti non descrivono più soltanto come il cliente valuta l'impresa, ma anche come l'impresa gestisce la relazione con chi ha scelto di esprimere pubblicamente quella valutazione. È in questo passaggio dalla valutazione all'interazione che si colloca la dimensione relazionale proposta nel presente lavoro.

Il sistema di indicatori viene pertanto articolato in tre gruppi complementari. Il primo riguarda come valutano i clienti e descrive il livello, il grado di concordanza e l'eventuale polarizzazione dei giudizi, consentendo di andare oltre il solo valore medio. Il secondo osserva come si esprimono i clienti, considerando la distribuzione temporale delle recensioni e la quantità di contenuto testuale associata ai diversi livelli di valutazione. Il terzo riguarda come risponde l'impresa e rende osservabile il suo comportamento nello spazio pubblico della recensione, misurandone la presenza, la selettività e la tempestività della risposta. Gli indicatori sono costruiti a partire da un numero limitato di informazioni direttamente osservabili (valutazione numerica, marca temporale, lunghezza del testo, presenza ed eventuale data della risposta aziendale) senza richiedere, in questa prima fase, alcuna interpretazione semantica del contenuto delle recensioni.



3.2. Come valutano i clienti

3.2.1. Quanto è affidabile il confronto? La valutazione corretta per la numerosità

La domanda manageriale.

Quando si confrontano marche, prodotti o punti vendita attraverso le recensioni online, una differenza spesso trascurata riguarda il numero di giudizi sui quali è calcolata la valutazione media. Una marca con poche decine di recensioni e una con diverse migliaia possono presentare lo stesso punteggio, ma quel valore non ha necessariamente la stessa solidità. La domanda rilevante non è quindi soltanto “qual è la valutazione media?”, ma anche “quanto possiamo considerarla affidabile ai fini del confronto?”.

Il limite della media tradizionale.

La media aritmetica attribuisce lo stesso significato a un punteggio indipendentemente dal numero di osservazioni che lo ha generato. Una valutazione di 4,8 ottenuta su venti recensioni viene così trattata allo stesso modo di una valutazione di 4,8 costruita su duemila recensioni, nonostante nel primo caso pochi nuovi giudizi possano modificarla sensibilmente. Il problema diventa particolarmente rilevante quando si confrontano imprese caratterizzate da volumi di recensioni molto differenti: le posizioni apparentemente migliori possono dipendere, almeno in parte, dalla maggiore instabilità delle medie calcolate su campioni ridotti.

L'indicatore proposto.

Per ridurre questo effetto viene utilizzata una valutazione media corretta per la numerosità. L'idea è intuitiva: quando le recensioni disponibili sono poche, il punteggio osservato viene prudenzialmente avvicinato alla valutazione media dell'intero insieme considerato; man mano che il numero delle recensioni aumenta, il valore della singola marca pesa sempre di più e la correzione diventa progressivamente irrilevante. L'indicatore non sostituisce quindi il giudizio espresso dai clienti, ma tiene conto della diversa quantità di evidenza sulla quale esso si basa.

L'utilità per la decisione.

La correzione rende più prudente e comparabile il benchmarking tra marche, prodotti o concorrenti con numerosità differenti. È particolarmente utile nella costruzione di ranking, cruscotti competitivi e sistemi di monitoraggio, perché riduce il rischio di attribuire eccessivo significato a valutazioni molto elevate — o molto basse — fondate su un numero limitato di osservazioni. Non indica se i clienti siano concordi o polarizzati, aspetto affrontato dagli indicatori successivi; consente però di stabilire una base più solida dalla quale partire per il confronto.

Nota tecnica.

La misura adottata è una media corretta verso il valore medio del dataset (c.d. media bayesiana):

$$RB = (C \cdot m + \sum r_i) / (C + n)$$

Dove r_i è la valutazione della recensione i -esima, n il numero di recensioni della marca, m la media dell'intero dataset e C una costante di smorzamento, interpretabile come un numero di osservazioni aggiuntive collocate al valore medio m . All'aumentare di n , il peso della correzione diminuisce e l'indicatore converge verso la media osservata della marca. Il valore di C deve essere esplicitato; nelle elaborazioni presentate nel lavoro è posto pari a 20.

3.2.2. Quanto sono concordi i clienti?

La domanda manageriale.

Conoscere la valutazione media di una marca non basta per capire quanto quell'opinione sia condivisa. Due marche possono presentare lo stesso punteggio medio, ma trovarsi in condizioni molto differenti: nel primo caso quasi tutti i clienti possono esprimere giudizi simili; nel secondo, lo stesso valore può essere il risultato di esperienze molto diverse tra loro. Per chi deve decidere, la domanda diventa quindi: quanto sono concordi i clienti nel valutare la marca?

Il limite della valutazione media.

La media sintetizza il livello complessivo dei giudizi, ma non ne rappresenta la dispersione. Un valore medio elevato può nascondere un consenso molto compatto oppure la presenza di gruppi di clienti con esperienze differenti. Questa distinzione è rilevante perché le

due situazioni richiedono letture manageriali diverse. Un giudizio ampiamente condiviso segnala un'esperienza relativamente uniforme; un elevato disaccordo suggerisce invece che esistano condizioni, segmenti di clientela o modalità di erogazione che producono esperienze differenti e che meritano quindi di essere approfondite.

L'indicatore proposto.

Per misurare il grado di accordo tra i recensori si utilizza l'entropia della distribuzione delle valutazioni. L'idea è semplice: quanto più i giudizi sono concentrati su uno o pochi livelli della scala, tanto maggiore è il consenso; quanto più sono distribuiti tra valutazioni differenti, tanto maggiore è il disaccordo. L'indicatore aggiunge quindi alla domanda "quanto siamo valutati?" una seconda informazione: "quanto i clienti sono d'accordo tra loro nel valutarci in quel modo?"

L'utilità per la decisione.

Il grado di accordo aiuta anzitutto a stabilire dove sia più utile approfondire l'analisi. Un'elevata concentrazione dei giudizi indica che le recensioni tendono a raccontare esperienze simili e che ulteriori analisi possono produrre rendimenti informativi decrescenti. Un elevato disaccordo segnala invece la presenza di una maggiore varietà di esperienze; diventa allora utile interrogare il testo delle recensioni per comprendere quali fattori distinguano i clienti soddisfatti da quelli meno soddisfatti, oppure verificare se le differenze siano associate a prodotti, canali, momenti o segmenti specifici.

L'indicatore assume quindi anche una funzione di orientamento dell'analisi successiva: non dice ancora perché i clienti siano in disaccordo, ma segnala dove valga la pena cercarne le ragioni. Né consente, da solo, di stabilire quale forma assuma il disaccordo. Valutazioni distribuite diffusamente lungo tutta la scala e giudizi concentrati soprattutto sui due estremi possono infatti produrre livelli elevati di entropia pur descrivendo fenomeni differenti. È questa la distinzione affrontata dall'indicatore successivo.

Nota tecnica.

Il grado di disaccordo viene misurato attraverso l'entropia di Shannon (1948), applicata alla distribuzione delle valutazioni sui cinque livelli:

$$H = -\sum p_k \log_2 p_k, \text{ con } p_k = n_k/n$$

dove p_k rappresenta la quota di recensioni collocata nel livello k . Su una scala a cinque livelli, l'indicatore varia da 0, quando tutte le recensioni esprimono la medesima valutazione, a $\log_2(5)$, pari a circa 2,32, quando le recensioni sono distribuite uniformemente tra tutti i livelli.

Valori crescenti indicano quindi una crescente eterogeneità dei giudizi, senza specificare quale forma assuma tale eterogeneità.

3.2.3. I giudizi sono moderati o polarizzati?

La domanda manageriale.

Sapere che i clienti sono in disaccordo costituisce un primo segnale, ma non è sufficiente per comprendere la natura del fenomeno. Il disaccordo può infatti assumere forme molto differenti. I giudizi possono distribuirsi gradualmente lungo la scala oppure concentrarsi soprattutto agli estremi: clienti fortemente soddisfatti da un lato e fortemente insoddisfatti dall'altro. Per il manager la domanda diventa quindi: le opinioni dei clienti si distribuiscono lungo diverse gradazioni oppure tendono a polarizzarsi tra entusiasmo e forte insoddisfazione?

Il limite dell'indicatore precedente.

L'entropia misura quanto le valutazioni siano disperse, ma non indica dove tale dispersione si concentri. Due distribuzioni possono presentare un grado simile di disaccordo e avere, tuttavia, significati manageriali molto diversi. Una può essere caratterizzata da giudizi distribuiti soprattutto nelle posizioni intermedie; un'altra può essere dominata dalle valutazioni estreme. Nel secondo caso il problema non è semplicemente che i clienti abbiano opinioni diverse: è che la stessa marca o lo stesso prodotto sembra generare esperienze fortemente contrapposte.

L'indicatore proposto.

Per distinguere queste situazioni si misura il peso relativo delle valutazioni estreme, cioè una e cinque stelle, rispetto a quelle intermedie. Quanto maggiore è la concentrazione agli estremi, tanto più il profilo delle recensioni assume una configurazione polarizzata. L'indicatore consente così di distinguere un generico disaccordo da una situazione nella quale chi decide di recensire tende soprattutto a esprimere entusiasmo oppure forte insoddisfazione.

L'utilità per la decisione.

Una polarizzazione elevata costituisce un segnale particolarmente interessante perché suggerisce di andare oltre il dato medio e cercare che cosa separi esperienze così differenti. La

stessa offerta potrebbe, ad esempio, funzionare molto bene in alcune circostanze e molto male in altre; soddisfare fortemente alcuni segmenti e deluderne altri; oppure produrre risultati differenti in funzione del canale, del prodotto o della fase del customer journey. L'indicatore non identifica direttamente quale di queste spiegazioni sia corretta, ma segnala una configurazione che merita di essere approfondita attraverso l'analisi delle recensioni e, ove disponibili, di altre informazioni sul cliente e sull'esperienza.

Anche in questo caso il valore manageriale dell'indicatore è quindi soprattutto diagnostico. Esso non spiega la causa della polarizzazione, ma rende visibile un fenomeno che la media può occultare e che la sola misura del disaccordo non riesce a qualificare. Media, entropia e polarizzazione rispondono così a tre domande diverse e complementari: qual è il giudizio complessivo? quanto è condiviso? e quale forma assume l'eventuale disaccordo?

Nota tecnica.

La polarizzazione viene misurata attraverso il rapporto tra la quota di valutazioni collocate agli estremi della scala e quella delle valutazioni intermedie:

$$B = (p_1 + p_5)/(p_2 + p_3 + p_4)$$

dove $p_1 \dots p_5$ rappresentano le quote di recensioni associate ai cinque livelli della scala. Un valore superiore a 1 indica che, complessivamente, le valutazioni estreme sono più numerose di quelle intermedie; valori crescenti segnalano una concentrazione progressivamente maggiore agli estremi. L'indicatore va tuttavia interpretato insieme alla distribuzione delle valutazioni, poiché non distingue, da solo, una concentrazione su entrambi gli estremi da una distribuzione fortemente sbilanciata verso uno soltanto di essi.

3.3. Come si esprimono i clienti

Le valutazioni considerate nel paragrafo precedente descrivono come i clienti giudicano la marca, ma le recensioni costituiscono anche un flusso di informazioni che si sviluppa nel tempo e assume caratteristiche diverse a seconda del contesto. Può essere regolare oppure concentrarsi in particolari momenti; può inoltre accadere che le recensioni più articolate si concentrino in corrispondenza di determinati livelli di valutazione. Osservare queste caratteristiche non consente ancora di comprendere il contenuto delle recensioni, ma aiuta a stabilire quando e dove il flusso meriti un approfondimento.

3.3.1. Le recensioni arrivano regolarmente o per picchi?

La domanda manageriale.

Il numero complessivo delle recensioni dice quanto i clienti si esprimano, ma non dice come tale attività si distribuisca nel tempo. Lo stesso numero di recensioni può infatti essere prodotto attraverso un flusso relativamente costante oppure concentrarsi in brevi periodi di intensa attività. Per il manager la domanda rilevante diventa quindi: le recensioni arrivano con una certa regolarità oppure si concentrano in picchi anomali?

Il limite degli indicatori tradizionali.

Il numero totale delle recensioni, o anche la loro media mensile, tende a nascondere la struttura temporale del fenomeno. Cento recensioni distribuite uniformemente nell'arco di un anno e cento recensioni concentrate prevalentemente in pochi giorni producono lo stesso volume complessivo, ma possono avere significati molto diversi. Un improvviso addensamento può coincidere con un episodio critico, con il lancio di un prodotto, con una campagna che stimola la raccolta di recensioni o con altri eventi che modificano temporaneamente la propensione dei clienti a pubblicare. Il dato aggregato non permette di distinguere queste configurazioni.

L'indicatore proposto.

Per descrivere la regolarità del flusso si osservano gli intervalli di tempo che separano una recensione dalla successiva. Quando tali intervalli sono relativamente simili, il flusso è regolare; quando invece si alternano periodi di scarsa attività e sequenze ravvicinate di pubblicazioni, le recensioni tendono ad arrivare "a raffiche". L'indicatore di burstiness sintetizza proprio questa caratteristica, consentendo di distinguere un flusso continuo da uno caratterizzato da addensamenti temporali.

L'utilità per la decisione.

Un valore elevato di burstiness costituisce soprattutto un segnale di attenzione. Non permette, da solo, di stabilire perché si sia verificato un picco, ma indica quando sia opportuno approfondirne l'origine. L'analista può quindi concentrare la lettura delle recensioni nei periodi interessati, verificare la presenza di eventi aziendali o di mercato e valutare se l'addensamento corrisponda, ad esempio, a una crisi, a un'iniziativa commerciale o a modalità anomale di generazione delle recensioni.

La funzione dell'indicatore è dunque diagnostica: non attribuisce una causa al fenomeno, ma individua i momenti nei quali il comportamento del flusso si discosta dalla sua normale regolarità e nei quali un'analisi più approfondita può produrre maggiore informazione.

Nota tecnica.

La regolarità temporale viene misurata attraverso l'indice di burstiness, calcolato sugli intervalli τ tra recensioni consecutive:

$$\beta = (\sigma_t - \mu_t) / (\sigma_t + \mu_t)$$

dove μ_t e σ_t rappresentano rispettivamente la media e la deviazione standard degli intervalli tra una recensione e la successiva. L'indice assume valori compresi tra -1 e $+1$. Valori prossimi a -1 indicano intervalli molto regolari; valori prossimi a zero sono compatibili con un flusso casuale di tipo poissoniano; valori crescenti verso $+1$ segnalano una maggiore concentrazione delle recensioni in raffiche. Il confronto tra marche richiede tuttavia cautela quando le finestre temporali di osservazione non sono omogenee; in tali condizioni l'indicatore è più affidabile per analizzare l'evoluzione della stessa marca nel tempo che per effettuare confronti diretti tra imprese.

3.3.2. Dove si concentra il contenuto informativo?

La domanda manageriale.

Una volta stabilito che non tutte le recensioni raccontano necessariamente la stessa esperienza, resta un problema pratico: da dove conviene iniziare a leggere? In un corpus composto da migliaia o centinaia di migliaia di testi, dedicare la stessa attenzione a tutte le recensioni può risultare poco efficiente. È quindi utile capire se i contributi più articolati si concentrino prevalentemente tra i clienti molto insoddisfatti, tra quelli soddisfatti oppure nelle valutazioni intermedie.

Il limite degli indicatori tradizionali.

La distribuzione dei rating indica quanti clienti assegnano una, due o cinque stelle, ma non dice quanto contenuto testuale producano nei diversi livelli della scala. Una categoria numericamente poco rappresentata può raccogliere recensioni molto più articolate e contenere quindi una quantità di materiale utile all'analisi superiore a quella suggerita dalla sua semplice numerosità. Allo stesso modo, una quota molto ampia di valutazioni può essere costituita da testi estremamente brevi e ripetitivi.

L'indicatore proposto.

Per rendere visibile questa differenza viene costruito un profilo della lunghezza delle recensioni per ciascun livello di valutazione. Non si tratta, in questo caso, di un singolo indice sintetico, ma di una sequenza di valori che mostra quanto siano mediamente articolati i testi

associati a una, due, tre, quattro e cinque stelle. La lunghezza non misura direttamente la qualità o la rilevanza del contenuto, ma costituisce un indicatore elementare della densità testuale disponibile nei diversi segmenti della distribuzione.

L'utilità per la decisione.

Il profilo consente di orientare l'analisi successiva verso le aree nelle quali si concentra il materiale testuale più ricco. Se, ad esempio, le recensioni negative risultano significativamente più lunghe delle positive, può essere ragionevole iniziare da esse per comprendere in dettaglio le cause dell'insoddisfazione; se invece i testi più articolati si concentrano nelle valutazioni intermedie, proprio quella parte apparentemente meno "estrema" della distribuzione può rivelarsi più utile per individuare punti di forza e criticità.

L'indicatore non stabilisce quindi quali recensioni siano più importanti, né può sostituire la loro lettura. Serve piuttosto a decidere dove concentrare per prima l'attenzione analitica. In questo senso costituisce un naturale punto di raccordo con l'impiego dell'intelligenza artificiale discusso nel seguito del lavoro: gli indicatori strutturati individuano le porzioni del corpus potenzialmente più promettenti e l'analisi semantica può essere successivamente indirizzata su di esse.

Un'ulteriore cautela riguarda la trasferibilità dei risultati. La relazione tra lunghezza del testo e valutazione può dipendere dalla piattaforma, dalla categoria merceologica e dalle modalità con cui la recensione viene sollecitata e pubblicata. Il profilo va quindi costruito sul contesto specifico e non trasformato in una regola generale valida per mercati o piattaforme differenti.

Nota tecnica.

Per ciascun livello della scala viene calcolata la lunghezza mediana delle recensioni, ottenendo il vettore:

$$L = (L_1, L_2, L_3, L_4, L_5)$$

Dove L_k rappresenta la lunghezza mediana dei testi associati alla valutazione k . L'impiego della mediana riduce l'influenza di recensioni eccezionalmente lunghe e rende il confronto più robusto rispetto alla media aritmetica. Il risultato non costituisce un indice sintetico, ma un profilo da interpretare all'interno del singolo dataset e del relativo contesto di piattaforma.

3.4. Come risponde l'impresa

Fin qui l'analisi ha riguardato ciò che i clienti producono nello spazio pubblico delle recensioni, ovvero come valutano la marca, quanto concordano tra loro, come si distribuiscono i giudizi e con quale andamento temporale vengono pubblicati. La presenza di una risposta aziendale introduce una prospettiva diversa. La recensione non è più soltanto una manifestazione dell'esperienza del cliente, ma può diventare l'avvio di un'interazione pubblica nella quale anche l'impresa sceglie se intervenire, a chi rivolgersi e con quale tempestività. È attraverso questi comportamenti osservabili che il sistema proposto affronta più direttamente la dimensione relazionale.

3.4.1. A quante recensioni risponde?

La domanda manageriale.

La prima domanda è la più semplice: quanto è presente l'impresa nello spazio pubblico delle recensioni? Due marche possono ricevere un numero simile di giudizi ma adottare comportamenti molto diversi: una può limitarsi a osservare ciò che viene pubblicato, mentre l'altra può rispondere sistematicamente ai clienti che hanno scelto di esprimersi. Il tasso di risposta fornisce una prima misura di questa presenza.

Il limite della sola osservazione delle valutazioni.

Rating, distribuzioni e volumi di recensioni descrivono esclusivamente il comportamento dei clienti. Non permettono di sapere se l'impresa consideri la piattaforma come un semplice luogo di raccolta del feedback oppure come uno spazio nel quale intervenire attivamente. Anche un numero elevato di recensioni, quindi, può convivere con una sostanziale assenza dell'impresa dalla conversazione pubblica.

L'indicatore proposto.

Il tasso di risposta misura la quota di recensioni alle quali l'impresa fornisce un riscontro pubblico. Si tratta dell'indicatore relazionale più elementare del sistema: non dice ancora a chi venga data risposta né quanto rapidamente, ma rende osservabile il grado di copertura della politica di risposta. Un valore elevato indica una presenza diffusa; un valore basso segnala invece che

l'intervento dell'impresa è selettivo o occasionale; un valore nullo identifica le imprese che non utilizzano la risposta pubblica come modalità di interazione sulla piattaforma considerata.

L'utilità per la decisione.

Il tasso di risposta consente di confrontare la propria presenza con quella dei concorrenti e di trasformare una generica intenzione di “presidiare le recensioni” in un obiettivo misurabile. Può inoltre essere monitorato nel tempo per verificare se l'introduzione di nuove procedure organizzative o strumenti di automazione produca un effettivo aumento della copertura.

Da solo, tuttavia, l'indicatore può essere fuorviante. Due imprese che rispondono alla stessa percentuale di recensioni possono adottare strategie profondamente diverse se una concentra l'attenzione sui clienti insoddisfatti e l'altra distribuisce le risposte indipendentemente dal segno della valutazione. Per comprendere quale logica governi la risposta è quindi necessario scomporre il tasso complessivo.

Nota tecnica.

Il tasso complessivo di risposta è definito come:

$$\rho = \#\{i : a_i \neq \emptyset\}/n$$

dove $a_i \neq \emptyset$ indica la presenza di una risposta aziendale alla recensione i -esima e n il numero complessivo di recensioni per le quali il campo relativo alla risposta è disponibile. L'indicatore varia quindi tra 0, in assenza di qualunque risposta, e 1, quando tutte le recensioni ricevono un riscontro.

3.4.2. A chi sceglie di rispondere?

La domanda manageriale.

Sapere quante recensioni ricevono una risposta non dice ancora come l'impresa distribuisca la propria attenzione. Un'azienda può intervenire soprattutto quando il cliente manifesta forte insoddisfazione, può rispondere sistematicamente indipendentemente dal giudizio espresso oppure, al contrario, può essere più presente nei confronti di chi formula valutazioni positive. La domanda rilevante diventa quindi: a chi scegliamo di rispondere?

Il limite del tasso complessivo di risposta.

Il tasso aggregato può nascondere strategie relazionali molto differenti. Due imprese possono, ad esempio, rispondere entrambe alla metà delle recensioni, ma la prima può concentrare quasi tutti i propri interventi sulle valutazioni negative, mentre la seconda può distribuire le risposte in

maniera equilibrata tra feedback positivi e negativi. Dal punto di vista quantitativo il tasso complessivo è identico; dal punto di vista manageriale, il comportamento osservato è profondamente diverso.

L'indicatore proposto.

Per rendere visibile questa differenza si calcolano separatamente il tasso di risposta alle recensioni negative e quello alle recensioni positive e se ne considera la differenza. L'asimmetria di risposta indica quindi verso quale parte della distribuzione l'impresa concentra maggiormente la propria attenzione.

Valori fortemente positivi sono coerenti con una strategia di contenimento, nella quale la risposta viene utilizzata soprattutto nei confronti di chi manifesta insoddisfazione. Valori prossimi allo zero, se associati a un tasso complessivo elevato, descrivono una strategia di presenza, nella quale l'impresa interviene indipendentemente dal segno della valutazione. Valori fortemente negativi sono invece coerenti con una strategia di vetrina, caratterizzata da una maggiore attenzione verso i giudizi positivi.

Queste etichette descrivono configurazioni osservabili e non intendono attribuire automaticamente all'impresa una specifica intenzione strategica: la causa del comportamento rilevato richiede informazioni ulteriori.

L'utilità per la decisione.

L'asimmetria rende il tasso di risposta una vera variabile diagnostica. Un'impresa può così verificare non soltanto quanto stia rispondendo, ma se la distribuzione delle risorse dedicate alle recensioni sia coerente con il ruolo che intende attribuire alla relazione pubblica con i clienti. Una forte concentrazione sulle recensioni negative può essere appropriata quando la risposta è concepita prevalentemente come strumento di service recovery; può invece segnalare un'opportunità non colta se l'obiettivo è riconoscere e alimentare anche il contributo dei clienti soddisfatti.

Il valore dell'indicatore è quindi soprattutto comparativo: consente di osservare differenze tra imprese che un semplice tasso aggregato non mostra e di verificare nel tempo se una modifica della politica di risposta produca il riequilibrio desiderato.

Nota tecnica.

Si definiscono anzitutto i tassi di risposta condizionati al segno della valutazione:

$$p_{neg} = \#\{i : r_i \leq 2 \wedge a_i \neq \emptyset\} / \#\{i : r_i \leq 2\}$$

$$p_{pos} = \#\{i : r_i \geq 4 \wedge a_i \neq \emptyset\} / \#\{i : r_i \geq 4\}$$

dove r_i rappresenta la valutazione associata alla recensione e $a_i \neq \emptyset$ la presenza di una risposta aziendale. Le valutazioni pari a 1 o 2 sono considerate negative, quelle pari a 4 o 5 positive; le valutazioni intermedie non entrano nel calcolo dell'asimmetria.

L'indice è quindi definito come:

$$\alpha = \rho_{neg} - \rho_{pos}$$

Valori positivi indicano una maggiore propensione a rispondere alle recensioni negative, valori negativi una maggiore propensione a rispondere alle positive, valori prossimi allo zero un comportamento sostanzialmente simmetrico.

Poiché α è una differenza tra due tassi, un'impresa che non risponde ad alcuna recensione produrrebbe formalmente lo stesso valore di un'impresa che risponde alle recensioni positive e negative nella stessa proporzione. Per questa ragione l'asimmetria deve essere interpretata congiuntamente al tasso complessivo di risposta; le imprese totalmente non rispondenti costituiscono una categoria distinta.

3.4.3 Quanto rapidamente risponde?

La domanda manageriale.

La presenza dell'impresa non dipende soltanto dal fatto che una risposta venga fornita, ma anche dal momento in cui essa arriva. Una risposta pubblicata poche ore dopo una recensione negativa e una pubblicata diversi giorni più tardi costituiscono formalmente due risposte, ma possono assumere un significato molto diverso per il cliente. Alla domanda "rispondiamo?" si aggiunge quindi una terza questione: "quanto rapidamente lo facciamo?"

Il limite del tasso di risposta.

Il tasso di risposta considera equivalenti tutti i riscontri, indipendentemente dal tempo trascorso dalla pubblicazione della recensione. Un'impresa può quindi presentare una copertura molto elevata e apparire particolarmente presente, pur rispondendo sistematicamente con forte ritardo. In questo caso l'indicatore di copertura sopravvaluterebbe la capacità dell'impresa di intervenire tempestivamente nella relazione con il cliente.

L'indicatore proposto.

La latenza di risposta misura il tempo che intercorre tra la pubblicazione della recensione e il riscontro dell'impresa. Non interessa soltanto sapere se l'azienda interviene, ma quanto rapidamente riesca a trasformare il feedback pubblico in un'occasione di interazione.

La tempestività assume particolare rilievo nel caso delle recensioni negative: quanto più la risposta è vicina all'episodio che ha generato l'insoddisfazione, tanto maggiori sono, almeno potenzialmente, le possibilità che essa svolga una funzione di recupero. Una risposta molto tardiva può continuare ad avere valore reputazionale nei confronti di chi leggerà successivamente la recensione, ma riduce la possibilità di intervenire tempestivamente sull'esperienza del cliente che l'ha pubblicata.

L'utilità per la decisione.

La latenza presenta un vantaggio manageriale particolarmente immediato: può essere tradotta direttamente in un obiettivo operativo. L'indicazione generica di "rispondere più rapidamente alle recensioni" può diventare, ad esempio, un tempo mediano obiettivo o uno standard differenziato per tipologia di feedback.

Considerata congiuntamente all'asimmetria, la latenza consente inoltre di distinguere configurazioni diverse. Una risposta rapida e distribuita tra feedback positivi e negativi segnala una presenza relazionale diffusa; una risposta rapida concentrata sui giudizi negativi è coerente con una gestione orientata al contenimento e al service recovery; una risposta sistematicamente tardiva segnala invece che l'esistenza formale del riscontro non coincide necessariamente con una reale capacità di intervento tempestivo.

Nota tecnica.

Quando la piattaforma rende disponibile anche la data della risposta, la latenza viene calcolata come differenza tra il momento del riscontro aziendale e quello della pubblicazione della recensione:

$$\Delta t = \text{mediana}[t(a_i) - t(r_i)]$$

Dove $t(r_i)$ indica il momento di pubblicazione della recensione e $t(a_i)$ quello della relativa risposta. Si utilizza la mediana anziché la media perché la distribuzione dei tempi di risposta può essere fortemente asimmetrica e contenere valori estremi molto elevati. L'indicatore può inoltre essere calcolato separatamente per recensioni positive e negative quando la numerosità delle osservazioni lo consente.

4. Dati e metodologia di analisi

4.1. I dataset

L'applicazione degli indicatori viene illustrata attraverso due dataset molto diversi per dimensione, piattaforma, mercato e natura delle recensioni. Come anticipato, questa eterogeneità è intenzionale, in quanto l'obiettivo non è costruire un campione rappresentativo dell'universo delle recensioni online, ma mostrare come il sistema di indicatori possa essere applicato in contesti differenti e quali cautele interpretative emergano al variare dei dati disponibili.

Il primo dataset è stato costruito mediante estrazione automatica delle recensioni pubblicamente disponibili su una piattaforma generalista italiana ([trustpilot.it](https://www.trustpilot.it)) e comprende undici marche appartenenti a cinque settori: servizi bancari e di pagamento, distribuzione beauty, gioielleria online, moda online e distribuzione culturale. Per ciascuna marca sono state acquisite, secondo l'ordinamento proposto dalla piattaforma, fino a un massimo di duecento recensioni. Dopo le operazioni di controllo e consolidamento illustrate nel paragrafo successivo, il dataset comprende complessivamente 2.142 recensioni uniche.

Il secondo dataset comprende 701.528 recensioni di prodotti della categoria beauty pubblicate su una piattaforma di commercio elettronico statunitense ([amazon.com](https://www.amazon.com)) tra il 2003 e il 2023. Per ciascuna recensione sono disponibili, oltre alla valutazione e al testo, l'identificativo del prodotto, la data di pubblicazione, il numero di voti di utilità ricevuti e l'indicazione relativa all'acquisto verificato. Quando l'analisi viene condotta a livello di singolo prodotto, sono considerati i 744 prodotti che dispongono di almeno cento recensioni, così da evitare che i confronti siano dominati da prodotti caratterizzati da una base informativa troppo limitata.

I due dataset svolgono quindi funzioni complementari. Il primo consente di osservare anche il comportamento di risposta delle imprese e di confrontare marche appartenenti a settori differenti; il secondo permette di verificare il comportamento degli indicatori valutativi e testuali su una base empirica di dimensione molto maggiore e in un contesto di piattaforma differente.

Tab. 1 — Composizione del dataset italiano

Settore	Marche	Recensioni
Servizi bancari e di pagamento	6	1.200
Distribuzione beauty	1	200
Gioielleria online	2	324
Moda online	1	199
Distribuzione culturale	1	219
Totale	11	2.142

4.2. Preparazione dei dati e condizioni di validità

Prima del calcolo degli indicatori è stato necessario verificare che i dati estratti potessero essere effettivamente interpretati nel modo previsto. Alcuni controlli possono apparire come semplici operazioni di pulizia del dataset, ma hanno in realtà un effetto diretto sulla validità dei risultati, in quanto un errore in questa fase può produrre indicatori formalmente corretti, ma riferiti a dati incompleti o conteggiati più volte.

Un primo problema riguarda la presenza di recensioni duplicate. Nel file relativo a una delle marche, ad esempio, le 1.200 righe estratte corrispondevano in realtà a sole 219 recensioni distinte, poiché alcuni record erano stati ripetuti fino a sette volte dalla procedura automatica di acquisizione. Se tali duplicazioni non fossero state riconosciute, quella marca avrebbe assunto nel dataset un peso molto superiore a quello reale, alterando non soltanto i suoi indicatori, ma anche le eventuali statistiche aggregate. Per questa ragione, prima di ogni elaborazione, le recensioni sono state ricondotte al loro identificativo univoco nativo e ogni duplicato è stato eliminato. Il punto ha una portata più generale: quando i dati vengono raccolti automaticamente, il numero di righe ottenute non coincide necessariamente con il numero di osservazioni effettivamente disponibili.

Un secondo problema riguarda la distinzione tra un valore assente e un'informazione non disponibile. I dati estratti per le undici marche non presentavano infatti una struttura perfettamente omogenea. Per due marche mancava il campo che indica la presenza di una risposta aziendale. Questo non significa che le imprese non abbiano risposto alle recensioni, ma soltanto che l'informazione non era stata rilevata nel processo di estrazione. Attribuire in questi casi un tasso di risposta pari a zero significherebbe trasformare un'informazione

manca in un comportamento dell'impresa che i dati non consentono di osservare. Gli indicatori relativi alla risposta vengono pertanto considerati non disponibili per queste marche e segnalati come tali nelle tabelle dei risultati.

È stata infine verificata l'uniformità delle date di pubblicazione. I diversi tracciati utilizzano infatti formati non sempre omogenei. Dove disponibile, è stato utilizzato il riferimento temporale espresso in secondi (epoch time), verificandone la corrispondenza con la data riportata in forma testuale. Una particolare attenzione è stata riservata alla convenzione statunitense mese-giorno-anno: interpretarla secondo l'ordine giorno-mese-anno avrebbe potuto modificare la corretta sequenza temporale delle recensioni e, di conseguenza, gli indicatori che dipendono dagli intervalli tra una pubblicazione e la successiva.

Questi controlli costituiscono quindi una condizione preliminare dell'analisi, per cui prima di chiedersi che cosa misurino gli indicatori, occorre assicurarsi che le osservazioni sulle quali vengono calcolati siano uniche, correttamente interpretate e realmente disponibili.

4.3. Soglie di interpretabilità

Alcuni indicatori vengono calcolati non sull'intero insieme delle recensioni, ma su gruppi più circoscritti. È il caso, ad esempio, del tasso di risposta alle sole valutazioni negative. In queste situazioni il numero complessivo di recensioni di una marca può essere relativamente elevato, mentre il sottoinsieme sul quale viene calcolato l'indicatore può risultare molto piccolo.

Un valore ottenuto su poche osservazioni è comunque matematicamente calcolabile, ma può essere fortemente influenzato da uno o due singoli casi e risultare quindi poco utile ai fini del confronto. Per evitare di attribuire eccessivo significato a valori instabili, nel presente lavoro gli indicatori riferiti alle sole recensioni negative vengono riportati soltanto quando sono disponibili almeno trenta osservazioni appartenenti a tale categoria.

La soglia non va interpretata come un valore universale oltre il quale un indicatore diventa automaticamente affidabile, ma come un criterio prudenziale di interpretabilità adottato nell'ambito di questa applicazione. Quando il numero di osservazioni non raggiunge

il livello stabilito, il valore non viene utilizzato nei confronti tra marche e viene indicato come non disponibile.

5. Risultati

Coerentemente con la finalità prevalentemente dimostrativa dell'analisi empirica, i risultati vengono letti soprattutto in funzione di ciò che i diversi indicatori consentono di rendere visibile rispetto alla sola osservazione della valutazione media. La presentazione segue i tre gruppi introdotti nel paragrafo 3: come valutano i clienti, come si esprimono e come risponde l'impresa.

5.1. Come valutano i clienti

La Tabella 2 riporta gli indicatori relativi alle undici marche del dataset italiano. Il primo risultato evidente riguarda il diverso grado di concordanza tra i clienti. Infatti, l'entropia varia da 0,56 a 1,84, su un massimo teorico di 2,32: dietro valutazioni medie apparentemente confrontabili possono quindi trovarsi situazioni molto diverse.

Tab. 2 — Come valutano i clienti: dataset italiano

Marca	N	$\bar{R}B$	H	B
BBVA	200	3,73	1,84	3,00
IBL Banca	200	4,68	0,75	9,53
ING	200	3,78	1,34	10,76
Santander Consumer	200	3,84	1,55	5,45
Satispay	200	3,91	1,42	7,70
Widiba	200	4,61	0,83	9,53
Sephora	200	3,37	1,44	10,76
Lapilli Shop	200	4,75	0,65	10,76
Sogni Gioielli	124	4,70	0,56	16,71
Yoox	199	4,38	1,13	6,65
La Feltrinelli	219	4,41	1,18	5,84

Le due marche di gioielleria online, Lapilli Shop e Sogni Gioielli, presentano valori di entropia inferiori a 0,7 e valutazioni medie corrette rispettivamente pari a 4,75 e 4,70. In questi casi il giudizio dei clienti è non soltanto elevato, ma anche relativamente omogeneo. La media descrive quindi abbastanza bene la direzione prevalente delle valutazioni, ma proprio l'elevato consenso ne riduce la capacità di differenziare marche che presentano configurazioni analoghe.

Una situazione diversa emerge, ad esempio, per BBVA, che presenta una valutazione media corretta di 3,73 e un'entropia pari a 1,84, la più elevata del dataset. In questo caso il valore medio sintetizza esperienze molto più eterogenee e rende particolarmente utile approfondire che cosa distingue i clienti che valutano positivamente la marca da quelli che ne hanno un'esperienza negativa. L'entropia non fornisce questa spiegazione, ma segnala dove cercarla.

Il dataset statunitense consente di osservare lo stesso problema su una scala molto più ampia. Sull'insieme delle 701.528 recensioni la valutazione media è pari a 3,96 e l'entropia a 1,74. Considerando i 744 prodotti con almeno cento recensioni, emerge una forte associazione inversa tra valutazione media ed entropia ($r = -0,82$): passando dal quartile dei prodotti con valutazioni più basse a quello con valutazioni più elevate, l'entropia media diminuisce da 1,99 a 0,97.

Tab. 3 — Concordanza e concentrazione agli estremi per quartile di valutazione media

Quartile di valutazione media	H medio	B medio
Primo quartile (più basso)	1,99	2,29
Secondo quartile	1,72	2,77
Terzo quartile	1,42	3,88
Quarto quartile (più alto)	0,97	7,21

Contemporaneamente aumenta il peso delle valutazioni estreme rispetto a quelle intermedie, in quanto l'indicatore B passa da 2,29 nel primo quartile a 7,21 nel quarto. I due risultati vanno letti congiuntamente. Nei prodotti meglio valutati, la crescita di B non indica necessariamente la presenza di due gruppi contrapposti, ma riflette soprattutto una crescente concentrazione verso l'estremo positivo della scala, coerente con la contemporanea riduzione dell'entropia.

L'utilità manageriale emerge proprio dalla combinazione dei diversi indicatori: la media dice quanto positivamente viene valutato un prodotto; l'entropia dice quanto quel giudizio è condiviso; la concentrazione agli estremi aiuta a comprendere quale forma assume la distribuzione. Nessuna delle tre informazioni, isolatamente, restituisce lo stesso quadro.

5.2. Come si esprimono i clienti

Gli indicatori relativi al flusso delle recensioni aggiungono una seconda prospettiva: non riguardano più soltanto il giudizio espresso, ma anche quando e quanto i clienti scrivono.

Per quanto riguarda la distribuzione temporale, l'indice di burstiness presenta valori compresi tra -0,01 e 0,67 nel dataset italiano. La maggior parte delle marche mostra valori relativamente contenuti, mentre La Feltrinelli presenta il valore più elevato (0,67), segnalando una maggiore tendenza delle recensioni a concentrarsi in raffiche anziché distribuirsi regolarmente nel tempo. Il dato non consente di stabilirne la causa, che potrebbe dipendere da eventi, campagne, episodi critici o altre circostanze, ma identifica un periodo o una configurazione sulla quale concentrare un eventuale approfondimento.

Tab. 4 — indice di burstiness

Marca	β
BBVA	0,31
IBL Banca	0,33
ING	0,28
Santander Consumer	0,39
Satispay	0,07
Widiba	0,16
Sephora	-0,01
Lapilli Shop	0,18
Sogni Gioielli	0,13
Yoox	0,23
La Feltrinelli	0,67

Il confronto diretto tra marche deve tuttavia essere effettuato con cautela, poiché le finestre temporali osservate non sono perfettamente allineate. In questa applicazione la

burstiness va quindi considerata soprattutto come indicatore utile a segnalare anomalie e a monitorare nel tempo la stessa marca, piuttosto che come base per costruire un ranking tra concorrenti.

Una seconda indicazione deriva dalla lunghezza delle recensioni. Come mostra la Tabella 5, il rapporto tra lunghezza del testo e valutazione assume configurazioni molto diverse nei due dataset.

Tab. 5 — Lunghezza mediana del testo per livello di valutazione

Dataset	1	2	3	4	5
Italiano (n = 2.142)	404	222	226	97	89
Statunitense (n = 701.528)	105	124	127	130	91

Nel dataset italiano le recensioni diventano progressivamente più brevi al crescere della valutazione: la mediana passa da 404 battute per le recensioni a una stella a 89 per quelle a cinque stelle. Le valutazioni fortemente negative contengono quindi, mediamente, una quantità di testo più di quattro volte superiore rispetto alle più positive.

Nel dataset statunitense il profilo è differente, in quanto la lunghezza cresce dalle valutazioni più basse fino alle quattro stelle, dove raggiunge il massimo di 130 battute, per poi diminuire a 91 nelle recensioni a cinque stelle. Non emerge quindi una regola generale secondo la quale il cliente insoddisfatto scriva sempre di più.

Una possibile interpretazione, da considerare come ipotesi e non come risultato dimostrato, riguarda la diversa natura delle piattaforme. In una piattaforma generalista la recensione può essere un atto deliberato, talvolta originato da un episodio problematico che il cliente sente la necessità di raccontare in dettaglio; in una piattaforma di commercio elettronico la recensione è invece inserita nel normale flusso successivo all'acquisto e può assumere maggiormente la funzione di consiglio rivolto ad altri consumatori.

Al di là della spiegazione, il risultato è metodologicamente rilevante e mostra come il profilo di lunghezza non possa essere trasferito automaticamente da una piattaforma a un'altra. Se lo scopo è utilizzare la lunghezza per orientare la successiva analisi del testo, il profilo deve essere costruito sullo specifico contesto osservato.

5.3. Come risponde l'impresa

La scomposizione delle risposte aziendali fornisce il risultato più direttamente collegato alla dimensione relazionale del lavoro. La Tabella 6 mostra che imprese apparentemente simili per presenza sulle piattaforme possono adottare comportamenti molto diversi a seconda del segno della recensione.

Tab. 6 — Come risponde l'impresa: dataset italiano

Marca	p_{neg}	p_{pos}	α	Δt
BBVA	94,4	100,0	-5,6	0,4
IBL Banca	75,0	99,5	-24,5	42,8
ING	96,7	97,8	-1,1	18,0
Santander Consumer	0,0	0,0	n.d.	—
Satispay	94,2	0,0	+94,2	120,9
Widiba	83,3	87,0	-3,7	21,3
Sephora	n.r.	n.r.	n.r.	—
Lapilli Shop	100,0	8,4	+91,6	106,4
Sogni Gioielli	100,0	100,0	0,0	83,1
Yoox	69,6	5,2	+64,4	—
La Feltrinelli	24,0	0,0	+24,0	261,1

Quattro delle nove marche per le quali l'asimmetria è interpretabile presentano un valore positivo: Satispay risponde al 94,2% delle recensioni negative e a nessuna di quelle positive; Lapilli Shop al 100% delle negative contro l'8,4% delle positive; Yoox al 69,6% contro il 5,2%; La Feltrinelli al 24% contro lo 0%. Sono configurazioni nelle quali l'attenzione dell'impresa si concentra quindi prevalentemente sul dissenso.

Tra le altre marche emergono comportamenti differenti. BBVA, ING, Widiba e Sogni Gioielli presentano tassi relativamente equilibrati tra recensioni positive e negative, mentre IBL Banca mostra un'asimmetria di segno opposto (-24,5), rispondendo più frequentemente ai giudizi positivi. Santander Consumer non risponde alle recensioni considerate, mentre per Sephora l'informazione relativa alla risposta non è disponibile.

L'aspetto più interessante è che configurazioni molto diverse convivono anche all'interno dello stesso settore. Nei servizi bancari e di pagamento si osservano imprese con risposta sostanzialmente simmetrica, un'impresa maggiormente orientata alle recensioni positive,

Satispay con un'asimmetria di +94,2 e Santander Consumer senza risposte. Nel dataset analizzato, quindi, la scelta di chi riceve una risposta non appare meccanicamente determinata dal settore, ma presenta un'importante componente riconducibile alle modalità con cui la singola impresa gestisce questo spazio di relazione.

Anche la tempestività differenzia fortemente i comportamenti osservati. La latenza mediana varia da 0,4 ore per BBVA a 261,1 ore per La Feltrinelli, cioè poco meno di undici giorni. Tra questi estremi si collocano configurazioni molto diverse, che comprendono imprese che rispondono diffusamente e rapidamente, e altre che privilegiano le recensioni negative ma intervengono dopo diversi giorni.

Asimmetria e latenza vanno quindi lette congiuntamente. Una forte propensione a rispondere alle recensioni negative non implica necessariamente una gestione tempestiva del cliente insoddisfatto: Satispay e Lapilli Shop, ad esempio, mostrano asimmetrie molto elevate, ma latenze mediane superiori alle cento ore. La combinazione dei due indicatori consente pertanto di distinguere a chi l'impresa dedica attenzione da quanto rapidamente riesce a trasformare tale attenzione in una risposta.

5.4. Rispondere di più significa avere clienti più concordi?

Un'ultima analisi esplorativa verifica se una maggiore presenza dell'impresa nelle recensioni sia associata a un maggiore accordo tra i clienti. L'ipotesi è plausibile alla luce della letteratura sul service recovery: se la risposta contribuisce a gestire efficacemente le esperienze negative, imprese più reattive potrebbero presentare nel tempo valutazioni meno eterogenee.

Nel dataset italiano, tuttavia, questa relazione non emerge. Per le marche sulle quali entrambi gli indicatori sono disponibili, la correlazione tra tasso complessivo di risposta ed entropia è debole: -0,14 secondo Pearson e -0,22 secondo Spearman. Con sole dieci osservazioni, entrambe le associazioni risultano però statisticamente non significative e non consentono conclusioni inferenziali.

Il risultato va quindi interpretato con prudenza e non dimostra che la risposta dell'impresa non produca effetti sulle valutazioni, né consente di stabilire relazioni causali. Mostra semplicemente che, in questa applicazione, rispondere più frequentemente non

coincide automaticamente con un maggiore consenso tra i recensori. È un'indicazione coerente con l'idea che gli indicatori relativi alle valutazioni dei clienti e quelli relativi al comportamento dell'impresa osservino dimensioni differenti e debbano essere letti congiuntamente, anziché considerati come misure alternative dello stesso fenomeno.

6. Il ruolo dell'intelligenza artificiale

Il calcolo degli indicatori presentati nei paragrafi precedenti non richiede l'impiego di alcun modello linguistico. Ciò non significa che l'intelligenza artificiale sia estranea al processo analitico, ma che il suo ruolo debba essere definito con precisione. In questa prospettiva si individuano tre possibili collocazioni, distinte ma complementari, accomunate da un principio: il modello opera come strumento di supporto e non come fonte diretta della misura pubblicata.

6.1. Collocazione “a monte”: validazione dell'integrità del dataset

Ognuno degli indicatori illustrati in precedenza poggia su un'assunzione implicita, cioè che le recensioni rappresentino espressioni autentiche di esperienza. Se una quota rilevante delle valutazioni massime fosse artificiale o priva di contenuto esperienziale, l'entropia non misurerebbe soltanto il consenso effettivo dei clienti, e anche la valutazione media corretta ne risulterebbe distorta.

È il punto nel quale l'aritmetica, da sola, non basta. Alcuni segnali elementari possono essere rilevati attraverso regole deterministiche e già consentono di dimensionare il fenomeno. Nel dataset italiano, 268 recensioni presentano un titolo identico al testo e 85 hanno una lunghezza non superiore a quindici battute; complessivamente si tratta di 322 unità, pari al 15,0 per cento del dataset. Queste recensioni presentano una valutazione media di 4,77, contro 4,18 dell'intero dataset, e nell'86,9 per cento dei casi attribuiscono la valutazione massima.

Il dato non dimostra l'esistenza di una manipolazione, in quanto una recensione molto breve può dipendere semplicemente dal limitato sforzo dedicato alla sua compilazione. Segnala però che le unità scarsamente informative non sono distribuite uniformemente lungo la scala, ma si concentrano soprattutto sulle valutazioni più elevate e possono quindi alterare gli indicatori calcolati sull'intero dataset. Inoltre, nessun testo superiore alle venti battute risulta duplicato: l'eventuale anomalia sembra quindi riguardare più la forma e la scarsa specificità del contenuto che la semplice ripetizione letterale, limite che i controlli deterministici faticano a intercettare.

Un possibile compito "a monte" del modello linguistico consiste allora nel supportare la verifica delle condizioni di validità del dataset: riconoscere testi templatizzati, recensioni prive di riferimenti esperienziali specifici o formulazioni fortemente ricorrenti associate a identità differenti. Il protocollo potrebbe classificare le recensioni, ad esempio, come "esperienziali", "non informative" o "sospette". La decisione finale sull'inclusione resterebbe comunque dell'analista: il modello segnala, non decide. Per preservare la riproducibilità, sia i criteri di classificazione sia il dataset utilizzato per il calcolo degli indicatori dovrebbero essere esplicitati e resi disponibili.

6.2. Collocazione "a valle": la lettura orientata

La seconda collocazione discende direttamente dagli indicatori relativi a come valutano i clienti. L'entropia può infatti funzionare come criterio di selezione, indicando dove l'analisi del testo abbia maggiori probabilità di produrre nuova informazione.

Quando il consenso è molto compatto, leggere un grande numero di recensioni può significare ritrovare molte volte indicazioni simili; l'analisi automatica rende questa lettura più rapida, ma non modifica la natura dell'informazione disponibile. Quando invece il disaccordo è elevato, coesistono esperienze diverse e diventa più interessante comprendere, attraverso il testo, che cosa le distingua. Nel dataset italiano l'entropia varia da 0,56 a 1,84, mostrando come la potenziale utilità dell'approfondimento testuale possa essere molto diversa da una marca all'altra.

L'antagonista della tesi qui sostenuta non è dunque l'intelligenza artificiale, ma il suo impiego indiscriminato. Gli indicatori orientano la selezione e consentono di concentrare

l'analisi semantica sulle aree del dataset che presentano maggiore eterogeneità o anomalie meritevoli di approfondimento. Il vantaggio non è soltanto economico, ma riguarda soprattutto il rapporto tra segnale e rumore. È in questo senso che la scelta di che cosa misurare precede, per importanza, la capacità di elaborare grandi quantità di testo.

6.3. Collocazione “a monte”: validazione dell'integrità del dataset

La terza collocazione riguarda l'azione anziché l'analisi. Gli indicatori non si limitano infatti a descrivere i comportamenti osservati, ma possono far emergere vuoti operativi. Nel dataset italiano, ad esempio, quattro marche concentrano fortemente le proprie risposte sui clienti insoddisfatti, lasciando senza riscontro gran parte di quelli soddisfatti.

In condizioni di risorse limitate, una simile scelta può rispondere a una logica di efficienza, per la quale, se non è possibile rispondere a tutti, appare razionale privilegiare i casi critici. La disponibilità di strumenti di generazione assistita modifica però almeno in parte questo vincolo, rendendo più sostenibile una maggiore copertura delle recensioni e una riduzione dei tempi di risposta. L'indicatore consente quindi di individuare uno spazio di intervento che la tecnologia può contribuire a rendere più facilmente gestibile.

Tuttavia, una risposta automatizzata che si limiti a riprodurre formule standard di cortesia rischia di ridurre il costo senza accrescere il valore relazionale dell'interazione. La leva diventa realmente utile quando restituisce al cliente l'evidenza che il contenuto della recensione è stato considerato e produce quindi un riscontro coerente e specifico. È questo, più della sola riduzione dei costi, il criterio con cui valutare l'impiego dell'intelligenza artificiale nella gestione operativa delle recensioni.

6.4. Il limite invariante

In tutte e tre le collocazioni il modello linguistico mantiene una funzione strumentale. Se una misura pubblicata dipendesse direttamente da un giudizio semantico del modello, sarebbe necessario verificarne la stabilità attraverso ripetizioni indipendenti e misure di concordanza esplicite. Finché, invece, gli indicatori restano calcolati su dati osservabili

mediante regole dichiarate, i risultati possono essere ricalcolati da chiunque disponga degli stessi dati e applichi la medesima procedura.

L'intelligenza artificiale può quindi intervenire a monte, contribuendo a verificare la qualità del materiale analizzato; a valle, orientando e approfondendo la lettura del testo; e in esecuzione, supportando la risposta dell'impresa. In nessuno di questi casi sostituisce però la scelta delle dimensioni da osservare, la definizione delle misure e l'interpretazione manageriale dei risultati.

7. Implicazioni per le decisioni di marketing

Gli indicatori proposti non hanno soltanto una funzione descrittiva. Letti congiuntamente, possono orientare sia alcune scelte di carattere strategico sia decisioni più direttamente operative. Le indicazioni che seguono vanno tuttavia interpretate in relazione al mercato, alla piattaforma e all'insieme competitivo considerato, evitando di trasformare i valori osservati nei due dataset in soglie valide in assoluto.

7.1. Livello strategico

Un primo risultato riguarda il significato del consenso. Una valutazione media elevata rappresenta certamente un segnale positivo, ma perde parte della propria capacità di differenziazione quando caratterizza in modo uniforme tutte le imprese di un determinato contesto. Nel dataset italiano, ad esempio, le due marche della gioielleria online presentano contemporaneamente valutazioni superiori a 4,7 e livelli di entropia inferiori a 0,7. In una situazione di questo tipo, costruire il posizionamento comunicativo sulla generica eccellenza del servizio rischia di enfatizzare un attributo sul quale i concorrenti appaiono altrettanto forti. Gli indicatori suggeriscono quindi di distinguere tra un risultato positivo per l'impresa e un elemento effettivamente discriminante sul piano competitivo.

Una maggiore eterogeneità dei giudizi apre, per converso, uno spazio di analisi più interessante. Un'entropia elevata non indica di per sé un problema né identifica le sue cause, ma segnala che clienti diversi stanno vivendo o valutando l'esperienza in modo differente. È in

queste situazioni che l'analisi del contenuto delle recensioni può risultare particolarmente utile per comprendere quali prodotti, processi, momenti della customer journey o segmenti di clientela siano associati alle esperienze più positive e più negative. Gli indicatori assumono quindi una funzione di selezione: non sostituiscono l'analisi qualitativa o semantica, ma contribuiscono a stabilire dove valga maggiormente la pena approfondire.

Anche l'asimmetria di risposta va interpretata come una scelta di allocazione dell'attenzione manageriale. Una forte concentrazione delle risposte sulle recensioni negative può riflettere una razionale strategia di service recovery, soprattutto quando le risorse disponibili sono limitate; la letteratura mostra infatti che priorità e modalità di risposta possono opportunamente variare in funzione della valenza della recensione e dei vincoli di budget (Wang et al., 2020; Lopes et al., 2023). Al tempo stesso, un'elevata asimmetria rende visibile una dimensione spesso trascurata: i clienti soddisfatti che dedicano tempo a esprimere pubblicamente il proprio apprezzamento possono non ricevere alcun riconoscimento dall'impresa. La scelta se riequilibrare tale comportamento dipende quindi dall'obiettivo assegnato alla risposta; se sia prevalentemente di recupero del disservizio oppure connesso ad una più ampia gestione della relazione. Gli effetti delle risposte a recensioni positive e negative non sono infatti necessariamente simmetrici (Wu, Ye e Chu, 2026).

Un'ulteriore implicazione riguarda la competitive intelligence. Gli indicatori utilizzati in questo lavoro sono costruiti, almeno nelle piattaforme che espongono i dati necessari, su informazioni pubblicamente osservabili. Possono quindi essere calcolati non soltanto per la propria impresa, ma anche per i concorrenti, consentendo di confrontare il modo in cui vengono valutati, come si presenta il flusso delle loro recensioni e come gestiscono pubblicamente le risposte. L'utilità non consiste tanto nell'attribuire un significato assoluto al singolo valore, quanto nel collocarlo all'interno di un insieme competitivo omogeneo e seguirne l'evoluzione nel tempo.

7.2. Livello operativo

Sul piano operativo, alcuni indicatori possono essere tradotti direttamente in obiettivi e criteri di intervento. La latenza di risposta, ad esempio, trasforma l'indicazione generica di

“rispondere rapidamente” in una misura osservabile e monitorabile. Può inoltre essere articolata per valenza della recensione, distinguendo il tempo di reazione ai feedback negativi da quello relativo ai riscontri ordinari. La rilevanza della tempestività è coerente con la letteratura sulla webcare, che individua nella risposta tempestiva e personalizzata una delle condizioni che ne accrescono l’efficacia (Lopes et al., 2023). Anche gli effetti osservati delle managerial responses sulla reputazione e sul successivo comportamento dei recensori confermano che la risposta non costituisce un atto neutro nello spazio pubblico della piattaforma (Proserpio e Zervas, 2017).

Il profilo di lunghezza offre invece un criterio per orientare la lettura del corpus. Come mostrato nel paragrafo 5.2, tuttavia, la relazione tra lunghezza e valutazione assume configurazioni differenti nei due dataset. Non è quindi opportuno assumere, ad esempio, che le recensioni negative siano sempre quelle più articolate o informative. Il profilo va costruito sul proprio contesto e utilizzato come criterio di priorità, non come regola universale.

La stessa logica vale più in generale per l’interpretazione degli indicatori. In assenza di benchmark consolidati e trasferibili, appare preferibile costruire riferimenti empirici “in casa”, confrontando l’impresa con concorrenti osservati sulla stessa piattaforma, nello stesso mercato e, per quanto possibile, nello stesso periodo. Un’asimmetria pari a +64, ad esempio, dice poco in assoluto; diventa più informativa se colloca l’impresa tra quelle che concentrano maggiormente le proprie risposte sui clienti insoddisfatti. Il benchmark rilevante è quindi soprattutto relativo e contestuale, più che fondato su soglie assolute.

Infine, la qualità degli indicatori dipende dalla qualità dei dati sui quali vengono calcolati. La verifica dell’integrità del dataset non dovrebbe quindi essere considerata un’attività preliminare svolta una tantum, ma una componente ricorrente del processo di analisi. Duplicazioni, recensioni scarsamente informative, anomalie temporali o variazioni nella struttura dei dati possono modificarsi nel tempo e alterare il significato delle misure. Il monitoraggio della qualità del flusso informativo diventa pertanto parte integrante dello stesso sistema di controllo (Jindal e Liu, 2008; Ott et al., 2011).

8. Limiti e agenda di ricerca

Il primo limite del lavoro riguarda la natura dei dati utilizzati. Il campionamento del dataset italiano non è probabilistico: per ciascuna marca sono state acquisite le recensioni restituite dalla piattaforma secondo il proprio ordinamento predefinito, fino al limite stabilito. A questo si aggiunge un problema più generale delle recensioni online, che riflettono per loro natura un processo di autoselezione da parte di chi decide di esprimere pubblicamente la propria esperienza (Hu, Pavlou e Zhang, 2006). I risultati non possono quindi essere interpretati come rappresentativi dell'intera popolazione dei clienti.

Una conseguenza specifica della procedura di estrazione riguarda inoltre la finestra temporale osservata. Per le marche caratterizzate da un elevato flusso di recensioni, le unità estratte coprono periodi relativamente brevi; per quelle con un flusso minore, la stessa numerosità si distribuisce su intervalli molto più lunghi, compresi nel dataset tra cinque giorni e sedici mesi. Ciò limita in particolare il confronto diretto dell'indicatore di burstiness tra marche diverse e ne rende più appropriato l'impiego per l'analisi della stessa marca nel tempo.

Un secondo limite riguarda la numerosità delle unità disponibili per i confronti tra imprese. Le correlazioni presentate nel paragrafo 5.4 poggiano su un numero molto ridotto di osservazioni e sono state pertanto interpretate esclusivamente in chiave esplorativa. Non consentono di formulare conclusioni inferenziali né, tanto meno, di stabilire relazioni causali tra il comportamento di risposta dell'impresa e la struttura delle valutazioni dei clienti. Un'estensione del numero di marche e settori costituisce quindi un naturale sviluppo dell'analisi.

Anche il perimetro empirico è necessariamente limitato. I dati provengono da due piattaforme, due mercati nazionali e un numero ristretto di categorie merceologiche; il dataset statunitense, in particolare, riguarda una sola categoria, quella beauty. Le differenze osservate nel profilo di lunghezza delle recensioni, discusse nel paragrafo 5.2, non consentono quindi di separare pienamente l'effetto della piattaforma da quello della categoria o del mercato. Replicare l'analisi su ulteriori categorie e piattaforme permetterebbe di comprendere meglio quali indicatori mantengano caratteristiche relativamente stabili e quali richiedano invece benchmark specifici per il contesto.

Un ulteriore limite riguarda la disponibilità dei dati necessari al calcolo degli indicatori relativi alla risposta dell'impresa. Non tutte le piattaforme rendono disponibile il riscontro aziendale e la relativa marca temporale; il sistema proposto è quindi pienamente applicabile soltanto negli ambienti nei quali queste informazioni sono pubblicamente osservabili. Anche questo vincolo suggerisce di considerare gli otto indicatori come un sistema modulare: alcuni possono essere utilizzati in qualunque ambiente di recensione, mentre altri dipendono dalle caratteristiche informative della piattaforma.

Una seconda direttrice dell'agenda di ricerca riguarda invece l'estensione del sistema dal comportamento osservabile al contenuto della relazione. Il presente lavoro misura dal lato del cliente come vengono formulate le valutazioni e dal lato dell'impresa se, a chi e con quale tempestività viene fornita una risposta, ma non classifica semanticamente la natura dell'esperienza raccontata. Un passo successivo potrebbe distinguere, ad esempio, recensioni riferite prevalentemente alla prestazione ricevuta da quelle nelle quali emerge una dimensione più propriamente relazionale (promessa disattesa, mancato riconoscimento, assenza di ascolto o percezione di trattamento inadeguato).

Una simile estensione richiederebbe inevitabilmente l'introduzione di un giudizio semantico e rappresenterebbe quindi anche un terreno naturale per l'impiego dei modelli linguistici. Coerentemente con la prospettiva adottata nel presente lavoro, tuttavia, il modello dovrebbe operare all'interno di un protocollo definito e verificabile: categorie esplicitate ex ante, codifica validata su un sottoinsieme di recensioni, confronto con il giudizio umano e verifica della stabilità delle classificazioni. La complementarità tra giudizio del ricercatore e capacità analitica degli LLM proposta dalla letteratura recente (Arora, Chakraborty e Nishimura, 2025) offre una possibile direzione in questo senso.

L'agenda di ricerca si sviluppa quindi lungo due direttrici complementari: estendere la base empirica, per verificare la trasferibilità e i limiti contestuali degli indicatori, e aumentarne progressivamente la profondità informativa, introducendo misure semantiche senza rinunciare alla trasparenza e alla riproducibilità del processo. È proprio in questo secondo passaggio che potrà essere ulteriormente verificato il principio su cui si fonda il lavoro: utilizzare l'intelligenza artificiale per ampliare la capacità di analisi, lasciando al ricercatore la definizione di ciò che vale la pena misurare e dei criteri con cui farlo.

Bibliografia

- Arora N., Chakraborty I. e Nishimura Y. (2025), AI-Human Hybrids for Marketing Research: Leveraging Large Language Models (LLMs) as Collaborators, *Journal of Marketing*, 89(2): 43-70.
- Bitner M. J., Booms B. H. e Tetreault M. S. (1990), The service encounter: diagnosing favorable and unfavorable incidents, *Journal of Marketing*, 54(1): 71-84.
- Chevalier J. A. e Mayzlin D. (2006), The effect of word of mouth on sales: online book reviews, *Journal of Marketing Research*, 43(3): 345-354.
- Dellarocas C. (2003), The digitization of word of mouth: promise and challenges of online feedback mechanisms, *Management Science*, 49(10): 1407-1424.
- Filieri R. (2015), What makes online reviews helpful? A diagnosticity-adoption framework to explain informational and normative influences in e-WOM, *Journal of Business Research*, 68(6): 1261-1270.
- Godes D. e Mayzlin D. (2004), Using online conversations to study word-of-mouth communication, *Marketing Science*, 23(4): 545-560.
- Grewal D., Satornino C.B., Davenport T. et al. (2025), How generative AI Is shaping the future of marketing, *Journal of the Academy of Marketing Science*, 53: 702-722.
- Hannah D. R., Treen E. R., Pitt L. e Berthon P. R. (2016), But you promised! Managing consumers' psychological contracts, *Business Horizons*, 59(4): 363-368.
- Hennig-Thurau T., Gwinner K. P., Walsh G. e Gremler D. D. (2004), Electronic word-of-mouth via consumer-opinion platforms: what motivates consumers to articulate themselves on the internet?, *Journal of Interactive Marketing*, 18(1): 38-52.
- Homburg C. e Fürst A. (2005), How organizational complaint handling drives customer loyalty: an analysis of the mechanistic and the organic approach, *Journal of Marketing*, 69(3): 95-114.
- Hu N., Pavlou P. A. e Zhang J. (2006), Can online reviews reveal a product's true quality? Empirical findings and analytical modeling of online word-of-mouth communication, *Proceedings of the 7th ACM Conference on Electronic Commerce*: 324-330.

- Jindal N. e Liu B. (2008), Opinion spam and analysis, *Proceedings of the International Conference on Web Search and Web Data Mining*: 219-230.
- Lopes A. I., Dens N., De Pelsmacker P. e Malthouse E. C. (2023), Managerial response strategies to eWOM: A framework and research agenda for webcare, *Tourism Management*, 98: 104739.
- Ma L., Sun B. e Kekre S. (2015), The squeaky wheel gets the grease: an empirical analysis of customer voice and firm intervention on Twitter, *Marketing Science*, 34(5): 627-645.
- Moe W. W. e Trusov M. (2011), The value of social dynamics in online product ratings forums, *Journal of Marketing Research*, 48(3): 444-456.
- Morgan R. M. e Hunt S. D. (1994), The commitment-trust theory of relationship marketing, *Journal of Marketing*, 58(3): 20-38.
- Mudambi S. M. e Schuff D. (2010), What makes a helpful online review? A study of customer reviews on Amazon.com, *MIS Quarterly*, 34(1): 185-200.
- Ott M., Choi Y., Cardie C. e Hancock J. T. (2011), Finding deceptive opinion spam by any stretch of the imagination, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*: 309-319.
- Proserpio D. e Zervas G. (2017), Online Reputation Management: Estimating the Impact of Management Responses on Consumer Reviews, *Marketing Science*, 36(5): 645-665.
- Rousseau D. M. (1989), Psychological and implied contracts in organizations, *Employee Responsibilities and Rights Journal*, 2(2): 121-139.
- Sarstedt M., Adler S.J., Rau L. e Schmitt B. (2024), Using large language models to generate silicon samples in consumer and marketing research: Challenges, opportunities, and guidelines, *Psychology & Marketing*, 41: 1254–1270.
- Shannon C. E. (1948), A mathematical theory of communication, *Bell System Technical Journal*, 27: 379-423, 623-656.
- Smith A. K., Bolton R. N. e Wagner J. (1999), A model of customer satisfaction with service encounters involving failure and recovery, *Journal of Marketing Research*, 36(3): 356-372.

- Sun M. (2012), How does the variance of product ratings matter?, *Management Science*, 58(4): 696-707.
- Van Noort G. e Willemsen L. M. (2012), Online damage control: the effects of proactive versus reactive webcare interventions in consumer-generated and brand-generated platforms, *Journal of Interactive Marketing*, 26(3): 131-140.
- Wang L., Ren X., Wan H. e Yan J. (2020), Managerial responses to online reviews under budget constraints: Whom to target and how, *Information & Management*, 57(8): 103382.
- Wu J., Ye J. e Chu J. (2026), Soothing the unsatisfied or pleasing the satisfied? The effects of managerial responses to positive versus negative reviews on customer ratings and financial performance, *Journal of the Academy of Marketing Science*, 54(1): 70-90.

La struttura informativa delle recensioni online
Un sistema di indicatori diagnostici e il ruolo
dell'intelligenza artificiale

di Enrico Bonetti, Josè Compagnone

IPE Working Paper

N. 34

September 17, 2026

ISSN 2284-1229

<https://doi.org/10.5281/zenodo.22813882>

La struttura informativa delle recensioni online

Un sistema di indicatori diagnostici e il ruolo dell'intelligenza artificiale

Enrico Bonetti, Josè Compagnone

Abstract

L'analisi delle recensioni online viene spesso presentata come un problema di scala, che i modelli linguistici di grandi dimensioni consentirebbero finalmente di superare grazie alla capacità di elaborare corpus troppo ampi per l'analisi manuale. Questo lavoro propone una prospettiva diversa, secondo la quale il principale vincolo non è soltanto computazionale, ma concettuale, perché il valore informativo dell'analisi dipende anzitutto dalla scelta di che cosa osservare e misurare.

Il contributo propone un sistema di otto indicatori elementari, calcolabili su informazioni direttamente osservabili e senza richiedere, nella prima fase, interpretazione semantica del testo. Gli indicatori sono organizzati in tre gruppi: come valutano i clienti, come si esprimono e come risponde l'impresa. La loro applicazione è illustrata su due dataset volutamente differenti: 2.142 recensioni italiane relative a undici brand in cinque settori e 701.528 recensioni statunitensi della categoria beauty. L'analisi empirica ha finalità prevalentemente dimostrativa ed esplorativa: mostra come misure semplici possano rendere visibili configurazioni che la sola valutazione media o indicatori aggregati tendono a occultare. In particolare, l'asimmetria di risposta (cioè la differenza tra il tasso di risposta alle valutazioni negative e quello alle positive) rende osservabili diverse modalità di allocazione dell'attenzione dell'impresa e, nel dataset analizzato, varia sensibilmente anche all'interno dello stesso settore. Inoltre, il lavoro individua tre ruoli dell'intelligenza artificiale nel processo: validazione dell'integrità del dataset, lettura orientata dagli indicatori e supporto all'esecuzione operativa. In tutti e tre i casi, il modello resta strumento di supporto, mentre la definizione delle misure e la loro interpretazione rimangono responsabilità del ricercatore.

Parole chiave: recensioni online; passaparola digitale; indicatori diagnostici; risposte aziendali; marketing relazionale; intelligenza artificiale; ricerche di marketing.

Online review analysis is often framed as a problem of scale, one that large language models would finally make it possible to overcome thanks to their ability to process corpora too large for manual analysis. This paper advances a different perspective: the main constraint is not only computational but conceptual, because the informational value of the analysis depends first and foremost on deciding what should be observed and measured.

The paper proposes a system of eight elementary indicators, calculated from directly observable information and requiring no semantic interpretation of the text in the initial stage. The indicators are organized into three groups: how customers evaluate, how they express themselves, and how firms respond. Their application is illustrated using two deliberately different datasets: 2,142 Italian reviews covering eleven brands across five sectors and 701,528 U.S. reviews in the beauty category. The empirical analysis is primarily demonstrative and exploratory: it shows how simple measures can reveal patterns that average ratings or aggregate indicators tend to conceal. In particular, response asymmetry—the difference between the response rate to negative ratings and the response rate to positive ratings—makes different patterns of managerial attention visible and, in the dataset analyzed, varies substantially even within the same sector. The paper also identifies three roles for artificial intelligence in the analytical process: validating dataset integrity, enabling indicator-guided reading, and supporting operational execution. In all three cases, the model remains a support tool, while the definition of the measures and the interpretation of their implications remain the responsibility of the researcher.

Keywords: online reviews; electronic word of mouth; diagnostic indicators; managerial responses; relationship marketing; artificial intelligence; marketing research.

Indice

1. Introduzione	6
2. Quadro di riferimento	9
3. Il sistema di indici	13
3.1 La media non basta	13
3.2 Come valutano i clienti	15
3.2.1. Quanto è affidabile il confronto? La valutazione corretta per la numerosità	15
3.2.2. Quanto sono concordi i clienti?	16
3.2.3. I giudizi sono moderati o polarizzati?	18
3.3 Come si esprimono i clienti	19
3.3.1. Le recensioni arrivano regolarmente o per picchi?	20
3.3.2. Dove si concentra il contenuto informativo?	21
3.4. Come risponde l'impresa	23
3.4.1. A quante recensioni risponde?	23
3.4.2. A chi sceglie di rispondere?	24
3.4.3. Quanto rapidamente risponde?	26
4. Dati e metodologia di analisi	28
4.1 I dataset	28
Tabella 1: Composizione del dataset italiano	29
4.2 Preparazione dei dati e condizioni di validità	29
4.3 Soglie di interpretabilità	30
5. Risultati	31
5.1. Come valutano i clienti	31
Tabella 2: Come valutano i clienti: dataset italiano	31
Tabella 3: Concordezza e concentrazione agli estremi per quartile di valutazione media	32
5.2. Come si esprimono i clienti	33
Tebella 4: indice di burstiness	33
Tabella 5: Lunghezza mediana del testo per livello di valutazione	34
5.3. Come risponde l'impresa	35

Tabella 6: Come risponde l'impresa: dataset italiano	35
5.4. Rispondere di più significa avere clienti più concordi?	36
6. Il ruolo dell'intelligenza artificiale	37
6.1. Collocazione "a monte": validazione dell'integrità del dataset	37
6.2. Collocazione "a valle": la lettura orientata	38
6.3. Collocazione "in esecuzione": dalla misura alla leva	39
6.4. Il limite invariante	39
7. Implicazioni per le decisioni di marketing	40
7.1. Livello strategico	40
7.2. Livello operativo	41
8. Limiti e agenda di ricerca	43
Bibliografia	45

1. Introduzione

Le recensioni online costituiscono da tempo una fonte rilevante per comprendere il comportamento dei consumatori, la reputazione delle imprese e le dinamiche del word of mouth elettronico (Dellarocas, 2003; Godes e Mayzlin, 2004; Chevalier e Mayzlin, 2006). Negli ultimi anni, tuttavia, il dibattito sulla loro analisi si è progressivamente spostato sulle opportunità offerte dall'intelligenza artificiale generativa e, in particolare, dai modelli linguistici di grandi dimensioni (LLM). La capacità degli LLM di elaborare grandi quantità di testo ha infatti ampliato in misura considerevole le possibilità della ricerca di marketing e dell'analisi dei dati non strutturati (Sarstedt et al., 2024; Arora, Chakraborty e Nishimura, 2025; Grewal et al., 2025). Corpus di migliaia o centinaia di migliaia di recensioni, impossibili da leggere manualmente, possono oggi essere analizzati in tempi e con costi prima impensabili.

La maggiore capacità di elaborazione, tuttavia, non coincide necessariamente con una maggiore capacità di produrre conoscenza utile alle decisioni. È questa l'osservazione da cui prende avvio il presente lavoro. Se applicata indistintamente a un insieme molto ampio di recensioni, l'analisi automatica tende innanzitutto a restituire ciò che nel dataset ricorre con maggiore frequenza: i temi dominanti, gli aspetti più apprezzati o criticati, le valutazioni prevalenti. Si tratta di informazioni certamente utili, ma spesso già note, almeno nelle loro linee essenziali, a chi gestisce la marca. Passare da mille a un milione di recensioni analizzate può quindi rendere più precisa la rappresentazione di fenomeni già evidenti senza necessariamente far emergere quelli meno visibili e, proprio per questo, potenzialmente più rilevanti per la decisione.

Il punto, dunque, non è soltanto leggere di più, ma sapere che cosa cercare. Prima della capacità di elaborare i dati viene la capacità di formulare domande alle quali quei dati possano fornire risposte utili. Da questa prospettiva, il principale limite dell'analisi delle recensioni non è necessariamente computazionale, ma può essere concettuale e riguarda la scelta delle dimensioni da osservare e degli indicatori attraverso i quali renderle misurabili. Un indicatore semplice, costruito per rispondere a una domanda manageriale precisa, può in alcuni casi fornire insight molto più utili di quelli derivanti da un'analisi molto sofisticata applicata all'intero corpus senza una preventiva selezione di ciò che si intende osservare.

Il presente lavoro propone, in questa prospettiva, un sistema di otto indici elementari, costruiti a partire da informazioni strutturate normalmente disponibili nelle piattaforme di

recensione: valutazione numerica, data di pubblicazione, presenza o assenza di una risposta dell'impresa ed eventuale momento in cui tale risposta viene fornita. Gli indici non richiedono, per essere calcolati, alcuna interpretazione semantica del testo. Alcuni descrivono la struttura delle valutazioni espresse dai clienti, evidenziando fenomeni che la sola media tende a comprimere; altri osservano invece il comportamento dell'impresa nei confronti di chi recensisce, permettendo di cogliere aspetti della relazione che il semplice tasso complessivo di risposta non rende visibili. L'obiettivo non è contrapporre questi indicatori all'analisi testuale né sostenere che misure semplici siano, in assoluto, superiori a tecniche più sofisticate. La proposta è piuttosto quella di utilizzare gli indicatori per stabilire dove guardare e che cosa cercare, riservando successivamente l'analisi semantica ai fenomeni e alle porzioni del corpus per i quali essa può generare un reale valore informativo.

L'analisi empirica condotta nel presente lavoro svolge, coerentemente con questa impostazione, una funzione prevalentemente dimostrativa ed esplorativa. Non è costruita per produrre stime rappresentative dell'universo delle recensioni online né per verificare relazioni causali tra le variabili considerate. Il suo scopo principale è mostrare come gli otto indici possano essere concretamente applicati a dati reali e quale capacità abbiano di mettere in evidenza configurazioni e comportamenti che gli indicatori aggregati normalmente utilizzati possono lasciare in secondo piano. In altri termini, i dati vengono utilizzati non tanto per dimostrare che determinate configurazioni caratterizzino in generale un settore o un mercato, quanto per verificare se il sistema di misurazione proposto sia in grado di rendere osservabili differenze manageriali altrimenti poco evidenti.

A questo scopo vengono utilizzati due dataset volutamente molto differenti. Il primo comprende 2.142 recensioni pubblicate su una piattaforma generalista italiana e relative a undici brand appartenenti a cinque settori: servizi bancari e di pagamento, distribuzione beauty, gioielleria online, moda online e distribuzione culturale. Il secondo comprende 701.528 recensioni di prodotti della categoria beauty pubblicate su una piattaforma di commercio elettronico statunitense nell'arco di vent'anni. La distanza tra i due dataset — per lingua, mercato, piattaforma, ordine di grandezza e natura dell'atto comunicativo — è intenzionale. Essa consente di mostrare l'applicazione degli stessi strumenti in contesti molto diversi e, al tempo stesso, di osservare quali indicatori mantengano un comportamento interpretabile al variare del contesto e quali, invece, richiedano criteri di lettura specifici. In quest'ultimo caso,

la dipendenza dal contesto non costituisce necessariamente un limite dello strumento, ma un'informazione utile per definirne correttamente le condizioni di impiego.

Il lavoro è articolato come segue. Il paragrafo 2 presenta il quadro teorico di riferimento, richiamando in particolare la letteratura sul word of mouth elettronico, sulla struttura delle valutazioni online e sul recupero del servizio. Il paragrafo 3 introduce il sistema degli indici, partendo dai limiti della valutazione media, tuttora dominante nella pratica manageriale ma incapace di rappresentare pienamente la struttura delle esperienze che la determinano. Il paragrafo 4 descrive i dataset e le procedure di preparazione dei dati, con particolare attenzione alle condizioni di validità poste dall'estrazione automatica. Il paragrafo 5 presenta i risultati dell'applicazione empirica. Il paragrafo 6 discute il ruolo che l'intelligenza artificiale può assumere nelle diverse fasi del processo analitico. Il paragrafo 7 traduce i risultati in implicazioni per le decisioni di marketing strategico e operativo. Il paragrafo 8, infine, presenta i principali limiti del lavoro e le possibili direttrici di ricerca futura.

2. Quadro di riferimento

La letteratura sulle recensioni online si sviluppa inizialmente nell'ambito degli studi sul word of mouth elettronico, concentrandosi soprattutto sul valore informativo delle valutazioni espresse dai consumatori e sui loro effetti sul mercato. I primi contributi mostrano che le recensioni influenzano in misura significativa le vendite e le scelte dei consumatori (Godes e Mayzlin, 2004; Chevalier e Mayzlin, 2006), mentre Dellarocas (2003) interpreta i sistemi di reputazione digitale come vere e proprie istituzioni economiche. Hennig-Thurau et al. (2004) spostano l'attenzione anche sulle motivazioni che inducono il consumatore a pubblicare una recensione, mettendo in evidenza come essa non costituisca soltanto un segnale informativo rivolto al mercato, ma anche un atto sociale attraverso il quale il cliente rende pubblica la propria esperienza con l'impresa.

Un secondo insieme di studi mostra che l'informazione contenuta nelle recensioni non è adeguatamente rappresentata dalla sola valutazione media. Sun (2012) evidenzia che la varianza delle valutazioni produce effetti sulle vendite indipendenti dalla media e che tali effetti dipendono dalle caratteristiche del prodotto; Hu, Pavlou e Zhang (2006) documentano la frequente bimodalità delle distribuzioni, riconducendola anche a meccanismi di autoselezione nella decisione di recensire; Moe e Trusov (2011) mostrano come le interazioni sociali tra recensori contribuiscano alla formazione di tali distribuzioni nel tempo. Nel loro insieme, questi contributi suggeriscono che la forma della distribuzione (consenso, dispersione, polarizzazione) contenga informazioni che la media tende a comprimere. È su questa base che si collocano gli indicatori di struttura valutativa utilizzati nel presente lavoro: non come nuove misure statistiche, ma come strumenti per tradurre proprietà già note delle distribuzioni in segnali più immediatamente utilizzabili nella diagnosi manageriale.

Anche le caratteristiche della singola recensione possono fornire informazioni ulteriori rispetto al rating. Mudambi e Schuff (2010) mostrano, ad esempio, che la relazione tra estrema della valutazione e utilità percepita non è lineare e varia in funzione della natura del prodotto, mentre Filieri (2015) evidenzia il ruolo della qualità informativa e della credibilità della recensione. Questa letteratura è rilevante per il presente lavoro soprattutto perché richiama l'attenzione sulla dipendenza dal contesto delle caratteristiche testuali, ma lunghezza, articolazione e contenuto non assumono necessariamente lo stesso significato in piattaforme,

mercati e categorie differenti. Ne discende l'opportunità di evitare soglie o regole interpretative trasferite automaticamente da un contesto all'altro.

Il passaggio più rilevante ai fini della tesi qui proposta avviene quando l'attenzione si sposta dal comportamento di chi recensisce a quello dell'impresa che decide se e come rispondere. La letteratura sul service failure e sul service recovery aveva già mostrato che la valutazione dell'esperienza dipende non soltanto dalla prestazione originaria, ma anche dal modo in cui l'impresa gestisce l'eccezione e il reclamo (Bitner, Booms e Tetreault, 1990; Smith, Bolton e Wagner, 1999; Homburg e Fürst, 2005). Con la diffusione delle piattaforme digitali, la risposta al cliente diventa però pubblica, e quindi non costituisce più soltanto un intervento di recupero rivolto al singolo consumatore, ma anche un comportamento osservabile dagli altri clienti e dal mercato.

Da questa prospettiva si sviluppa la letteratura sulle managerial responses e sulla webcare. In tal senso, Van Noort e Willemsen (2012) analizzano gli effetti delle risposte aziendali nei contesti digitali, mentre Ma, Sun e Kekre (2015) mostrano che l'intervento pubblico dell'impresa può produrre effetti differenti in funzione delle caratteristiche del messaggio del cliente. Proserpio e Zervas (2017) documentano inoltre come la scelta di rispondere alle recensioni incida sulla successiva reputazione online dell'impresa. Gli studi successivi approfondiscono la natura selettiva della risposta: Wang et al. (2020) analizzano a quali recensioni convenga destinare risorse limitate; Lopes et al. (2023), attraverso una revisione sistematica della letteratura sulla webcare, mostrano l'importanza della valenza della recensione, della tempestività e della personalizzazione della risposta; Wu, Ye e Chu (2026) distinguono esplicitamente gli effetti dei tassi di risposta alle recensioni positive e negative. La domanda non è quindi più soltanto se l'impresa risponda, ma a chi scelga di rispondere, con quale tempestività e secondo quale logica.

È precisamente in questo passaggio che la gestione delle recensioni assume una dimensione relazionale. La teoria del relationship marketing attribuisce a fiducia, impegno e reciprocità un ruolo centrale nella costruzione di rapporti durevoli tra impresa e cliente (Morgan e Hunt, 1994). In modo complementare, la nozione di contratto psicologico, sviluppata originariamente in ambito organizzativo (Rousseau, 1989) e successivamente estesa alla relazione impresa-consumatore (Hannah et al., 2016), consente di interpretare alcune manifestazioni di insoddisfazione non soltanto come giudizi negativi sulla singola prestazione, ma come reazioni alla percezione che una promessa o un'obbligazione implicita della relazione

sia stata disattesa. In questa prospettiva, anche il comportamento dell'impresa successivo alla recensione — rispondere o tacere, intervenire rapidamente o tardivamente, privilegiare chi protesta o riconoscere anche chi esprime apprezzamento — diventa un segnale osservabile del modo in cui l'impresa gestisce la relazione.

Un ulteriore problema riguarda la qualità dei dati sui quali tali misure vengono calcolate. La letteratura sull'integrità dei dataset di recensioni ha sviluppato sia approcci fondati su segnali comportamentali (Jindal e Liu, 2008), sia metodi basati sull'analisi linguistica del testo (Ott et al., 2011). La questione assume oggi nuova rilevanza con la disponibilità dei modelli linguistici di grandi dimensioni. La ricerca più recente sottolinea infatti le potenzialità degli LLM come strumenti di supporto alla ricerca di marketing e, al tempo stesso, l'opportunità di integrarli con il giudizio e il controllo del ricercatore anziché considerarli sostituti autonomi del processo analitico (Arora, Chakraborty e Nishimura, 2025). Per l'analisi delle recensioni ciò apre uno spazio importante per l'intelligenza artificiale, ma pone anche il problema di distinguere le attività nelle quali il modello può assistere l'analista da quelle nelle quali la misura deve restare direttamente verificabile e riproducibile.

Il presente lavoro si colloca all'intersezione di questi contributi, ma ne sposta il punto di osservazione. La letteratura ha ampiamente studiato gli effetti delle recensioni e delle risposte aziendali: sulle vendite, sulla reputazione, sulle valutazioni successive e sul comportamento dei consumatori. Molto meno centrale è stata la costruzione di un sistema parsimonioso di indicatori finalizzato non tanto a spiegare gli esiti di mercato, quanto a diagnosticare la configurazione osservabile della relazione tra impresa e recensori a partire da dati pubblicamente disponibili. La tesi proposta è che una parte significativa di questa diagnosi possa essere condotta prima di qualunque interpretazione semantica, attraverso un numero limitato di misure semplici e riproducibili.

Il contributo è dunque prevalentemente di natura metodologica e manageriale. Gli indicatori proposti non intendono sostituire i costrutti e le tecniche sviluppati dalla letteratura precedente, ma organizzarne alcune intuizioni in un sistema di lettura orientato alla decisione: distinguere consenso e polarizzazione, osservare se l'impresa risponde e a chi risponde, misurarne la tempestività e individuare le porzioni del corpus nelle quali un'analisi più approfondita può produrre maggiore informazione. In questa architettura anche il ruolo dell'intelligenza artificiale risulta meglio definito: non come fonte autonoma della misura, ma come strumento utilizzabile per verificarne le condizioni di validità, approfondire i fenomeni

segnalati dagli indicatori e supportare l'esecuzione operativa delle decisioni che ne derivano. In questa prospettiva, la crescente capacità di elaborazione offerta dall'intelligenza artificiale non riduce il ruolo del ricercatore, ma ne ridefinisce e, per certi versi, ne rafforza la centralità. Se la macchina può assumere una quota crescente del lavoro di lettura, classificazione ed esecuzione, resta infatti all'analista il compito non delegabile di formulare le domande, scegliere le dimensioni rilevanti, costruire le misure e interpretarne le implicazioni. Quanto più aumenta la capacità di elaborare i dati, tanto più il valore dell'analisi dipende dalla capacità di decidere che cosa valga la pena misurare. È in questo spazio che si colloca la proposta sviluppata nel presente lavoro.

3. Il sistema di indici

3.1. La media non basta

La valutazione media è l'indicatore attorno al quale è organizzata gran parte dell'esperienza pubblica delle recensioni online: è il numero mostrato accanto al nome della marca o del prodotto, quello utilizzato per costruire classifiche e confronti e, spesso, quello che entra nei cruscotti direzionali. La ragione della sua diffusione è evidente, in quanto riesce a concentrare in un unico valore un insieme anche molto ampio di giudizi ed è immediatamente comprensibile. Proprio questa capacità di sintesi costituisce, tuttavia, anche il suo principale limite.

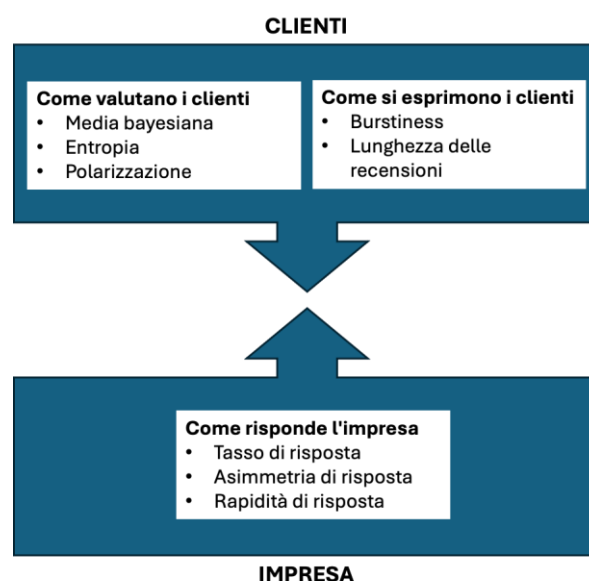
Due distribuzioni profondamente diverse possono infatti produrre la stessa media. Quattro valutazioni da cinque stelle e una da una stella restituiscono una media di 4,2; lo stesso valore si ottiene con quattro valutazioni da quattro stelle e una da cinque. Dal punto di vista manageriale, le due situazioni raccontano storie differenti. Nel primo caso il servizio soddisfa pienamente la maggioranza dei clienti ma produce, per una minoranza, un'esperienza fortemente negativa: il problema consiste allora nell'individuare le condizioni nelle quali si verifica il fallimento e intervenire sulle eccezioni. Nel secondo caso le esperienze sono invece omogeneamente positive e non emerge un analogo segnale di criticità. La media è la stessa, ma le decisioni che ne dovrebbero derivare sono diverse.

Il problema non è quindi che la media sia un indicatore errato, ma che risponde a una domanda soltanto: qual è il livello medio delle valutazioni? Non dice quanto i clienti concordino tra loro, se i giudizi siano concentrati o polarizzati, né quale struttura si nasconda dietro lo stesso valore sintetico. Per rispondere a queste domande è necessario osservare anche la distribuzione delle valutazioni.

A questa perdita di informazione si aggiunge un secondo problema. Quando il numero di recensioni è ridotto, poche osservazioni possono modificare sensibilmente la media e rendere poco affidabile il confronto tra marche o prodotti caratterizzati da numerosità molto diverse. Anche in questo caso non si tratta di rinunciare alla media, ma di correggerla e affiancarla ad altre misure capaci di descrivere ciò che essa, per costruzione, non può mostrare.

Esiste poi una dimensione che la distribuzione delle valutazioni, da sola, non può cogliere. Una recensione non è soltanto un giudizio espresso dal cliente, ma una volta pubblicata, può diventare anche l'inizio di un'interazione pubblica con l'impresa. L'azienda può rispondere o non rispondere, può farlo rapidamente o dopo molti giorni, può concentrare la propria attenzione sui clienti insoddisfatti oppure rivolgersi anche a quelli soddisfatti. Questi comportamenti non descrivono più soltanto come il cliente valuta l'impresa, ma anche come l'impresa gestisce la relazione con chi ha scelto di esprimere pubblicamente quella valutazione. È in questo passaggio dalla valutazione all'interazione che si colloca la dimensione relazionale proposta nel presente lavoro.

Il sistema di indicatori viene pertanto articolato in tre gruppi complementari. Il primo riguarda come valutano i clienti e descrive il livello, il grado di concordanza e l'eventuale polarizzazione dei giudizi, consentendo di andare oltre il solo valore medio. Il secondo osserva come si esprimono i clienti, considerando la distribuzione temporale delle recensioni e la quantità di contenuto testuale associata ai diversi livelli di valutazione. Il terzo riguarda come risponde l'impresa e rende osservabile il suo comportamento nello spazio pubblico della recensione, misurandone la presenza, la selettività e la tempestività della risposta. Gli indicatori sono costruiti a partire da un numero limitato di informazioni direttamente osservabili (valutazione numerica, marca temporale, lunghezza del testo, presenza ed eventuale data della risposta aziendale) senza richiedere, in questa prima fase, alcuna interpretazione semantica del contenuto delle recensioni.



3.2. Come valutano i clienti

3.2.1. Quanto è affidabile il confronto? La valutazione corretta per la numerosità

La domanda manageriale.

Quando si confrontano marche, prodotti o punti vendita attraverso le recensioni online, una differenza spesso trascurata riguarda il numero di giudizi sui quali è calcolata la valutazione media. Una marca con poche decine di recensioni e una con diverse migliaia possono presentare lo stesso punteggio, ma quel valore non ha necessariamente la stessa solidità. La domanda rilevante non è quindi soltanto “qual è la valutazione media?”, ma anche “quanto possiamo considerarla affidabile ai fini del confronto?”.

Il limite della media tradizionale.

La media aritmetica attribuisce lo stesso significato a un punteggio indipendentemente dal numero di osservazioni che lo ha generato. Una valutazione di 4,8 ottenuta su venti recensioni viene così trattata allo stesso modo di una valutazione di 4,8 costruita su duemila recensioni, nonostante nel primo caso pochi nuovi giudizi possano modificarla sensibilmente. Il problema diventa particolarmente rilevante quando si confrontano imprese caratterizzate da volumi di recensioni molto differenti: le posizioni apparentemente migliori possono dipendere, almeno in parte, dalla maggiore instabilità delle medie calcolate su campioni ridotti.

L'indicatore proposto.

Per ridurre questo effetto viene utilizzata una valutazione media corretta per la numerosità. L'idea è intuitiva: quando le recensioni disponibili sono poche, il punteggio osservato viene prudenzialmente avvicinato alla valutazione media dell'intero insieme considerato; man mano che il numero delle recensioni aumenta, il valore della singola marca pesa sempre di più e la correzione diventa progressivamente irrilevante. L'indicatore non sostituisce quindi il giudizio espresso dai clienti, ma tiene conto della diversa quantità di evidenza sulla quale esso si basa.

L'utilità per la decisione.

La correzione rende più prudente e comparabile il benchmarking tra marche, prodotti o concorrenti con numerosità differenti. È particolarmente utile nella costruzione di ranking, cruscotti competitivi e sistemi di monitoraggio, perché riduce il rischio di attribuire eccessivo significato a valutazioni molto elevate — o molto basse — fondate su un numero limitato di osservazioni. Non indica se i clienti siano concordi o polarizzati, aspetto affrontato dagli indicatori successivi; consente però di stabilire una base più solida dalla quale partire per il confronto.

Nota tecnica.

La misura adottata è una media corretta verso il valore medio del dataset (c.d. media bayesiana):

$$RB = (C \cdot m + \sum r_i) / (C + n)$$

Dove r_i è la valutazione della recensione i -esima, n il numero di recensioni della marca, m la media dell'intero dataset e C una costante di smorzamento, interpretabile come un numero di osservazioni aggiuntive collocate al valore medio m . All'aumentare di n , il peso della correzione diminuisce e l'indicatore converge verso la media osservata della marca. Il valore di C deve essere esplicitato; nelle elaborazioni presentate nel lavoro è posto pari a 20.

3.2.2. Quanto sono concordi i clienti?

La domanda manageriale.

Conoscere la valutazione media di una marca non basta per capire quanto quell'opinione sia condivisa. Due marche possono presentare lo stesso punteggio medio, ma trovarsi in condizioni molto differenti: nel primo caso quasi tutti i clienti possono esprimere giudizi simili; nel secondo, lo stesso valore può essere il risultato di esperienze molto diverse tra loro. Per chi deve decidere, la domanda diventa quindi: quanto sono concordi i clienti nel valutare la marca?

Il limite della valutazione media.

La media sintetizza il livello complessivo dei giudizi, ma non ne rappresenta la dispersione. Un valore medio elevato può nascondere un consenso molto compatto oppure la presenza di gruppi di clienti con esperienze differenti. Questa distinzione è rilevante perché le due situazioni richiedono letture manageriali diverse. Un giudizio ampiamente condiviso segnala un'esperienza relativamente uniforme; un elevato disaccordo suggerisce invece che

esistano condizioni, segmenti di clientela o modalità di erogazione che producono esperienze differenti e che meritano quindi di essere approfondite.

L'indicatore proposto.

Per misurare il grado di accordo tra i recensori si utilizza l'entropia della distribuzione delle valutazioni. L'idea è semplice: quanto più i giudizi sono concentrati su uno o pochi livelli della scala, tanto maggiore è il consenso; quanto più sono distribuiti tra valutazioni differenti, tanto maggiore è il disaccordo. L'indicatore aggiunge quindi alla domanda "quanto siamo valutati?" una seconda informazione: "quanto i clienti sono d'accordo tra loro nel valutarci in quel modo?"

L'utilità per la decisione.

Il grado di accordo aiuta anzitutto a stabilire dove sia più utile approfondire l'analisi. Un'elevata concentrazione dei giudizi indica che le recensioni tendono a raccontare esperienze simili e che ulteriori analisi possono produrre rendimenti informativi decrescenti. Un elevato disaccordo segnala invece la presenza di una maggiore varietà di esperienze; diventa allora utile interrogare il testo delle recensioni per comprendere quali fattori distinguano i clienti soddisfatti da quelli meno soddisfatti, oppure verificare se le differenze siano associate a prodotti, canali, momenti o segmenti specifici.

L'indicatore assume quindi anche una funzione di orientamento dell'analisi successiva: non dice ancora perché i clienti siano in disaccordo, ma segnala dove valga la pena cercarne le ragioni. Né consente, da solo, di stabilire quale forma assuma il disaccordo. Valutazioni distribuite diffusamente lungo tutta la scala e giudizi concentrati soprattutto sui due estremi possono infatti produrre livelli elevati di entropia pur descrivendo fenomeni differenti. È questa la distinzione affrontata dall'indicatore successivo.

Nota tecnica.

Il grado di disaccordo viene misurato attraverso l'entropia di Shannon (1948), applicata alla distribuzione delle valutazioni sui cinque livelli:

$$H = -\sum p_k \log_2 p_k, \text{ con } p_k = n_k/n$$

dove p_k rappresenta la quota di recensioni collocata nel livello k . Su una scala a cinque livelli, l'indicatore varia da 0, quando tutte le recensioni esprimono la medesima valutazione, a $\log_2(5)$, pari a circa 2,32, quando le recensioni sono distribuite uniformemente tra tutti i livelli. Valori crescenti indicano quindi una crescente eterogeneità dei giudizi, senza specificare quale forma assuma tale eterogeneità.

3.2.3. I giudizi sono moderati o polarizzati?

La domanda manageriale.

Sapere che i clienti sono in disaccordo costituisce un primo segnale, ma non è sufficiente per comprendere la natura del fenomeno. Il disaccordo può infatti assumere forme molto differenti. I giudizi possono distribuirsi gradualmente lungo la scala oppure concentrarsi soprattutto agli estremi: clienti fortemente soddisfatti da un lato e fortemente insoddisfatti dall'altro. Per il manager la domanda diventa quindi: le opinioni dei clienti si distribuiscono lungo diverse gradazioni oppure tendono a polarizzarsi tra entusiasmo e forte insoddisfazione?

Il limite dell'indicatore precedente.

L'entropia misura quanto le valutazioni siano disperse, ma non indica dove tale dispersione si concentri. Due distribuzioni possono presentare un grado simile di disaccordo e avere, tuttavia, significati manageriali molto diversi. Una può essere caratterizzata da giudizi distribuiti soprattutto nelle posizioni intermedie; un'altra può essere dominata dalle valutazioni estreme. Nel secondo caso il problema non è semplicemente che i clienti abbiano opinioni diverse: è che la stessa marca o lo stesso prodotto sembra generare esperienze fortemente contrapposte.

L'indicatore proposto.

Per distinguere queste situazioni si misura il peso relativo delle valutazioni estreme, cioè una e cinque stelle, rispetto a quelle intermedie. Quanto maggiore è la concentrazione agli estremi, tanto più il profilo delle recensioni assume una configurazione polarizzata. L'indicatore consente così di distinguere un generico disaccordo da una situazione nella quale chi decide di recensire tende soprattutto a esprimere entusiasmo oppure forte insoddisfazione.

L'utilità per la decisione.

Una polarizzazione elevata costituisce un segnale particolarmente interessante perché suggerisce di andare oltre il dato medio e cercare che cosa separi esperienze così differenti. La stessa offerta potrebbe, ad esempio, funzionare molto bene in alcune circostanze e molto male in altre; soddisfare fortemente alcuni segmenti e deluderne altri; oppure produrre risultati differenti in funzione del canale, del prodotto o della fase del customer journey. L'indicatore non identifica direttamente quale di queste spiegazioni sia corretta, ma segnala una configurazione

che merita di essere approfondita attraverso l'analisi delle recensioni e, ove disponibili, di altre informazioni sul cliente e sull'esperienza.

Anche in questo caso il valore manageriale dell'indicatore è quindi soprattutto diagnostico. Esso non spiega la causa della polarizzazione, ma rende visibile un fenomeno che la media può occultare e che la sola misura del disaccordo non riesce a qualificare. Media, entropia e polarizzazione rispondono così a tre domande diverse e complementari: qual è il giudizio complessivo? quanto è condiviso? e quale forma assume l'eventuale disaccordo?

Nota tecnica.

La polarizzazione viene misurata attraverso il rapporto tra la quota di valutazioni collocate agli estremi della scala e quella delle valutazioni intermedie:

$$B = (p_1 + p_5)/(p_2 + p_3 + p_4)$$

dove $p_1 \dots p_5$ rappresentano le quote di recensioni associate ai cinque livelli della scala. Un valore superiore a 1 indica che, complessivamente, le valutazioni estreme sono più numerose di quelle intermedie; valori crescenti segnalano una concentrazione progressivamente maggiore agli estremi. L'indicatore va tuttavia interpretato insieme alla distribuzione delle valutazioni, poiché non distingue, da solo, una concentrazione su entrambi gli estremi da una distribuzione fortemente sbilanciata verso uno soltanto di essi.

3.3. Come si esprimono i clienti

Le valutazioni considerate nel paragrafo precedente descrivono come i clienti giudicano la marca, ma le recensioni costituiscono anche un flusso di informazioni che si sviluppa nel tempo e assume caratteristiche diverse a seconda del contesto. Può essere regolare oppure concentrarsi in particolari momenti; può inoltre accadere che le recensioni più articolate si concentrino in corrispondenza di determinati livelli di valutazione. Osservare queste caratteristiche non consente ancora di comprendere il contenuto delle recensioni, ma aiuta a stabilire quando e dove il flusso meriti un approfondimento.

3.3.1. Le recensioni arrivano regolarmente o per picchi?

La domanda manageriale.

Il numero complessivo delle recensioni dice quanto i clienti si esprimano, ma non dice come tale attività si distribuisca nel tempo. Lo stesso numero di recensioni può infatti essere prodotto attraverso un flusso relativamente costante oppure concentrarsi in brevi periodi di intensa attività. Per il manager la domanda rilevante diventa quindi: le recensioni arrivano con una certa regolarità oppure si concentrano in picchi anomali?

Il limite degli indicatori tradizionali.

Il numero totale delle recensioni, o anche la loro media mensile, tende a nascondere la struttura temporale del fenomeno. Cento recensioni distribuite uniformemente nell'arco di un anno e cento recensioni concentrate prevalentemente in pochi giorni producono lo stesso volume complessivo, ma possono avere significati molto diversi. Un improvviso addensamento può coincidere con un episodio critico, con il lancio di un prodotto, con una campagna che stimola la raccolta di recensioni o con altri eventi che modificano temporaneamente la propensione dei clienti a pubblicare. Il dato aggregato non permette di distinguere queste configurazioni.

L'indicatore proposto.

Per descrivere la regolarità del flusso si osservano gli intervalli di tempo che separano una recensione dalla successiva. Quando tali intervalli sono relativamente simili, il flusso è regolare; quando invece si alternano periodi di scarsa attività e sequenze ravvicinate di pubblicazioni, le recensioni tendono ad arrivare "a raffiche". L'indicatore di burstiness sintetizza proprio questa caratteristica, consentendo di distinguere un flusso continuo da uno caratterizzato da addensamenti temporali.

L'utilità per la decisione.

Un valore elevato di burstiness costituisce soprattutto un segnale di attenzione. Non permette, da solo, di stabilire perché si sia verificato un picco, ma indica quando sia opportuno approfondirne l'origine. L'analista può quindi concentrare la lettura delle recensioni nei periodi interessati, verificare la presenza di eventi aziendali o di mercato e valutare se l'addensamento corrisponda, ad esempio, a una crisi, a un'iniziativa commerciale o a modalità anomale di generazione delle recensioni.

La funzione dell'indicatore è dunque diagnostica: non attribuisce una causa al fenomeno, ma individua i momenti nei quali il comportamento del flusso si discosta dalla sua normale regolarità e nei quali un'analisi più approfondita può produrre maggiore informazione.

Nota tecnica.

La regolarità temporale viene misurata attraverso l'indice di burstiness, calcolato sugli intervalli τ tra recensioni consecutive:

$$\beta = (\sigma_t - \mu_t) / (\sigma_t + \mu_t)$$

dove μ_t e σ_t rappresentano rispettivamente la media e la deviazione standard degli intervalli tra una recensione e la successiva. L'indice assume valori compresi tra -1 e $+1$. Valori prossimi a -1 indicano intervalli molto regolari; valori prossimi a zero sono compatibili con un flusso casuale di tipo poissoniano; valori crescenti verso $+1$ segnalano una maggiore concentrazione delle recensioni in raffiche. Il confronto tra marche richiede tuttavia cautela quando le finestre temporali di osservazione non sono omogenee; in tali condizioni l'indicatore è più affidabile per analizzare l'evoluzione della stessa marca nel tempo che per effettuare confronti diretti tra imprese.

3.3.2. Dove si concentra il contenuto informativo?

La domanda manageriale.

Una volta stabilito che non tutte le recensioni raccontano necessariamente la stessa esperienza, resta un problema pratico: da dove conviene iniziare a leggere? In un corpus composto da migliaia o centinaia di migliaia di testi, dedicare la stessa attenzione a tutte le recensioni può risultare poco efficiente. È quindi utile capire se i contributi più articolati si concentrino prevalentemente tra i clienti molto insoddisfatti, tra quelli soddisfatti oppure nelle valutazioni intermedie.

Il limite degli indicatori tradizionali.

La distribuzione dei rating indica quanti clienti assegnano una, due o cinque stelle, ma non dice quanto contenuto testuale producano nei diversi livelli della scala. Una categoria numericamente poco rappresentata può raccogliere recensioni molto più articolate e contenere quindi una quantità di materiale utile all'analisi superiore a quella suggerita dalla sua semplice numerosità. Allo stesso modo, una quota molto ampia di valutazioni può essere costituita da testi estremamente brevi e ripetitivi.

L'indicatore proposto.

Per rendere visibile questa differenza viene costruito un profilo della lunghezza delle recensioni per ciascun livello di valutazione. Non si tratta, in questo caso, di un singolo indice sintetico, ma di una sequenza di valori che mostra quanto siano mediamente articolati i testi associati a una, due, tre, quattro e cinque stelle. La lunghezza non misura direttamente la qualità o la rilevanza del contenuto,

ma costituisce un indicatore elementare della densità testuale disponibile nei diversi segmenti della distribuzione.

L'utilità per la decisione.

Il profilo consente di orientare l'analisi successiva verso le aree nelle quali si concentra il materiale testuale più ricco. Se, ad esempio, le recensioni negative risultano significativamente più lunghe delle positive, può essere ragionevole iniziare da esse per comprendere in dettaglio le cause dell'insoddisfazione; se invece i testi più articolati si concentrano nelle valutazioni intermedie, proprio quella parte apparentemente meno "estrema" della distribuzione può rivelarsi più utile per individuare punti di forza e criticità.

L'indicatore non stabilisce quindi quali recensioni siano più importanti, né può sostituire la loro lettura. Serve piuttosto a decidere dove concentrare per prima l'attenzione analitica. In questo senso costituisce un naturale punto di raccordo con l'impiego dell'intelligenza artificiale discusso nel seguito del lavoro: gli indicatori strutturati individuano le porzioni del corpus potenzialmente più promettenti e l'analisi semantica può essere successivamente indirizzata su di esse.

Un'ulteriore cautela riguarda la trasferibilità dei risultati. La relazione tra lunghezza del testo e valutazione può dipendere dalla piattaforma, dalla categoria merceologica e dalle modalità con cui la recensione viene sollecitata e pubblicata. Il profilo va quindi costruito sul contesto specifico e non trasformato in una regola generale valida per mercati o piattaforme differenti.

Nota tecnica.

Per ciascun livello della scala viene calcolata la lunghezza mediana delle recensioni, ottenendo il vettore:

$$L = (L_1, L_2, L_3, L_4, L_5)$$

Dove L_k rappresenta la lunghezza mediana dei testi associati alla valutazione k . L'impiego della mediana riduce l'influenza di recensioni eccezionalmente lunghe e rende il confronto più robusto rispetto alla media aritmetica. Il risultato non costituisce un indice sintetico, ma un profilo da interpretare all'interno del singolo dataset e del relativo contesto di piattaforma.

3.4. Come risponde l'impresa

Fin qui l'analisi ha riguardato ciò che i clienti producono nello spazio pubblico delle recensioni, ovvero come valutano la marca, quanto concordano tra loro, come si distribuiscono i giudizi e con quale andamento temporale vengono pubblicati. La presenza di una risposta aziendale introduce una prospettiva diversa. La recensione non è più soltanto una manifestazione dell'esperienza del cliente, ma può diventare l'avvio di un'interazione pubblica nella quale anche l'impresa sceglie se intervenire, a chi rivolgersi e con quale tempestività. È attraverso questi comportamenti osservabili che il sistema proposto affronta più direttamente la dimensione relazionale.

3.4.1. A quante recensioni risponde?

La domanda manageriale.

La prima domanda è la più semplice: quanto è presente l'impresa nello spazio pubblico delle recensioni? Due marche possono ricevere un numero simile di giudizi ma adottare comportamenti molto diversi: una può limitarsi a osservare ciò che viene pubblicato, mentre l'altra può rispondere sistematicamente ai clienti che hanno scelto di esprimersi. Il tasso di risposta fornisce una prima misura di questa presenza.

Il limite della sola osservazione delle valutazioni.

Rating, distribuzioni e volumi di recensioni descrivono esclusivamente il comportamento dei clienti. Non permettono di sapere se l'impresa consideri la piattaforma come un semplice luogo di raccolta del feedback oppure come uno spazio nel quale intervenire attivamente. Anche un numero elevato di recensioni, quindi, può convivere con una sostanziale assenza dell'impresa dalla conversazione pubblica.

L'indicatore proposto.

Il tasso di risposta misura la quota di recensioni alle quali l'impresa fornisce un riscontro pubblico. Si tratta dell'indicatore relazionale più elementare del sistema: non dice ancora a chi venga data risposta né quanto rapidamente, ma rende osservabile il grado di copertura della politica di risposta. Un valore elevato indica una presenza diffusa; un valore basso segnala invece che l'intervento dell'impresa è selettivo o occasionale; un valore nullo identifica le imprese che non utilizzano la risposta pubblica come modalità di interazione sulla piattaforma considerata.

L'utilità per la decisione.

Il tasso di risposta consente di confrontare la propria presenza con quella dei concorrenti e di trasformare una generica intenzione di “presidiare le recensioni” in un obiettivo misurabile. Può inoltre essere monitorato nel tempo per verificare se l’introduzione di nuove procedure organizzative o strumenti di automazione produca un effettivo aumento della copertura.

Da solo, tuttavia, l’indicatore può essere fuorviante. Due imprese che rispondono alla stessa percentuale di recensioni possono adottare strategie profondamente diverse se una concentra l’attenzione sui clienti insoddisfatti e l’altra distribuisce le risposte indipendentemente dal segno della valutazione. Per comprendere quale logica governi la risposta è quindi necessario scomporre il tasso complessivo.

Nota tecnica.

Il tasso complessivo di risposta è definito come:

$$\rho = \#\{i : a_i \neq \emptyset\}/n$$

dove $a_i \neq \emptyset$ indica la presenza di una risposta aziendale alla recensione i -esima e n il numero complessivo di recensioni per le quali il campo relativo alla risposta è disponibile. L’indicatore varia quindi tra 0, in assenza di qualunque risposta, e 1, quando tutte le recensioni ricevono un riscontro.

3.4.2. A chi sceglie di rispondere?

La domanda manageriale.

Sapere quante recensioni ricevono una risposta non dice ancora come l’impresa distribuisca la propria attenzione. Un’azienda può intervenire soprattutto quando il cliente manifesta forte insoddisfazione, può rispondere sistematicamente indipendentemente dal giudizio espresso oppure, al contrario, può essere più presente nei confronti di chi formula valutazioni positive. La domanda rilevante diventa quindi: a chi scegliamo di rispondere?

Il limite del tasso complessivo di risposta.

Il tasso aggregato può nascondere strategie relazionali molto differenti. Due imprese possono, ad esempio, rispondere entrambe alla metà delle recensioni, ma la prima può concentrare quasi tutti i propri interventi sulle valutazioni negative, mentre la seconda può distribuire le risposte in maniera equilibrata tra feedback positivi e negativi. Dal punto di vista quantitativo il tasso complessivo è identico; dal punto di vista manageriale, il comportamento osservato è profondamente diverso.

L’indicatore proposto.

Per rendere visibile questa differenza si calcolano separatamente il tasso di risposta alle recensioni negative e quello alle recensioni positive e se ne considera la differenza. L’asimmetria di

risposta indica quindi verso quale parte della distribuzione l'impresa concentra maggiormente la propria attenzione.

Valori fortemente positivi sono coerenti con una strategia di contenimento, nella quale la risposta viene utilizzata soprattutto nei confronti di chi manifesta insoddisfazione. Valori prossimi allo zero, se associati a un tasso complessivo elevato, descrivono una strategia di presenza, nella quale l'impresa interviene indipendentemente dal segno della valutazione. Valori fortemente negativi sono invece coerenti con una strategia di vetrina, caratterizzata da una maggiore attenzione verso i giudizi positivi.

Queste etichette descrivono configurazioni osservabili e non intendono attribuire automaticamente all'impresa una specifica intenzione strategica: la causa del comportamento rilevato richiede informazioni ulteriori.

L'utilità per la decisione.

L'asimmetria rende il tasso di risposta una vera variabile diagnostica. Un'impresa può così verificare non soltanto quanto stia rispondendo, ma se la distribuzione delle risorse dedicate alle recensioni sia coerente con il ruolo che intende attribuire alla relazione pubblica con i clienti. Una forte concentrazione sulle recensioni negative può essere appropriata quando la risposta è concepita prevalentemente come strumento di service recovery; può invece segnalare un'opportunità non colta se l'obiettivo è riconoscere e alimentare anche il contributo dei clienti soddisfatti.

Il valore dell'indicatore è quindi soprattutto comparativo: consente di osservare differenze tra imprese che un semplice tasso aggregato non mostra e di verificare nel tempo se una modifica della politica di risposta produca il riequilibrio desiderato.

Nota tecnica.

Si definiscono anzitutto i tassi di risposta condizionati al segno della valutazione:

$$\rho_{neg} = \#\{i : r_i \leq 2 \wedge a_i \neq \emptyset\} / \#\{i : r_i \leq 2\}$$

$$\rho_{pos} = \#\{i : r_i \geq 4 \wedge a_i \neq \emptyset\} / \#\{i : r_i \geq 4\}$$

dove r_i rappresenta la valutazione associata alla recensione e $a_i \neq \emptyset$ la presenza di una risposta aziendale. Le valutazioni pari a 1 o 2 sono considerate negative, quelle pari a 4 o 5 positive; le valutazioni intermedie non entrano nel calcolo dell'asimmetria.

L'indice è quindi definito come:

$$\alpha = \rho_{neg} - \rho_{pos}$$

Valori positivi indicano una maggiore propensione a rispondere alle recensioni negative, valori negativi una maggiore propensione a rispondere alle positive, valori prossimi allo zero un comportamento sostanzialmente simmetrico.

Poiché α è una differenza tra due tassi, un'impresa che non risponde ad alcuna recensione produrrebbe formalmente lo stesso valore di un'impresa che risponde alle recensioni positive e negative nella stessa proporzione. Per questa ragione l'asimmetria deve essere interpretata congiuntamente al tasso complessivo di risposta; le imprese totalmente non rispondenti costituiscono una categoria distinta.

3.4.3 Quanto rapidamente risponde?

La domanda manageriale.

La presenza dell'impresa non dipende soltanto dal fatto che una risposta venga fornita, ma anche dal momento in cui essa arriva. Una risposta pubblicata poche ore dopo una recensione negativa e una pubblicata diversi giorni più tardi costituiscono formalmente due risposte, ma possono assumere un significato molto diverso per il cliente. Alla domanda “rispondiamo?” si aggiunge quindi una terza questione: “quanto rapidamente lo facciamo?”

Il limite del tasso di risposta.

Il tasso di risposta considera equivalenti tutti i riscontri, indipendentemente dal tempo trascorso dalla pubblicazione della recensione. Un'impresa può quindi presentare una copertura molto elevata e apparire particolarmente presente, pur rispondendo sistematicamente con forte ritardo. In questo caso l'indicatore di copertura sopravvaluterebbe la capacità dell'impresa di intervenire tempestivamente nella relazione con il cliente.

L'indicatore proposto.

La latenza di risposta misura il tempo che intercorre tra la pubblicazione della recensione e il riscontro dell'impresa. Non interessa soltanto sapere se l'azienda interviene, ma quanto rapidamente riesca a trasformare il feedback pubblico in un'occasione di interazione.

La tempestività assume particolare rilievo nel caso delle recensioni negative: quanto più la risposta è vicina all'episodio che ha generato l'insoddisfazione, tanto maggiori sono, almeno potenzialmente, le possibilità che essa svolga una funzione di recupero. Una risposta molto tardiva può continuare ad avere valore reputazionale nei confronti di chi leggerà successivamente la recensione, ma riduce la possibilità di intervenire tempestivamente sull'esperienza del cliente che l'ha pubblicata.

L'utilità per la decisione.

La latenza presenta un vantaggio manageriale particolarmente immediato: può essere tradotta direttamente in un obiettivo operativo. L'indicazione generica di “rispondere più rapidamente alle

recensioni” può diventare, ad esempio, un tempo mediano obiettivo o uno standard differenziato per tipologia di feedback.

Considerata congiuntamente all’asimmetria, la latenza consente inoltre di distinguere configurazioni diverse. Una risposta rapida e distribuita tra feedback positivi e negativi segnala una presenza relazionale diffusa; una risposta rapida concentrata sui giudizi negativi è coerente con una gestione orientata al contenimento e al service recovery; una risposta sistematicamente tardiva segnala invece che l’esistenza formale del riscontro non coincide necessariamente con una reale capacità di intervento tempestivo.

Nota tecnica.

Quando la piattaforma rende disponibile anche la data della risposta, la latenza viene calcolata come differenza tra il momento del riscontro aziendale e quello della pubblicazione della recensione:

$$\Delta t = \text{mediana}[t(a_i) - t(r_i)]$$

Dove $t(r_i)$ indica il momento di pubblicazione della recensione e $t(a_i)$ quello della relativa risposta. Si utilizza la mediana anziché la media perché la distribuzione dei tempi di risposta può essere fortemente asimmetrica e contenere valori estremi molto elevati. L’indicatore può inoltre essere calcolato separatamente per recensioni positive e negative quando la numerosità delle osservazioni lo consente.

4. Dati e metodologia di analisi

4.1. I dataset

L'applicazione degli indicatori viene illustrata attraverso due dataset molto diversi per dimensione, piattaforma, mercato e natura delle recensioni. Come anticipato, questa eterogeneità è intenzionale, in quanto l'obiettivo non è costruire un campione rappresentativo dell'universo delle recensioni online, ma mostrare come il sistema di indicatori possa essere applicato in contesti differenti e quali cautele interpretative emergano al variare dei dati disponibili.

Il primo dataset è stato costruito mediante estrazione automatica delle recensioni pubblicamente disponibili su una piattaforma generalista italiana ([trustpilot.it](https://www.trustpilot.it)) e comprende undici marche appartenenti a cinque settori: servizi bancari e di pagamento, distribuzione beauty, gioielleria online, moda online e distribuzione culturale. Per ciascuna marca sono state acquisite, secondo l'ordinamento proposto dalla piattaforma, fino a un massimo di duecento recensioni. Dopo le operazioni di controllo e consolidamento illustrate nel paragrafo successivo, il dataset comprende complessivamente 2.142 recensioni uniche.

Il secondo dataset comprende 701.528 recensioni di prodotti della categoria beauty pubblicate su una piattaforma di commercio elettronico statunitense ([amazon.com](https://www.amazon.com)) tra il 2003 e il 2023. Per ciascuna recensione sono disponibili, oltre alla valutazione e al testo, l'identificativo del prodotto, la data di pubblicazione, il numero di voti di utilità ricevuti e l'indicazione relativa all'acquisto verificato. Quando l'analisi viene condotta a livello di singolo prodotto, sono considerati i 744 prodotti che dispongono di almeno cento recensioni, così da evitare che i confronti siano dominati da prodotti caratterizzati da una base informativa troppo limitata.

I due dataset svolgono quindi funzioni complementari. Il primo consente di osservare anche il comportamento di risposta delle imprese e di confrontare marche appartenenti a settori differenti; il secondo permette di verificare il comportamento degli indicatori valutativi e testuali su una base empirica di dimensione molto maggiore e in un contesto di piattaforma differente.

Tab. 1 — Composizione del dataset italiano

Settore	Marche	Recensioni
Servizi bancari e di pagamento	6	1.200
Distribuzione beauty	1	200
Gioielleria online	2	324
Moda online	1	199
Distribuzione culturale	1	219
Totale	11	2.142

4.2. Preparazione dei dati e condizioni di validità

Prima del calcolo degli indicatori è stato necessario verificare che i dati estratti potessero essere effettivamente interpretati nel modo previsto. Alcuni controlli possono apparire come semplici operazioni di pulizia del dataset, ma hanno in realtà un effetto diretto sulla validità dei risultati, in quanto un errore in questa fase può produrre indicatori formalmente corretti, ma riferiti a dati incompleti o conteggiati più volte.

Un primo problema riguarda la presenza di recensioni duplicate. Nel file relativo a una delle marche, ad esempio, le 1.200 righe estratte corrispondevano in realtà a sole 219 recensioni distinte, poiché alcuni record erano stati ripetuti fino a sette volte dalla procedura automatica di acquisizione. Se tali duplicazioni non fossero state riconosciute, quella marca avrebbe assunto nel dataset un peso molto superiore a quello reale, alterando non soltanto i suoi indicatori, ma anche le eventuali statistiche aggregate. Per questa ragione, prima di ogni elaborazione, le recensioni sono state ricondotte al loro identificativo univoco nativo e ogni duplicato è stato eliminato. Il punto ha una portata più generale: quando i dati vengono raccolti automaticamente, il numero di righe ottenute non coincide necessariamente con il numero di osservazioni effettivamente disponibili.

Un secondo problema riguarda la distinzione tra un valore assente e un'informazione non disponibile. I dati estratti per le undici marche non presentavano infatti una struttura perfettamente omogenea. Per due marche mancava il campo che indica la presenza di una risposta aziendale. Questo non significa che le imprese non abbiano risposto alle recensioni, ma soltanto che l'informazione non era stata rilevata nel processo di estrazione. Attribuire in questi casi un tasso di risposta pari a zero significherebbe trasformare un'informazione mancante in un comportamento dell'impresa che i dati non consentono di osservare. Gli indicatori relativi

alla risposta vengono pertanto considerati non disponibili per queste marche e segnalati come tali nelle tabelle dei risultati.

È stata infine verificata l'uniformità delle date di pubblicazione. I diversi tracciati utilizzano infatti formati non sempre omogenei. Dove disponibile, è stato utilizzato il riferimento temporale espresso in secondi (epoch time), verificandone la corrispondenza con la data riportata in forma testuale. Una particolare attenzione è stata riservata alla convenzione statunitense mese-giorno-anno: interpretarla secondo l'ordine giorno-mese-anno avrebbe potuto modificare la corretta sequenza temporale delle recensioni e, di conseguenza, gli indicatori che dipendono dagli intervalli tra una pubblicazione e la successiva.

Questi controlli costituiscono quindi una condizione preliminare dell'analisi, per cui prima di chiedersi che cosa misurino gli indicatori, occorre assicurarsi che le osservazioni sulle quali vengono calcolati siano uniche, correttamente interpretate e realmente disponibili.

4.3. Soglie di interpretabilità

Alcuni indicatori vengono calcolati non sull'intero insieme delle recensioni, ma su gruppi più circoscritti. È il caso, ad esempio, del tasso di risposta alle sole valutazioni negative. In queste situazioni il numero complessivo di recensioni di una marca può essere relativamente elevato, mentre il sottoinsieme sul quale viene calcolato l'indicatore può risultare molto piccolo.

Un valore ottenuto su poche osservazioni è comunque matematicamente calcolabile, ma può essere fortemente influenzato da uno o due singoli casi e risultare quindi poco utile ai fini del confronto. Per evitare di attribuire eccessivo significato a valori instabili, nel presente lavoro gli indicatori riferiti alle sole recensioni negative vengono riportati soltanto quando sono disponibili almeno trenta osservazioni appartenenti a tale categoria.

La soglia non va interpretata come un valore universale oltre il quale un indicatore diventa automaticamente affidabile, ma come un criterio prudenziale di interpretabilità adottato nell'ambito di questa applicazione. Quando il numero di osservazioni non raggiunge il livello stabilito, il valore non viene utilizzato nei confronti tra marche e viene indicato come non disponibile.

5. Risultati

Coerentemente con la finalità prevalentemente dimostrativa dell'analisi empirica, i risultati vengono letti soprattutto in funzione di ciò che i diversi indicatori consentono di rendere visibile rispetto alla sola osservazione della valutazione media. La presentazione segue i tre gruppi introdotti nel paragrafo 3: come valutano i clienti, come si esprimono e come risponde l'impresa.

5.1. Come valutano i clienti

La Tabella 2 riporta gli indicatori relativi alle undici marche del dataset italiano. Il primo risultato evidente riguarda il diverso grado di concordanza tra i clienti. Infatti, l'entropia varia da 0,56 a 1,84, su un massimo teorico di 2,32: dietro valutazioni medie apparentemente confrontabili possono quindi trovarsi situazioni molto diverse.

Tab. 2 — Come valutano i clienti: dataset italiano

Marca	N	RB	H	B
BBVA	200	3,73	1,84	3,00
IBL Banca	200	4,68	0,75	9,53
ING	200	3,78	1,34	10,76
Santander Consumer	200	3,84	1,55	5,45
Satispay	200	3,91	1,42	7,70
Widiba	200	4,61	0,83	9,53
Sephora	200	3,37	1,44	10,76
Lapilli Shop	200	4,75	0,65	10,76
Sogni Gioielli	124	4,70	0,56	16,71
Yoox	199	4,38	1,13	6,65
La Feltrinelli	219	4,41	1,18	5,84

Le due marche di gioielleria online, Lapilli Shop e Sogni Gioielli, presentano valori di entropia inferiori a 0,7 e valutazioni medie corrette rispettivamente pari a 4,75 e 4,70. In questi casi il giudizio dei clienti è non soltanto elevato, ma anche relativamente omogeneo. La media

descrive quindi abbastanza bene la direzione prevalente delle valutazioni, ma proprio l'elevato consenso ne riduce la capacità di differenziare marche che presentano configurazioni analoghe.

Una situazione diversa emerge, ad esempio, per BBVA, che presenta una valutazione media corretta di 3,73 e un'entropia pari a 1,84, la più elevata del dataset. In questo caso il valore medio sintetizza esperienze molto più eterogenee e rende particolarmente utile approfondire che cosa distingua i clienti che valutano positivamente la marca da quelli che ne hanno un'esperienza negativa. L'entropia non fornisce questa spiegazione, ma segnala dove cercarla.

Il dataset statunitense consente di osservare lo stesso problema su una scala molto più ampia. Sull'insieme delle 701.528 recensioni la valutazione media è pari a 3,96 e l'entropia a 1,74. Considerando i 744 prodotti con almeno cento recensioni, emerge una forte associazione inversa tra valutazione media ed entropia ($r = -0,82$): passando dal quartile dei prodotti con valutazioni più basse a quello con valutazioni più elevate, l'entropia media diminuisce da 1,99 a 0,97.

Tab. 3 — Concordanza e concentrazione agli estremi per quartile di valutazione media

Quartile di valutazione media	H medio	B medio
Primo quartile (più basso)	1,99	2,29
Secondo quartile	1,72	2,77
Terzo quartile	1,42	3,88
Quarto quartile (più alto)	0,97	7,21

Contemporaneamente aumenta il peso delle valutazioni estreme rispetto a quelle intermedie, in quanto l'indicatore B passa da 2,29 nel primo quartile a 7,21 nel quarto. I due risultati vanno letti congiuntamente. Nei prodotti meglio valutati, la crescita di B non indica necessariamente la presenza di due gruppi contrapposti, ma riflette soprattutto una crescente concentrazione verso l'estremo positivo della scala, coerente con la contemporanea riduzione dell'entropia.

L'utilità manageriale emerge proprio dalla combinazione dei diversi indicatori: la media dice quanto positivamente viene valutato un prodotto; l'entropia dice quanto quel giudizio è condiviso; la concentrazione agli estremi aiuta a comprendere quale forma assume la distribuzione. Nessuna delle tre informazioni, isolatamente, restituisce lo stesso quadro.

5.2. Come si esprimono i clienti

Gli indicatori relativi al flusso delle recensioni aggiungono una seconda prospettiva: non riguardano più soltanto il giudizio espresso, ma anche quando e quanto i clienti scrivono.

Per quanto riguarda la distribuzione temporale, l'indice di burstiness presenta valori compresi tra -0,01 e 0,67 nel dataset italiano. La maggior parte delle marche mostra valori relativamente contenuti, mentre La Feltrinelli presenta il valore più elevato (0,67), segnalando una maggiore tendenza delle recensioni a concentrarsi in raffiche anziché distribuirsi regolarmente nel tempo. Il dato non consente di stabilirne la causa, che potrebbe dipendere da eventi, campagne, episodi critici o altre circostanze, ma identifica un periodo o una configurazione sulla quale concentrare un eventuale approfondimento.

Il confronto diretto tra marche deve tuttavia essere effettuato con cautela, poiché le finestre temporali osservate non sono perfettamente allineate. In questa applicazione la burstiness va quindi considerata soprattutto come indicatore utile a segnalare anomalie e a monitorare nel tempo la stessa marca, piuttosto che come base per costruire un ranking tra concorrenti.

Tab. 4 — indice di burstiness

Marca	β
BBVA	0,31
IBL Banca	0,33
ING	0,28
Santander Consumer	0,39
Satispay	0,07
Widiba	0,16
Sephora	-0,01
Lapilli Shop	0,18
Sogni Gioielli	0,13
Yoox	0,23
La Feltrinelli	0,67

Una seconda indicazione deriva dalla lunghezza delle recensioni. Come mostra la Tabella 5, il rapporto tra lunghezza del testo e valutazione assume configurazioni molto diverse nei due dataset.

Tab. 5 — Lunghezza mediana del testo per livello di valutazione

Dataset	1	2	3	4	5
Italiano (n = 2.142)	404	222	226	97	89
Statunitense (n = 701.528)	105	124	127	130	91

Nel dataset italiano le recensioni diventano progressivamente più brevi al crescere della valutazione: la mediana passa da 404 battute per le recensioni a una stella a 89 per quelle a cinque stelle. Le valutazioni fortemente negative contengono quindi, mediamente, una quantità di testo più di quattro volte superiore rispetto alle più positive.

Nel dataset statunitense il profilo è differente, in quanto la lunghezza cresce dalle valutazioni più basse fino alle quattro stelle, dove raggiunge il massimo di 130 battute, per poi diminuire a 91 nelle recensioni a cinque stelle. Non emerge quindi una regola generale secondo la quale il cliente insoddisfatto scriva sempre di più.

Una possibile interpretazione, da considerare come ipotesi e non come risultato dimostrato, riguarda la diversa natura delle piattaforme. In una piattaforma generalista la recensione può essere un atto deliberato, talvolta originato da un episodio problematico che il cliente sente la necessità di raccontare in dettaglio; in una piattaforma di commercio elettronico la recensione è invece inserita nel normale flusso successivo all'acquisto e può assumere maggiormente la funzione di consiglio rivolto ad altri consumatori.

Al di là della spiegazione, il risultato è metodologicamente rilevante e mostra come il profilo di lunghezza non possa essere trasferito automaticamente da una piattaforma a un'altra. Se lo scopo è utilizzare la lunghezza per orientare la successiva analisi del testo, il profilo deve essere costruito sullo specifico contesto osservato.

5.3. Come risponde l'impresa

La scomposizione delle risposte aziendali fornisce il risultato più direttamente collegato alla dimensione relazionale del lavoro. La Tabella 6 mostra che imprese apparentemente simili per presenza sulle piattaforme possono adottare comportamenti molto diversi a seconda del segno della recensione.

Tab. 6 — Come risponde l'impresa: dataset italiano

Marca	p_{neg}	p_{pos}	α	Δt
BBVA	94,4	100,0	-5,6	0,4
IBL Banca	75,0	99,5	-24,5	42,8
ING	96,7	97,8	-1,1	18,0
Santander Consumer	0,0	0,0	n.d.	—
Satispay	94,2	0,0	+94,2	120,9
Widiba	83,3	87,0	-3,7	21,3
Sephora	n.r.	n.r.	n.r.	—
Lapilli Shop	100,0	8,4	+91,6	106,4
Sogni Gioielli	100,0	100,0	0,0	83,1
Yoox	69,6	5,2	+64,4	—
La Feltrinelli	24,0	0,0	+24,0	261,1

Quattro delle nove marche per le quali l'asimmetria è interpretabile presentano un valore positivo: Satispay risponde al 94,2% delle recensioni negative e a nessuna di quelle positive; Lapilli Shop al 100% delle negative contro l'8,4% delle positive; Yoox al 69,6% contro il 5,2%; La Feltrinelli al 24% contro lo 0%. Sono configurazioni nelle quali l'attenzione dell'impresa si concentra quindi prevalentemente sul dissenso.

Tra le altre marche emergono comportamenti differenti. BBVA, ING, Widiba e Sogni Gioielli presentano tassi relativamente equilibrati tra recensioni positive e negative, mentre IBL Banca mostra un'asimmetria di segno opposto (-24,5), rispondendo più frequentemente ai giudizi positivi. Santander Consumer non risponde alle recensioni considerate, mentre per Sephora l'informazione relativa alla risposta non è disponibile.

L'aspetto più interessante è che configurazioni molto diverse convivono anche all'interno dello stesso settore. Nei servizi bancari e di pagamento si osservano imprese con risposta

sostanzialmente simmetrica, un'impresa maggiormente orientata alle recensioni positive, Satispay con un'asimmetria di +94,2 e Santander Consumer senza risposte. Nel dataset analizzato, quindi, la scelta di chi riceve una risposta non appare meccanicamente determinata dal settore, ma presenta un'importante componente riconducibile alle modalità con cui la singola impresa gestisce questo spazio di relazione.

Anche la tempestività differenzia fortemente i comportamenti osservati. La latenza mediana varia da 0,4 ore per BBVA a 261,1 ore per La Feltrinelli, cioè poco meno di undici giorni. Tra questi estremi si collocano configurazioni molto diverse, che comprendono imprese che rispondono diffusamente e rapidamente, e altre che privilegiano le recensioni negative ma intervengono dopo diversi giorni.

Asimmetria e latenza vanno quindi lette congiuntamente. Una forte propensione a rispondere alle recensioni negative non implica necessariamente una gestione tempestiva del cliente insoddisfatto: Satispay e Lapilli Shop, ad esempio, mostrano asimmetrie molto elevate, ma latenze mediane superiori alle cento ore. La combinazione dei due indicatori consente pertanto di distinguere a chi l'impresa dedica attenzione da quanto rapidamente riesce a trasformare tale attenzione in una risposta.

5.4. Rispondere di più significa avere clienti più concordi?

Un'ultima analisi esplorativa verifica se una maggiore presenza dell'impresa nelle recensioni sia associata a un maggiore accordo tra i clienti. L'ipotesi è plausibile alla luce della letteratura sul service recovery: se la risposta contribuisce a gestire efficacemente le esperienze negative, imprese più reattive potrebbero presentare nel tempo valutazioni meno eterogenee.

Nel dataset italiano, tuttavia, questa relazione non emerge. Per le marche sulle quali entrambi gli indicatori sono disponibili, la correlazione tra tasso complessivo di risposta ed entropia è debole: -0,14 secondo Pearson e -0,22 secondo Spearman. Con sole dieci osservazioni, entrambe le associazioni risultano però statisticamente non significative e non consentono conclusioni inferenziali.

Il risultato va quindi interpretato con prudenza e non dimostra che la risposta dell'impresa non produca effetti sulle valutazioni, né consente di stabilire relazioni causali. Mostra semplicemente che, in questa applicazione, rispondere più frequentemente non coincide automaticamente con un maggiore consenso tra i recensori. È un'indicazione coerente

con l'idea che gli indicatori relativi alle valutazioni dei clienti e quelli relativi al comportamento dell'impresa osservino dimensioni differenti e debbano essere letti congiuntamente, anziché considerati come misure alternative dello stesso fenomeno.

6. Il ruolo dell'intelligenza artificiale

Il calcolo degli indicatori presentati nei paragrafi precedenti non richiede l'impiego di alcun modello linguistico. Ciò non significa che l'intelligenza artificiale sia estranea al processo analitico, ma che il suo ruolo debba essere definito con precisione. In questa prospettiva si individuano tre possibili collocazioni, distinte ma complementari, accomunate da un principio: il modello opera come strumento di supporto e non come fonte diretta della misura pubblicata.

6.1. Collocazione “a monte”: validazione dell'integrità del dataset

Ognuno degli indicatori illustrati in precedenza poggia su un'assunzione implicita, cioè che le recensioni rappresentino espressioni autentiche di esperienza. Se una quota rilevante delle valutazioni massime fosse artificiale o priva di contenuto esperienziale, l'entropia non misurerebbe soltanto il consenso effettivo dei clienti, e anche la valutazione media corretta ne risulterebbe distorta.

È il punto nel quale l'aritmetica, da sola, non basta. Alcuni segnali elementari possono essere rilevati attraverso regole deterministiche e già consentono di dimensionare il fenomeno. Nel dataset italiano, 268 recensioni presentano un titolo identico al testo e 85 hanno una lunghezza non superiore a quindici battute; complessivamente si tratta di 322 unità, pari al 15,0 per cento del dataset. Queste recensioni presentano una valutazione media di 4,77, contro 4,18 dell'intero dataset, e nell'86,9 per cento dei casi attribuiscono la valutazione massima.

Il dato non dimostra l'esistenza di una manipolazione, in quanto una recensione molto breve può dipendere semplicemente dal limitato sforzo dedicato alla sua compilazione. Segnala però che le unità scarsamente informative non sono distribuite uniformemente lungo la scala,

ma si concentrano soprattutto sulle valutazioni più elevate e possono quindi alterare gli indicatori calcolati sull'intero dataset. Inoltre, nessun testo superiore alle venti battute risulta duplicato: l'eventuale anomalia sembra quindi riguardare più la forma e la scarsa specificità del contenuto che la semplice ripetizione letterale, limite che i controlli deterministici faticano a intercettare.

Un possibile compito "a monte" del modello linguistico consiste allora nel supportare la verifica delle condizioni di validità del dataset: riconoscere testi templatizzati, recensioni prive di riferimenti esperienziali specifici o formulazioni fortemente ricorrenti associate a identità differenti. Il protocollo potrebbe classificare le recensioni, ad esempio, come "esperienziali", "non informative" o "sospette". La decisione finale sull'inclusione resterebbe comunque dell'analista: il modello segnala, non decide. Per preservare la riproducibilità, sia i criteri di classificazione sia il dataset utilizzato per il calcolo degli indicatori dovrebbero essere esplicitati e resi disponibili.

6.2. Collocazione "a valle": la lettura orientata

La seconda collocazione discende direttamente dagli indicatori relativi a come valutano i clienti. L'entropia può infatti funzionare come criterio di selezione, indicando dove l'analisi del testo abbia maggiori probabilità di produrre nuova informazione.

Quando il consenso è molto compatto, leggere un grande numero di recensioni può significare ritrovare molte volte indicazioni simili; l'analisi automatica rende questa lettura più rapida, ma non modifica la natura dell'informazione disponibile. Quando invece il disaccordo è elevato, coesistono esperienze diverse e diventa più interessante comprendere, attraverso il testo, che cosa le distingue. Nel dataset italiano l'entropia varia da 0,56 a 1,84, mostrando come la potenziale utilità dell'approfondimento testuale possa essere molto diversa da una marca all'altra.

L'antagonista della tesi qui sostenuta non è dunque l'intelligenza artificiale, ma il suo impiego indiscriminato. Gli indicatori orientano la selezione e consentono di concentrare l'analisi semantica sulle aree del dataset che presentano maggiore eterogeneità o anomalie meritevoli di approfondimento. Il vantaggio non è soltanto economico, ma riguarda soprattutto il rapporto tra segnale e rumore. È in questo senso che la scelta di che cosa misurare precede, per importanza, la capacità di elaborare grandi quantità di testo.

6.3. Collocazione “a monte”: validazione dell'integrità del dataset

La terza collocazione riguarda l'azione anziché l'analisi. Gli indicatori non si limitano infatti a descrivere i comportamenti osservati, ma possono far emergere vuoti operativi. Nel dataset italiano, ad esempio, quattro marche concentrano fortemente le proprie risposte sui clienti insoddisfatti, lasciando senza riscontro gran parte di quelli soddisfatti.

In condizioni di risorse limitate, una simile scelta può rispondere a una logica di efficienza, per la quale, se non è possibile rispondere a tutti, appare razionale privilegiare i casi critici. La disponibilità di strumenti di generazione assistita modifica però almeno in parte questo vincolo, rendendo più sostenibile una maggiore copertura delle recensioni e una riduzione dei tempi di risposta. L'indicatore consente quindi di individuare uno spazio di intervento che la tecnologia può contribuire a rendere più facilmente gestibile.

Tuttavia, una risposta automatizzata che si limiti a riprodurre formule standard di cortesia rischia di ridurre il costo senza accrescere il valore relazionale dell'interazione. La leva diventa realmente utile quando restituisce al cliente l'evidenza che il contenuto della recensione è stato considerato e produce quindi un riscontro coerente e specifico. È questo, più della sola riduzione dei costi, il criterio con cui valutare l'impiego dell'intelligenza artificiale nella gestione operativa delle recensioni.

6.4. Il limite invariante

In tutte e tre le collocazioni il modello linguistico mantiene una funzione strumentale. Se una misura pubblicata dipendesse direttamente da un giudizio semantico del modello, sarebbe necessario verificarne la stabilità attraverso ripetizioni indipendenti e misure di concordanza esplicite. Finché, invece, gli indicatori restano calcolati su dati osservabili mediante regole dichiarate, i risultati possono essere ricalcolati da chiunque disponga degli stessi dati e applichi la medesima procedura.

L'intelligenza artificiale può quindi intervenire a monte, contribuendo a verificare la qualità del materiale analizzato; a valle, orientando e approfondendo la lettura del testo; e in

esecuzione, supportando la risposta dell'impresa. In nessuno di questi casi sostituisce però la scelta delle dimensioni da osservare, la definizione delle misure e l'interpretazione manageriale dei risultati.

7. Implicazioni per le decisioni di marketing

Gli indicatori proposti non hanno soltanto una funzione descrittiva. Letti congiuntamente, possono orientare sia alcune scelte di carattere strategico sia decisioni più direttamente operative. Le indicazioni che seguono vanno tuttavia interpretate in relazione al mercato, alla piattaforma e all'insieme competitivo considerato, evitando di trasformare i valori osservati nei due dataset in soglie valide in assoluto.

7.1. Livello strategico

Un primo risultato riguarda il significato del consenso. Una valutazione media elevata rappresenta certamente un segnale positivo, ma perde parte della propria capacità di differenziazione quando caratterizza in modo uniforme tutte le imprese di un determinato contesto. Nel dataset italiano, ad esempio, le due marche della gioielleria online presentano contemporaneamente valutazioni superiori a 4,7 e livelli di entropia inferiori a 0,7. In una situazione di questo tipo, costruire il posizionamento comunicativo sulla generica eccellenza del servizio rischia di enfatizzare un attributo sul quale i concorrenti appaiono altrettanto forti. Gli indicatori suggeriscono quindi di distinguere tra un risultato positivo per l'impresa e un elemento effettivamente discriminante sul piano competitivo.

Una maggiore eterogeneità dei giudizi apre, per converso, uno spazio di analisi più interessante. Un'entropia elevata non indica di per sé un problema né identifica le sue cause, ma segnala che clienti diversi stanno vivendo o valutando l'esperienza in modo differente. È in queste situazioni che l'analisi del contenuto delle recensioni può risultare particolarmente utile per comprendere quali prodotti, processi, momenti della customer journey o segmenti di clientela siano associati alle esperienze più positive e più negative. Gli indicatori assumono

quindi una funzione di selezione: non sostituiscono l'analisi qualitativa o semantica, ma contribuiscono a stabilire dove valga maggiormente la pena approfondire.

Anche l'asimmetria di risposta va interpretata come una scelta di allocazione dell'attenzione manageriale. Una forte concentrazione delle risposte sulle recensioni negative può riflettere una razionale strategia di service recovery, soprattutto quando le risorse disponibili sono limitate; la letteratura mostra infatti che priorità e modalità di risposta possono opportunamente variare in funzione della valenza della recensione e dei vincoli di budget (Wang et al., 2020; Lopes et al., 2023). Al tempo stesso, un'elevata asimmetria rende visibile una dimensione spesso trascurata: i clienti soddisfatti che dedicano tempo a esprimere pubblicamente il proprio apprezzamento possono non ricevere alcun riconoscimento dall'impresa. La scelta se riequilibrare tale comportamento dipende quindi dall'obiettivo assegnato alla risposta; se sia prevalentemente di recupero del disservizio oppure connesso ad una più ampia gestione della relazione. Gli effetti delle risposte a recensioni positive e negative non sono infatti necessariamente simmetrici (Wu, Ye e Chu, 2026).

Un'ulteriore implicazione riguarda la competitive intelligence. Gli indicatori utilizzati in questo lavoro sono costruiti, almeno nelle piattaforme che espongono i dati necessari, su informazioni pubblicamente osservabili. Possono quindi essere calcolati non soltanto per la propria impresa, ma anche per i concorrenti, consentendo di confrontare il modo in cui vengono valutati, come si presenta il flusso delle loro recensioni e come gestiscono pubblicamente le risposte. L'utilità non consiste tanto nell'attribuire un significato assoluto al singolo valore, quanto nel collocarlo all'interno di un insieme competitivo omogeneo e seguirne l'evoluzione nel tempo.

7.2. Livello operativo

Sul piano operativo, alcuni indicatori possono essere tradotti direttamente in obiettivi e criteri di intervento. La latenza di risposta, ad esempio, trasforma l'indicazione generica di "rispondere rapidamente" in una misura osservabile e monitorabile. Può inoltre essere articolata per valenza della recensione, distinguendo il tempo di reazione ai feedback negativi da quello relativo ai riscontri ordinari. La rilevanza della tempestività è coerente con la letteratura sulla webcare, che individua nella risposta tempestiva e personalizzata una delle condizioni che ne accrescono l'efficacia (Lopes et al., 2023). Anche gli effetti osservati delle

managerial responses sulla reputazione e sul successivo comportamento dei recensori confermano che la risposta non costituisce un atto neutro nello spazio pubblico della piattaforma (Proserpio e Zervas, 2017).

Il profilo di lunghezza offre invece un criterio per orientare la lettura del corpus. Come mostrato nel paragrafo 5.2, tuttavia, la relazione tra lunghezza e valutazione assume configurazioni differenti nei due dataset. Non è quindi opportuno assumere, ad esempio, che le recensioni negative siano sempre quelle più articolate o informative. Il profilo va costruito sul proprio contesto e utilizzato come criterio di priorità, non come regola universale.

La stessa logica vale più in generale per l'interpretazione degli indicatori. In assenza di benchmark consolidati e trasferibili, appare preferibile costruire riferimenti empirici "in casa", confrontando l'impresa con concorrenti osservati sulla stessa piattaforma, nello stesso mercato e, per quanto possibile, nello stesso periodo. Un'asimmetria pari a +64, ad esempio, dice poco in assoluto; diventa più informativa se colloca l'impresa tra quelle che concentrano maggiormente le proprie risposte sui clienti insoddisfatti. Il benchmark rilevante è quindi soprattutto relativo e contestuale, più che fondato su soglie assolute.

Infine, la qualità degli indicatori dipende dalla qualità dei dati sui quali vengono calcolati. La verifica dell'integrità del dataset non dovrebbe quindi essere considerata un'attività preliminare svolta una tantum, ma una componente ricorrente del processo di analisi. Duplicazioni, recensioni scarsamente informative, anomalie temporali o variazioni nella struttura dei dati possono modificarsi nel tempo e alterare il significato delle misure. Il monitoraggio della qualità del flusso informativo diventa pertanto parte integrante dello stesso sistema di controllo (Jindal e Liu, 2008; Ott et al., 2011).

8. Limiti e agenda di ricerca

Il primo limite del lavoro riguarda la natura dei dati utilizzati. Il campionamento del dataset italiano non è probabilistico: per ciascuna marca sono state acquisite le recensioni restituite dalla piattaforma secondo il proprio ordinamento predefinito, fino al limite stabilito. A questo si aggiunge un problema più generale delle recensioni online, che riflettono per loro natura un processo di autoselezione da parte di chi decide di esprimere pubblicamente la propria esperienza (Hu, Pavlou e Zhang, 2006). I risultati non possono quindi essere interpretati come rappresentativi dell'intera popolazione dei clienti.

Una conseguenza specifica della procedura di estrazione riguarda inoltre la finestra temporale osservata. Per le marche caratterizzate da un elevato flusso di recensioni, le unità estratte coprono periodi relativamente brevi; per quelle con un flusso minore, la stessa numerosità si distribuisce su intervalli molto più lunghi, compresi nel dataset tra cinque giorni e sedici mesi. Ciò limita in particolare il confronto diretto dell'indicatore di burstiness tra marche diverse e ne rende più appropriato l'impiego per l'analisi della stessa marca nel tempo.

Un secondo limite riguarda la numerosità delle unità disponibili per i confronti tra imprese. Le correlazioni presentate nel paragrafo 5.4 poggiano su un numero molto ridotto di osservazioni e sono state pertanto interpretate esclusivamente in chiave esplorativa. Non consentono di formulare conclusioni inferenziali né, tanto meno, di stabilire relazioni causali tra il comportamento di risposta dell'impresa e la struttura delle valutazioni dei clienti. Un'estensione del numero di marche e settori costituisce quindi un naturale sviluppo dell'analisi.

Anche il perimetro empirico è necessariamente limitato. I dati provengono da due piattaforme, due mercati nazionali e un numero ristretto di categorie merceologiche; il dataset statunitense, in particolare, riguarda una sola categoria, quella beauty. Le differenze osservate nel profilo di lunghezza delle recensioni, discusse nel paragrafo 5.2, non consentono quindi di separare pienamente l'effetto della piattaforma da quello della categoria o del mercato. Replicare l'analisi su ulteriori categorie e piattaforme permetterebbe di comprendere meglio quali indicatori mantengano caratteristiche relativamente stabili e quali richiedano invece benchmark specifici per il contesto.

Un ulteriore limite riguarda la disponibilità dei dati necessari al calcolo degli indicatori relativi alla risposta dell'impresa. Non tutte le piattaforme rendono disponibile il riscontro

aziendale e la relativa marca temporale; il sistema proposto è quindi pienamente applicabile soltanto negli ambienti nei quali queste informazioni sono pubblicamente osservabili. Anche questo vincolo suggerisce di considerare gli otto indicatori come un sistema modulare: alcuni possono essere utilizzati in qualunque ambiente di recensione, mentre altri dipendono dalle caratteristiche informative della piattaforma.

Una seconda direttrice dell'agenda di ricerca riguarda invece l'estensione del sistema dal comportamento osservabile al contenuto della relazione. Il presente lavoro misura dal lato del cliente come vengono formulate le valutazioni e dal lato dell'impresa se, a chi e con quale tempestività viene fornita una risposta, ma non classifica semanticamente la natura dell'esperienza raccontata. Un passo successivo potrebbe distinguere, ad esempio, recensioni riferite prevalentemente alla prestazione ricevuta da quelle nelle quali emerge una dimensione più propriamente relazionale (promessa disattesa, mancato riconoscimento, assenza di ascolto o percezione di trattamento inadeguato).

Una simile estensione richiederebbe inevitabilmente l'introduzione di un giudizio semantico e rappresenterebbe quindi anche un terreno naturale per l'impiego dei modelli linguistici. Coerentemente con la prospettiva adottata nel presente lavoro, tuttavia, il modello dovrebbe operare all'interno di un protocollo definito e verificabile: categorie esplicitate ex ante, codifica validata su un sottoinsieme di recensioni, confronto con il giudizio umano e verifica della stabilità delle classificazioni. La complementarità tra giudizio del ricercatore e capacità analitica degli LLM proposta dalla letteratura recente (Arora, Chakraborty e Nishimura, 2025) offre una possibile direzione in questo senso.

L'agenda di ricerca si sviluppa quindi lungo due direttrici complementari: estendere la base empirica, per verificare la trasferibilità e i limiti contestuali degli indicatori, e aumentarne progressivamente la profondità informativa, introducendo misure semantiche senza rinunciare alla trasparenza e alla riproducibilità del processo. È proprio in questo secondo passaggio che potrà essere ulteriormente verificato il principio su cui si fonda il lavoro: utilizzare l'intelligenza artificiale per ampliare la capacità di analisi, lasciando al ricercatore la definizione di ciò che vale la pena misurare e dei criteri con cui farlo.

Bibliografia

Arora N., Chakraborty I. e Nishimura Y. (2025), AI-Human Hybrids for Marketing Research: Leveraging Large Language Models (LLMs) as Collaborators, *Journal of Marketing*, 89(2): 43-70.

Bitner M. J., Booms B. H. e Tetreault M. S. (1990), The service encounter: diagnosing favorable and unfavorable incidents, *Journal of Marketing*, 54(1): 71-84.

Chevalier J. A. e Mayzlin D. (2006), The effect of word of mouth on sales: online book reviews, *Journal of Marketing Research*, 43(3): 345-354.

Dellarocas C. (2003), The digitization of word of mouth: promise and challenges of online feedback mechanisms, *Management Science*, 49(10): 1407-1424.

Filieri R. (2015), What makes online reviews helpful? A diagnosticity-adoption framework to explain informational and normative influences in e-WOM, *Journal of Business Research*, 68(6): 1261-1270.

Godes D. e Mayzlin D. (2004), Using online conversations to study word-of-mouth communication, *Marketing Science*, 23(4): 545-560.

Grewal D., Saturnino C.B., Davenport T. et al. (2025), How generative AI Is shaping the future of marketing , *Journal of the Academy of Marketing Science*, 53: 702-722.

Hannah D. R., Treen E. R., Pitt L. e Berthon P. R. (2016), But you promised! Managing consumers' psychological contracts, *Business Horizons*, 59(4): 363-368.

Hennig-Thurau T., Gwinner K. P., Walsh G. e Gremler D. D. (2004), Electronic word-of-mouth via consumer-opinion platforms: what motivates consumers to articulate themselves on the internet?, *Journal of Interactive Marketing*, 18(1): 38-52.

Homburg C. e Fürst A. (2005), How organizational complaint handling drives customer loyalty: an analysis of the mechanistic and the organic approach, *Journal of Marketing*, 69(3): 95-114.

Hu N., Pavlou P. A. e Zhang J. (2006), Can online reviews reveal a product's true quality? Empirical findings and analytical modeling of online word-of-mouth communication, *Proceedings of the 7th ACM Conference on Electronic Commerce*: 324-330.

Jindal N. e Liu B. (2008), Opinion spam and analysis, *Proceedings of the International Conference on Web Search and Web Data Mining*: 219-230.

Lopes A. I., Dens N., De Pelsmacker P. e Malthouse E. C. (2023), Managerial response strategies to eWOM: A framework and research agenda for webcare, *Tourism Management*, 98: 104739.

Ma L., Sun B. e Kekre S. (2015), The squeaky wheel gets the grease: an empirical analysis of customer voice and firm intervention on Twitter, *Marketing Science*, 34(5): 627-645.

Moe W. W. e Trusov M. (2011), The value of social dynamics in online product ratings forums, *Journal of Marketing Research*, 48(3): 444-456.

Morgan R. M. e Hunt S. D. (1994), The commitment-trust theory of relationship marketing, *Journal of Marketing*, 58(3): 20-38.

Mudambi S. M. e Schuff D. (2010), What makes a helpful online review? A study of customer reviews on Amazon.com, *MIS Quarterly*, 34(1): 185-200.

Ott M., Choi Y., Cardie C. e Hancock J. T. (2011), Finding deceptive opinion spam by any stretch of the imagination, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*: 309-319.

Proserpio D. e Zervas G. (2017), Online Reputation Management: Estimating the Impact of Management Responses on Consumer Reviews, *Marketing Science*, 36(5): 645-665.

Rousseau D. M. (1989), Psychological and implied contracts in organizations, *Employee Responsibilities and Rights Journal*, 2(2): 121-139.

Sarstedt M., Adler S.J., Rau L. e Schmitt B. (2024), Using large language models to generate silicon samples in consumer and marketing research: Challenges, opportunities, and guidelines, *Psychology & Marketing*, 41: 1254-1270.

Shannon C. E. (1948), A mathematical theory of communication, *Bell System Technical Journal*, 27: 379-423, 623-656.

Smith A. K., Bolton R. N. e Wagner J. (1999), A model of customer satisfaction with service encounters involving failure and recovery, *Journal of Marketing Research*, 36(3): 356-372.

Sun M. (2012), How does the variance of product ratings matter?, *Management Science*, 58(4): 696-707.

Van Noort G. e Willemsen L. M. (2012), Online damage control: the effects of proactive versus reactive webcare interventions in consumer-generated and brand-generated platforms, *Journal of Interactive Marketing*, 26(3): 131-140.

Wang L., Ren X., Wan H. e Yan J. (2020), Managerial responses to online reviews under budget constraints: Whom to target and how, *Information & Management*, 57(8): 103382.

Wu J., Ye J. e Chu J. (2026), Soothing the unsatisfied or pleasing the satisfied? The effects of managerial responses to positive versus negative reviews on customer ratings and financial performance, *Journal of the Academy of Marketing Science*, 54(1): 70-90.