Università degli Studi di Padova

Dipartimento di Studi Linguistici e Letterari

Corso di Laurea Magistrale in
Strategie di Comunicazione
Classe LM-92

Tesi di Laurea

Reddit e il dibattito sull'intelligenza artificiale: un'analisi statistica dei commenti online per l'esplorazione dell'opinione pubblica

Relatore
Prof. Arjuna Tuzzi

Laureando
Aurora Cilione
n° matr.2110469 / LMSGC

Anno Accademico 2025 / 2026

Indice

Introduzione

Nell'attuale *information* e *network society*, lo sviluppo delle tecnologie digitali ha rivoluzionato l'ambiente entro il quale si svolgono le interazioni sociali. Ridefinendo le modalità di produzione, organizzazione e trasmissione delle informazioni, i media digitali hanno inciso profondamente sulla società nel suo complesso e condotto alla nascita di nuove forme di socialità. Oltre a imporsi come canale primario di accesso all'informazione, le piattaforme online costituiscono oggi la nuova forma assunta dalla *Great Community* di Dewey (1927) e dalla sfera pubblica di Habermas (1962): i social media si contraddistinguono in quanto ambiente digitale di partecipazione attiva alla creazione, condivisione e discussione dei contenuti (*User-Generated Content*, UGC), nel quale l'uomo (o meglio, l'utente) può confrontarsi apertamente con prospettive differenti e, attraverso il dibattito libero e razionale, affinare la propria capacità critica.

Il capitolo introduttivo di questa tesi si occuperà di porre le basi teoriche per comprendere al meglio l'affermarsi dei social media come spazio privilegiato per l'espressione dell'opinione pubblica. In uno scenario caratterizzato dalla diffusa possibilità di ricercare informazioni online e di partecipare attivamente alla creazione e diffusione delle stesse, l'atto di scrivere e leggere commenti online è parte integrante del consumo quotidiano delle piattaforme social, in molte delle quali le discussioni online nella forma di commenti scritti costituiscono un elemento cardine del loro funzionamento. Se i primi studi nell'ambito delle scienze sociali computazionali si sono focalizzati sull'indagare la comunicazione *one-to-one*, la nascita e lo sviluppo di canali di comunicazione fondati sulla cultura partecipativa hanno reso possibile lo studio delle discussioni collettive che hanno origine nell'ambiente digitale. Si comprende così come, da un punto di vista analitico, i commenti pubblicati online e, in particolare, *le threaded online conversations* tipiche dei forum di discussione, acquisiscano rilevanza in quanto vasta fonte di dati testuali per l'analisi di tematiche socialmente controverse. Una raccolta di commenti online si configura in tal senso come un corpus testuale di grandi dimensioni, la cui analisi richiede l'applicazione di metodi statistici in grado di rilevarne le strutture latenti.

Il presente lavoro si propone pertanto di evidenziare il potenziale dell'analisi statistica dei dati testuali (ASDT), un'analisi quantitativa basata sulla codifica *ex post* e automatica di unità testuali, come strumento metodologico per l'esplorazione

dell'opinione pubblica. Al fine di conseguire questo primo obiettivo, si è assunto come caso studio il dibattito online relativo all'intelligenza artificiale. Tale scelta è da ricondurre a due motivazioni principali. In primo luogo, l'intelligenza artificiale si contraddistingue in quanto argomento di interesse sociale in grado di soddisfare i criteri di attualità e opinabilità del tema e di disponibilità dei dati testuali: sin dal rilascio di ChatGPT, avvenuto a novembre 2022, l'AI generativa (GenAI) si è posta al centro di un dibattito pubblico in continua espansione e fortemente polarizzato. La coesistenza di opinioni divergenti è tuttavia frutto di una percezione personale derivata dalla fruizione quotidiana di contenuti online. È dunque questa osservazione preliminare che ha stimolato l'interesse a esplorare le opinioni espresse dagli utenti e che ha, secondariamente alla rilevanza del tema, guidato il lavoro di ricerca. L'applicazione di metodi statistici consente difatti di superare i limiti dell'analisi qualitativa e analizzare corpora di grandi dimensioni senza richiedere la lettura integrale del materiale oggetto di studio. Nel caso in esame, il corpus si compone di un totale di 5.245 commenti estratti da Reddit, il più grande forum di discussione online: il Social Network, costruito attorno ai blocchi funzionali di *conversations*, *sharing*, *raputation* e *groups*, si presta come esempio di spazio digitale in cui le interazioni tra utenti (i redditors) alimentano il dibattito collettivo, interpretabile come espressione dell'opinione pubblica. L'utilità dell'analisi statistica dei dati testuali, l'utilizzo dei commenti dei redditors come corpora testuali e l'eterogeneità del dibattito pubblico sull'intelligenza artificiale convergono nella seguente domanda di ricerca: in che modo si articola il discorso dei redditors intorno all'intelligenza artificiale? Quali sono i maggiori nuclei tematici del dibattito online? Quali relazioni intercorrono tra i temi e le parole più ricorrenti e/o discriminanti?

Il contesto empirico di partenza e la metodologia adottata per rispondere alla domanda di ricerca saranno ampiamente discussi nel secondo capitolo dell'elaborato. Il presente studio ricorre difatti a due metodi appartenenti al ramo dei metodi *unsupervised*: il *text clustering* e l'analisi delle corrispondenze, che consentono di identificare le diverse declinazioni assunte dal discorso e il modo in cui queste si interconnettono o meno tra loro. Dopo averne delineato il quadro teorico, la seconda parte del lavoro sarà interamente dedicata alla componente empirica dell'analisi e si svilupperà conformemente alle tre fasi canoniche di un percorso di analisi. In particolare, nel terzo capitolo saranno illustrati i criteri adottati per l'acquisizione dei dati testuali e presentati i primi parametri relativi al

corpus, assieme alla loro rielaborazione in proprietà misurabili. Saranno infine il quarto e il quinto capitolo della ricerca, a forte connotazione qualitativa, a dedicarsi alla rappresentazione, alla contestualizzazione e all'interpretazione dei risultati ottenuti, passaggi fondamentali affinché questi siano utilizzabili e utilizzati.

1. Il ruolo dei media digitali nella costruzione dell'opinione pubblica

1.1 L'opinione pubblica

Nonostante una lunga storia di studi, che affonda le proprie radici nei primi decenni del Novecento, fornire una definizione di "opinione pubblica" risulta ancora oggi problematico. In *Lezioni brevi sull'opinione pubblica*, un volume transdisciplinare che riunisce alcuni dei contributi italiani più recenti sul tema, Laura Gherardi (2022) ne sottolinea la natura enigmatica e la paragona ad "[…] un fantasma che occupa quotidianamente la mente di politici e politologi, giornalisti, comunicatori, sociologi, psicologi sociali, ma anche persone comuni" (*ivi*, p. 19). Si tratta difatti di un termine-ombrello che, proprio a causa della sua pervasività e pluralità di ambiti di impiego, resiste ad ogni tentativo di cristallizzazione lessicale e non si lascia ingabbiare in una limitazione definitoria.

Sin dal 1940, il politologo americano Harwood L. Childs richiamava l'attenzione sulla molteplicità di accezioni del termine e sulle difficoltà di giungere a una definizione univoca che potesse ritenersi accettabile, evidenziando come la questione fosse ampiamente discussa a livello internazionale. Molti autori hanno infatti tentato, attraverso una rassegna della letteratura del settore, di ricondurre a ordine questo caos concettuale, producendo tuttavia nuove definizioni che hanno solo alimentato la complessità del dibattito. Pur riconoscendo questi limiti, anche Childs (1940) apporta il proprio contributo e definisce l'opinione pubblica come segue:

> Public opinion is any collection of individual opinions, regardless of the degree of agreement or uniformity. The degree of uniformity is a matter to be investigated, not something to be arbitrarily set up as a condition for the existence of public opinion. (*ivi*, p. 48)[1]

È importante enfatizzare come, in questo caso, il significato attribuito al termine sia frutto di un'analisi critica delle formulazioni fino ad allora presenti in letteratura. Per motivare la scelta di riportare questa definizione è opportuno ripercorrere brevemente lo smantellamento attuato dall'autore. Nodo centrale sono gli stessi componenti del sintagma "opinione pubblica": Childs (1940) si sofferma sulla necessità di mettere in

Nota sulle traduzioni: Salvo diversa indicazione, tutte le traduzioni dall'inglese sono dell'autrice.

[1] L'opinione pubblica è un qualsiasi insieme di opinioni individuali, indipendentemente dal grado di accordo o uniformità. Il grado di uniformità costituisce un aspetto da indagare, non una condizione da assumere arbitrariamente come requisito per l'esistenza dell'opinione pubblica.

relazione l'opinione pubblica a *opinioni* precise riguardo a un determinato oggetto e ad un *pubblico* specifico che le esprime.

In merito alle opinioni, occorre distinguere tre aspetti fondamentali. Innanzitutto, sebbene non siano sempre indici affidabili dell'atteggiamento di un individuo, le opinioni sono espressioni verbali oggettive e significative di per sé, perché gettano luce sul suo comportamento atteso. In secondo luogo, è necessario precisare che "[o]pinions differ from one another in many respects, such as content, the form in which they are expressed, their quality, their stability, their intensity, and the way in which they have been formed or elicited" (*ivi*, p. 43)[2]. Infine, si può ritenere implicito che, nel parlare di un'opinione, si faccia sempre riferimento all'opinione di un'unica persona, e non di un gruppo: è solo raccogliendo le opinioni dei singoli individui che è possibile comprendere lo stato dell'opinione pubblica. Contrariamente, quando si parla di pubblico l'immagine che generalmente si ha è quella di un gruppo di individui, spesso con interessi simili e talvolta accomunati dall'essere parte di un'organizzazione. Tuttavia, come osservato da Childs (1940), "[…] the public in which we are interested may be composed of a very heterogeneous collection of individuals without organization, lacking identifying symbols and attributes" (*ivi*, p. 41)[3]. Definire il concetto di pubblico, dunque, non equivale a ridurlo a un gruppo di individui caratterizzati da un attributo comune e, in virtù di questo, reputarli un'entità distinta. Tale limitazione risulta ancor più evidente se si considera l'esistenza di diverse tipologie di pubblico e la possibilità che ogni individuo appartenga simultaneamente a più gruppi. Si potrebbe pertanto concludere che, se per pubblico si intende generalmente un qualsiasi aggregato di individui che si riunisce temporaneamente intorno ad un tema di interesse, l'opinione pubblica corrisponde a qualsiasi insieme di opinioni individuali.

Avendo chiarito questo punto, è possibile ora soffermarsi sull'elemento principale su cui Childs (1940) incentra la propria definizione. Alcuni autori concordano sul fatto che si possa parlare di opinione pubblica solo se l'insieme di opinioni in questione è condivisa dalla totalità – un livello di accordo che è, per di più, difficilmente raggiunto – o dalla maggioranza del pubblico di riferimento: il grado di accordo o uniformità tra le opinioni

[2] "[l]e opinioni differiscono tra loro sotto molti aspetti, come il contenuto, la forma in cui vengono espresse, la loro qualità, la loro stabilità, la loro intensità e il modo in cui sono state formate o suscitate".
[3] "[…] il pubblico di nostro interesse può essere costituito da un insieme molto eterogeneo di individui, privo di organizzazione e di simboli o attributi identificativi".

individuali è considerato condizione di esistenza dell'opinione pubblica. Childs mette in discussione questa visione, osservando che un ricercatore che si propone di individuare un livello definito di accordo finirebbe per concentrarsi su un solo aspetto dell'opinione pubblica e, di conseguenza, condurrebbe un'indagine infruttuosa. Alla luce di queste argomentazioni, si può concludere che non sia necessaria la presenza di un accordo uniformante per poter parlare di opinione pubblica: la coesistenza di opinioni diverse, o persino contrastanti, non compromette la possibilità di definire il loro insieme come "opinione pubblica". La diversità ne è anzi un elemento intrinseco. A conferma delle osservazioni di Childs (1940), a circa otto decenni di distanza, Laura Gherardi (2022) riassume:

> […] l'opinione pubblica (indipendentemente dal fatto che la si consideri come un mero aggregato di preferenze individuali o come un processo emergente da discussione collettiva) è sempre l'*insieme* delle correnti di opinione, che sono tra loro *diverse*, in un dato momento storico. […] Corretto sarebbe, dunque, parlare sempre di *opinioni pubbliche* al plurale: l'opinione pubblica non è la maggioranza delle opinioni rispetto a un tema di natura pubblica […]. Occorre poi, certo, che queste opinioni siano manifeste, che siano cioè verbalizzate o scritte, giacché un conto è pensare una cosa, un altro è dichiararla.
> (*ivi*, pp. 19-20)

Tra i vari aspetti vagliati da Gherardi (2022) sull'enigmaticità dell'opinione pubblica, assume rilievo la domanda circa la sua natura passiva o attiva, un tema ricorrente nella storia di questo concetto e tuttora irrisolto. Le seguenti sottosezioni si occuperanno di indagare le due visioni contrastanti che ne hanno segnato l'origine, considerate fondamentali per comprendere l'evoluzione dell'idea di opinione pubblica: quella di Walter Lippmann, incentrata sulla manipolazione mediatica, e quella di John Dewey, fondata sul valore del confronto e del dibattito collettivo.

1.1.1 Lippmann e Dewey: tra manipolazione mediatica e dibattito collettivo

Non è ancora possibile stabilire con certezza se l'opinione pubblica sia intrinsecamente manipolabile e manipolata o, al contrario, critica e razionale. Il dibattito moderno su questo tema prende forma nel primo dopoguerra, a partire dalle posizioni opposte di Walter Lippmann e John Dewey, due autori che si confrontano sulla crescente manipolazione e strumentalizzazione delle informazioni. Genesi di questa preoccupazione fu la massiccia propaganda messa in atto, durante la Prima Guerra

Mondiale, dal governo americano con lo scopo di orientare l'opinione pubblica verso l'interventismo bellico. Essa rivelò in maniera inequivocabile come l'opinione pubblica, in assenza di un accesso diretto agli eventi, potesse essere plasmata dall'alto attraverso strumenti di comunicazione di massa.

In questo contesto, entrambi gli autori si interrogano sulla capacità del cittadino medio di comprendere e partecipare consapevolmente ai processi democratici. Il giornalista e politologo Walter Lippmann, precursore degli studi sulla manipolazione mediatica, espresse una visione fortemente critica sulla capacità cognitiva del pubblico. L'assunto fondamentale dell'autore è che l'opinione sia essenzialmente passiva, e dunque manipolabile. In *L'opinione pubblica* (1922), egli afferma:

> Non c'è il tempo, né la possibilità per una conoscenza profonda. E così ci limitiamo a notare un tratto, che caratterizza un tipo ben conosciuto, e riempiamo il resto dell'immagine grazie agli stereotipi che ci portiamo in testa. […] Le più sottili e contagianti influenze sono quelle che creano e conservano il repertorio degli stereotipi. Sentiamo parlare del mondo prima di vederlo. Immaginiamo la maggior parte delle cose prima di averne esperienza. E questi preconcetti, se non siamo stati molto avvertiti dall'educazione, incidono profondamente nell'intero processo di percezione.
>
> (*ivi*, 1922; tr. it. 2004, p. 68)

Secondo Lippmann, è impossibile fare esperienza diretta dell'intero ambiente e, conseguentemente, di tutto ciò che le nostre opinioni coprono. Per questo, per orientarsi nella confusione del mondo, l'uomo è costretto a ricorrere a modelli semplificati della realtà, denominati pseudo-ambienti. Tali rappresentazioni sono costruite da ciascun individuo a partire da "immagini che egli si forma o che gli vengono date" (*ivi*, p. 24), stereotipi e preconcetti che incidono profondamente sul suo comportamento e sulla percezione del mondo circostante. Il ricorso agli stereotipi è però ingannevole: se, da un lato, questi modelli semplificati ci consentono di ridurre la complessità del mondo e, dunque, ci aiutano a interpretarlo, dall'altro riducono e distorcono la realtà. La distorsione è innanzitutto da ricondurre alla credulità con cui vengono adoperati gli stereotipi, i quali, ricorda Lippmann, "sono carichi di preferenze, soffusi di simpatia o antipatia, abbarbicati a timori, brame, passioni, orgoglio, speranze" (*ivi*, p. 88) che attenuano la ricettività dell'individuo a informazioni nuove e, talvolta, contrarie. In aggiunta, gli stereotipi, base di costruzione degli pseudo-ambienti, sono trasmessi da fonti esterne, che selezionano e filtrano le informazioni accessibili. Ne consegue che le nostre opinioni individuali – e,

nel loro insieme, l'opinione pubblica – si formino in gran parte su ciò che ci viene riferito dagli altri e, in una prospettiva più ampia, dai media. In un ambiente "troppo grande, troppo complesso e troppo fuggevole per consentire una conoscenza diretta" (*ivi*, p. 18), la censura, la segretezza e l'informazione distorta da fonti esterne trovano terreno fertile per la creazione di influenze mediatiche. È così che l'opinione pubblica risulta particolarmente suscettibile alla manipolazione di chi controlla la produzione e la diffusione delle informazioni. La democrazia, quindi, non può più fondarsi sulla presunta competenza del pubblico: per Lippmann, la soluzione risiede nell'affidare il potere decisionale ad un governo di esperti capaci di giudicare gli interessi comuni.

A collocarsi al polo opposto delle osservazioni di Lippmann è il filosofo e pedagogista John Dewey, che sostiene l'idea dell'opinione pubblica come forza attiva. Pur concordando con Lippmann in merito al ruolo degli stereotipi e all'inevitabile mancanza di conoscenza dell'uomo, Dewey presenta una visione più ottimistica della capacità cognitiva del pubblico. Nel suo saggio *The Public and Its Problems* (1927), egli risponde direttamente alle tesi di Lippmann e osserva:

> The essential need […] is the improvement of the methods and conditions of debate, discussion and persuasion. That is *the* problem of the public. […] It is not necessary that the many should have the knowledge and skill to carry on the needed investigations; what is required is that they have the ability to judge of the bearing of the knowledge supplied by others upon common concerns.
> (*ivi*, p. 208)[4]

Dewey accetta la critica che Lippmann muove all'opinione pubblica, ma ne rifiuta le conclusioni. Mentre il secondo propone una soluzione di tipo elitista, il primo ne avanza una pedagogica: se gli uomini sono inevitabilmente portati ad agire sulla base di rappresentazioni parziali e incomplete del mondo, piuttosto che sulla realtà oggettiva, la chiave risiede nell'educazione, non intesa come finalizzata alla formazione di un individuo onnicompetente (*omnicompetent citizen*), ma come finalizzata allo sviluppo di una capacità che permetta all'uomo di valutare criticamente la conoscenza fornita da altri. L'autore difatti evidenzia come l'idea, coniata da Lippmann, di individuo onnicompetente sia un'illusione, e che "[f]aculties of effectual observation, reflection and desire are habits

[4] La necessità essenziale […] consiste nel miglioramento dei metodi e delle condizioni del dibattito, della discussione e della persuasione. Questo è il problema del pubblico. […] Non è necessario che la maggioranza possieda la conoscenza e le competenze per condurre le indagini necessarie; ciò che è richiesto è che abbia la capacità di valutare l'influenza della conoscenza fornita da altri sui temi di interesse comune.

acquired under the influence of the culture and institutions of society, not ready-made inherent powers" (*ivi*, 158)[5]. In termini riassuntivi, la prospettiva deweyana non condanna l'assenza nell'uomo di una capacità critica: essa non è innata, ma può solo essere acquisita e affinata attraverso l'educazione e le pratiche comunicative e sociali.

Il problema dell'opinione pubblica non è pertanto l'incompetenza del pubblico né la sua difficoltà a comprendere la complessità del mondo, bensì l'assenza di condizioni che favoriscano il dibattito per lo sviluppo delle suddette facoltà critiche: l'opinione pubblica può essere critica solo se immersa in un contesto che promuova lo scambio, la partecipazione e le esperienze sociali. Per Dewey, il risveglio del pubblico è affidato alla transizione dalla *Great Society* alla *Great Community*, una società nella quale non esistano ostacoli alla comunicazione e che permetta, oltre alla libera diffusione delle informazioni, un confronto aperto tra saperi e vissuti differenti: "[t]ill the Great Society is converted into a Great Community, the Public will remain in eclipse. Communication can alone create a great community" (*ivi*, p. 142)[6]. La prospettiva deweyna pone dunque l'accento sul ruolo della comunicazione, che costituisce la condizione stessa per l'emergere di una comunità consapevole, capace di deliberazione critica e meno vulnerabile alla manipolazione.

1.1.2 Jürgen Habermas: dalla sfera pubblica alla sfera digitale

Enfatizzando la necessità di erigere spazi sociali in cui ciascuno possa esprimere la propria opinione, alimentare il dibattito e sviluppare il proprio potenziale critico e partecipativo, è lecito affermare che già negli anni Venti Dewey accennava a quella che, qualche decennio più tardi, sarebbe stata definita "sfera pubblica", presupposto dell'esistenza di un'opinione pubblica criticamente razionale. Da molti considerato il più influente teorico della sfera pubblica, il filosofo e sociologo tedesco Jürgen Habermas la definisce nel suo *Storia e critica dell'opinione pubblica* (1962) come uno spazio sociale in cui l'opinione pubblica può formarsi attraverso discussione critica e razionale. Habermas, difatti, riconduce la nascita dell'opinione pubblica in Europa alla nascita della

[5] "[l]e facoltà di osservazione efficace, riflessione e desiderio sono abitudini acquisite sotto l'influenza della cultura e delle istituzioni della società, non poteri innati già pronti".
[6] "[f]inché la Grande Società non sarà trasformata in una Grande Comunità, il Pubblica rimarrà in eclissi. Solo la comunicazione può creare una grande comunità".

sfera pubblica borghese tra il XVII e il XVIII secolo, con l'emergere della borghesia – razionale, consapevole ed indipendente rispetto al potere – e la diffusione della stampa. Si tratta di una sfera – che l'autore distingue dalla sfera pubblica rappresentativa, più istituzionalizzata – in cui gli individui partecipano attivamente, liberamente e razionalmente a dialoghi riguardanti questioni di interesse comune, influenzando indirettamente le decisioni politiche. La sfera pubblica è però soggetta a trasformazioni, attribuibili alla costante innovazione nel campo dei media: da un lato, la nascita della stampa e l'avvento della radio e della televisione hanno progressivamente ampliato le modalità di accesso all'informazione e di partecipazione al pubblico; dall'altro, tuttavia, l'espansione del capitalismo ha conferito sempre più potere ai suddetti media tradizionali, che si sono evoluti in mezzi di controllo in grado di erodere l'autonomia critica della sfera pubblica. Stampa, radio e televisione si sono gradualmente discostati dalla loro funzione informativa per assumere un ruolo di influenza su quelle stesse idee e punti di vista che in passato erano unicamente deputati a diffondere. L'autore esplica questa trasformazione come segue:

> In confronto alla stampa dell'epoca liberale, i mass-media da un lato hanno raggiunto una forza di penetrazione e un'efficacia incomparabilmente maggiori (con essi si è estesa la sfera stessa della dimensione pubblica), dall'altra, si sono allontanati sempre più da questa sfera per penetrare in quella, un tempo privata, dello scambio di merci. […] Mentre prima la stampa poteva soltanto mediare e rafforzare il dibattito dei privati raccolti nel pubblico, adesso, viceversa, esso è plasmato dai mass-media.
> (*ivi*; tr. it. 1974, p. 225)

Habermas denuncia come, nella società contemporanea, persino la sfera pubblica diventi uno spazio dominato dalle logiche di mercato e dalle grandi corporazioni, che mercificano il dibattito plasmando il discorso pubblico secondo i propri interessi. In tal senso, è inevitabile osservare una corrispondenza con il pensiero di Lippmann, che già al tempo attribuiva un'importanza straordinaria ai media nella manipolazione dell'opinione.

Quello della sfera pubblica habermasiana resta comunque un modello ideale di spazio sociale che, nel contesto odierno, assume la forma più ampia di ambiente digitale, all'interno del quale l'avvento delle tecnologie digitali modifica le condizioni stesse di esistenza della sfera pubblica. A partire dagli anni Settanta, difatti, la digitalizzazione e la diffusione di Internet determinano la transizione verso quella che è stata definita "società

dell'informazione", uno spazio che ridefinisce tanto il modo in cui l'opinione pubblica si forma, quanto il modo in cui essa si esprime.

1.2 La società dell'informazione e la *network society*

La rapida evoluzione tecnologica degli ultimi decenni ha influito in modo significativo su diversi campi della società, dell'economia e della vita umana nel suo complesso, al punto tale da condurre molti studiosi a parlare di una vera e propria rivoluzione digitale. Le tecnologie digitali stanno difatti "[…] provocando sulle nostre vite conseguenze paragonabili a quelle che l'automazione – con la macchina a vapore, l'elettrificazione, la raffinazione del petrolio – ha avuto in altri momenti della nostra storia che definiamo 'rivoluzioni'" (Quarta & Smorto 2020, p. 1). Comunemente concettualizzata come Terza Rivoluzione Industriale, nella prospettiva economica la rivoluzione digitale è sancita dalla nascita società dell'informazione.

Prima di approfondire il significato di "società dell'informazione", è tuttavia opportuno premettere come la letteratura non sia unanime nel definire un numero esatto di rivoluzioni industriali: la loro periodizzazione è spesso al centro di un dibattito che vede, da un lato, storici che tendono a distinguere tre grandi fasi fondamentali e, dall'altro, studi più recenti che propongono l'introduzione di una Quarta Rivoluzione Industriale per designare il salto qualitativo della fase attuale (Schwab 2016). Ciononostante, pur riconoscendo quest'ultima come un fenomeno autonomo, molti osservatori concordano nell'interpretarla come diretta emanazione della Terza, il cui avvio viene convenzionalmente ricondotto agli anni Settanta con l'affermazione dell'informatica e dell'automazione. La Terza Rivoluzione Industriale definisce difatti l'avvento dei computer, delle telecomunicazioni e, successivamente, di Internet – innovazioni che ne giustificano l'appellativo di "rivoluzione digitale" e che, al contempo, la configurano come motore di sviluppo della Quarta.

La letteratura economica riconduce le innovazioni ad ampio raggio – come la macchina a vapore della Prima Rivoluzione Industriale e l'elettrificazione della Seconda – alle *General Purpose Technologies* (GPTs), un concetto introdotto dagli economisti Manuel Trajtenberg e Timothy F. Bresnahan (1995) e che identifica le tecnologie di portata generale, ossia quelle tecnologie i cui effetti si manifestano in molteplici settori dell'economia e della società. Nello specifico, nel loro articolo *General purpose*

technologies: 'Engines of growth?', che costituisce la prima sistematizzazione teorica delle GPTs, i due autori le definiscono come innovazioni tecnologiche caratterizzate da tre elementi fondamentali: "[…] pervasiveness, inherent potential for technological improvements, and 'innovational complementarities' […]" (*ivi*, p. 83). Per essere annoverata tra le GPTs, oltre a possedere la capacità di diffondersi trasversalmente in più ambiti, un'innovazione dovrebbe dunque presentare un potenziale intrinseco di miglioramento, ovvero la possibilità di essere oggetto di continui perfezionamenti tecnologici che ne amplino le capacità. Il terzo requisito riguarda invece la presenza di complementarietà con altri ambiti: una *General Purpose Technology* non solo stimola processi innovativi nel proprio settore di origine, ma favorisce lo sviluppo di nuove applicazioni e opportunità produttive negli ambiti in cui essa viene adottata (da qui il titolo dell'articolo, *'Engines of Growth?'*).

Alla luce di queste considerazioni, si comprende come le tecnologie digitali possano essere ascritte a pieno titolo alle GPTs. La loro *pervasiveness* è evidente: il digitale ha rivoluzionato tanto i mercati quanto le relazioni sociali, incidendo profondamente sulla qualità di vita delle persone e persino sui processi politici e di democratizzazione. Allo stesso modo, l'*inherent potential for technological improvements* si manifesta in una capacità evolutiva che raggiunge ritmi sempre più frenetici, al punto tale – come già accennato – da riconoscere la fase odierna come una nuova, Quarta Rivoluzione Industriale. Il digitale è difatti un fenomeno dinamico che, a partire dalla diffusione capillare di Internet, si è progressivamente esteso a una pluralità di dispositivi (Internet of Things), per continuare a evolversi in forme sempre più sofisticate in cui l'intelligenza artificiale e il machine learning ricoprono un ruolo costantemente in crescita. Inoltre, grazie alle sue *innovational complementarities*, il digitale contribuisce in gran misura all'aumento della produttività in tutti i comparti che investe, oltre che alla nascita di nuovi modelli di produzione e di nuovi settori, come l'e-commerce, le piattaforme digitali e l'economia dei dati.

Con l'affermarsi delle tecnologie digitali, la Terza Rivoluzione Industriale segna una trasformazione radicale non solo nei processi produttivi, ma anche nelle modalità di produzione, organizzazione e trasmissione delle informazioni. Il termine di "società dell'informazione", diffusosi largamente nel corso degli anni Novanta con la crescente accessibilità di Internet, è utilizzato nella sua accezione generale per indicare un nuovo

modello organizzativo improntato sulla centralità dell'informazione. Nonostante non vi sia un accordo unanime circa la definizione di "società dell'informazione", dalla letteratura dell'ultimo mezzo secolo emergono cinque principali approcci interpretativi: tecnologico, economico, occupazionale, spaziale e culturale (Webster, 2014). Ciascuno di questi, tuttavia, oltre che a cogliere unicamente un aspetto parziale della società dell'informazione, si limita a giustificarne l'esistenza con un semplice cambiamento quantitativo nella mole di informazione disponibile: in sintesi, oggi la quantità di informazioni è notevolmente aumentata, ed è per questo che si ritiene si possa parlare di società dell'informazione. In contrapposizione, il sociologo Webster sostiene una definizione più critica, di carattere marcatamente qualitativo piuttosto che quantitativo: "[…] [the information society's] main claim is not that there is more information today (there obviously is), but rather that the character of information is such as to have transformed how we live. The suggestion here is that *theoretical knowledge/information* is at the core of how we conduct ourselves these days" (*ivi*, p. 11)[7].

A dominare il dibattito sull'ascesa della società dell'informazione è il sociologo catalano Manuel Castells (1996), che negli anni Novanta sancisce l'inizio dell'era del capitalismo informazionale, mettendo in luce il cambiamento di paradigma all'interno di una società in trasformazione. Alla base del potere economico e politico non troviamo più la produzione di beni materiali tipica delle società industriali, bensì asset informativi o comunque intangibili, quali marchi, innovazione e conoscenza (i diritti di proprietà intellettuale, ad esempio, assumono un'importanza senza precedenti). Ne emerge, secondo l'autore, un sistema economico che è *informational*, *global* e *networked*: il suo successo è determinato dalla "[…] capacità di generare, elaborare e applicare con efficienza informazione basata sulla conoscenza", da attività "[…] di produzione, consumo e circolazione […] organizzate su scala globale", e da una concorrenza che ha luogo "[…] in una ragnatela globale di interazione tra reti aziendali" (*ivi*; tr. it. 2002, p. 83). In questo contesto, Castells (1996) descrive i networks come strutture organizzative fondamentali, ed introduce il concetto di *network society*.

[7] "[…] la principale tesi [della società dell'informazione] non è che oggi esista una quantità maggiore di informazioni (il che è ovvio), ma piuttosto che la natura stessa dell'informazione sia tale da aver trasformato il nostro modo di vivere. L'idea di fondo è che la *conoscenza teorica/l'informazione* costituisca il nucleo di come oggi conduciamo la nostra vita".

La centralità dei networks – intesi come una "forma organizzativa in reti" (*ivi*; tr. it. 2002, p. 83) – nell'ambito della produzione economica si riflette nella dimensione sociale. Lo sviluppo della comunicazione digitale ha infatti trasformato la spazialità dell'interazione sociale: essa assume una nuova forma, che l'autore concettualizza come *space of flows*. Nella loro *Introduction to Digital Media* (2019), Delfanti e Arvidsson riassumono il pensiero castellsiano:

> He described a "space of flows" composed of the spaces, both physical and mediated, where information, money, and people circulate. This space is configured as an open network, in which national borders and boundaries among organizations, communities, and groups are becoming less important. […] [I]ndividuals who have access to the space of flows are the ones who possess the skills necessary to efficiently and productively exchange information or move freely in different places, or between one organization and another. (*ivi*, p. 26)[8]

Mentre Habermas descriveva una sfera pubblica legata a luoghi fisici e media tradizionali, la *network society* trasferisce la discussione in uno spazio digitale. Nella *network society*, i media digitali non si riducono a semplici strumenti tecnologici: essi costituiscono un nuovo ambiente sociale, l'infrastruttura materiale che rende tangibile il suddetto *space of flows* e su cui la sfera pubblica si esprime.

1.2.1 I media digitali: il nuovo ambiente sociale della *network society*

Distinguiamo una moltitudine di fattori che concorrono alla costruzione dell'opinione pubblica. Essa si configura difatti come un processo dinamico e articolato che intreccia istruzione e ideologie politiche, norme e valori culturali, caratteristiche demografiche del pubblico e bias psicologici a cui esso è esposto (Anju 2024). Tra questi, agenti d'influenza sono innegabilmente i media digitali, che detengono un potere non trascurabile: in linea con le osservazioni di Lippmann (1922), la selezione delle tematiche trattate e le modalità con cui queste sono formulate contribuiscono ad influenzare le idee e il pensiero del pubblico che ne fruisce. A tal proposito, per una migliore comprensione del potere

[8] Descriveva uno "spazio di flussi" composto dagli spazi, sia fisici che mediati, dove circolano informazioni, denaro e persone. Questo spazio è configurato come una reta aperta, in cui le frontiere nazionali e i confini tra organizzazioni, comunità e gruppi stanno diventando meno rilevanti. […] Gli individui che hanno accesso allo spazio di flussi sono quelli che possiedono le competenze necessarie per scambiare informazioni in modo efficiente e produttivo o per spostarsi liberamente in luoghi diversi, o tra un'organizzazione e l'altra.

d'influenza dei media digitali, appare essenziale richiamare brevemente due concetti teorici più recenti e di grande rilievo in ambito comunicativo: l'agenda-setting e il framing.

Sebbene la prima formulazione di agenda-setting sia da attribuire a Cohen (1963), sono McCombs & Shaw (1972) a teorizzare il modello e a coniarne il termine: l'assunto di fondo dell'*agenda-setting theory* è che, i media, nel selezionare e gerarchizzare i temi dell'agenda pubblica, determinino la rilevanza percepita di tali temi dal pubblico. Organizzando l'orizzonte tematico del dibattito, i media non dicono al pubblico "cosa pensare", bensì "su cosa pensare". Shaw (1976) espone il concetto con chiarezza:

> The agenda setting hypothesis does not say that the media are trying to persuade […] The media, by describing and detailing what is out there, present people with a list of what to think about and talk about. […] For the basic claim of agenda-setting theory is that people's understanding of much of social reality is copied from the media.
> (*ivi*, pp. 96-101)[9]

Strettamente correlata all'agenda-setting è la *framing theory*, che riguarda invece il modo in cui i contenuti vengono – come indica il termine stesso – 'incorniciati' o 'inquadrati'. Dapprima teorizzato da Goffman (1974), il framing fu in seguito ampliato da altri autori, tra i quali Robert Entman (1993) si distinse per la sua applicazione ai media. L'autore afferma:

> Framing essentially involves *selection* and *salience*[10]. To frame is to *select some aspects of a perceived reality and make them more salient in a communicating text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation* for the item described. […] Frames, then, *define problems* – determine what a causal agent is doing with what costs and benefits […]; *diagnose causes* – identify the forces creating the problem; *make moral judgments* – evaluate causal agents and their effects; and *suggest remedies* – offer and justify treatments for the problems and predict their likely effects.
> (*ivi*, p. 52)[11]

---

[9] L'ipotesi dell'agenda setting non sostiene che i media cercano di persuadere […] I media, descrivendo e precisando la realtà esterna, presentano al pubblico una lista di ciò intorno a cui avere un'opinione e discutere. L'assunto fondamentale dell'agenda setting è che la comprensione che la gente ha di gran parte della realtà sociale è mutuata dai media. (Traduzione di Luca Corchia, *dispensa universitaria*, Università di Pisa, 2011)

[10] Con il termine *salience*, l'autore si riferisce all'atto d "making a piece of information more noticeable, meaningful, or memorable to audiences", ossia all'atto di rendere un'informazione più evidente, significativa e memorabile per il pubblico.

[11] Il framing implica essenzialmente *selezione e rilevanza*. "Inquadrare" significa *selezionare alcuni aspetti di una realtà percepita e renderli rilevanti in un testo comunicativo, in modo tale da promuovere una*

Il framing, dunque, si riferisce alla costruzione discorsiva delle informazioni: il modo in cui un'informazione, un contenuto o un tema vengono presentati "[…] indirizza verso una particolare 'chiave di lettura' degli eventi, dei soggetti o dei problemi raffigurati" (Corchia, 2011), contribuendo a plasmare la percezione pubblica delle questioni sociali presentate. L'*agenda-setting theory* e la *framing theory* furono originariamente formulate in riferimento ai media tradizionali, ma nello scenario odierno dominato dalle piattaforme digitali producono effetti egualmente impattanti, al punto da essere riattualizzate nelle forme della *networked agenda setting* e del *networked framing*. Alla luce di quanto affermato, risulta ora fondamentale chiarire il significato di "media digitali".

Cellulari e smartphone, personal computer, tablet e fotocamere digitali sono solo alcuni dei dispositivi che si integrano nella definizione di "media digitali", vale a dire tecnologie in grado di processare e diffondere informazione rappresentata da sequenze numeriche binarie, successivamente elaborata e tradotta in linguaggio umano. Secondo Delfanti e Arvidsson (2019), con il termine "media digitali" (o *digital media*) si identificano le tecnologie di comunicazione basate sui computer e sulle reti – o, più in generale, le tecnologie dell'informazione e della comunicazione basate sul codice digitale – diffusesi a partire dagli ultimi decenni del XX secolo, prima affiancando i mass media tradizionali e progressivamente integrandosi con essi.

I primi studi sull'ascesa dei media digitali accostavano la denominazione di *digital media* a quella di *new media* (o "nuovi media") un'interscambiabilità terminologica che oggi risulta problematica. Per quanto l'aggettivo "nuovo" consentisse di distinguerli dagli *old media* (o "media tradizionali") – i canali di comunicazione di massa pre-digitali come giornali, televisione o radio – esso risultava ambiguo in particolare per due motivi. In primo luogo, il concetto di novità resta fine a sé stesso se si considera che, da un lato, ogni medium è "nuovo" al momento della sua introduzione e, dall'altro, ogni "nuovo" medium è destinato a diventare "vecchio" se sostituito da tecnologie più recenti; in secondo luogo, il termine presupponeva una concezione lineare dell'evoluzione dei media, implicando che i *new media* fossero in qualche modo migliori degli *old media*.

particolare definizione del problema, un'interpretazione causale, una valutazione morale e/o un suggerimento su come affrontare il tema descritto. […] I frame, dunque, *definiscono i problemi* – determinano cosa un agente causale stia facendo, con quali costi e benefici […]; *diagnosticano le cause* – identificano i fattori che generano il problema; *formulano giudizi morali* – valutano gli agenti causali e i loro effetti; *e suggeriscono rimedi* – propongono e giustificano soluzioni ai problemi e ne prevedono i probabili effetti.

Ciononostante, la storia dell'evoluzione dei media dimostra come l'introduzione di un nuovo medium non comporti necessariamente la scomparsa di quelli precedenti: il nuovo non soverchia il vecchio, bensì lo integra o lo cambia. A tal proposito, Delfanti e Arvidsson (2019) enfatizzano come "[t]he introduction of television has not caused the disappearance of the newspaper. The introduction of the tablet did not cause the disappearance of the book. Rather, books evolved into different technological formats" (*ivi*, p. 7)[12]. Nell'illustrare questo processo evolutivo, gli autori prendono in prestito un termine coniato da Bolter & Grusin (2000), quello di *remediation*, un concetto che guarda all'evoluzione dei media come un fenomeno dinamico, caratterizzato non solo dalla competizione tra media differenti, ma anche dalla coevoluzione e cooperazione tra gli stessi:

> A practice, content, or format can be re-mediated by a new technology that mimics or reworks previous formats. [...] The concept of remediation allows the recognition that the history of media is a continuous, non-linear process that can go in several directions, as old and new media continue to influence each other.
> (Delfanti & Arvidsson 2019, p. 7)[13]

Sebbene sia fondamentale riconoscere come i media contemporanei non siano poi così distanti dai propri predecessori, non si può trascurare la caratteristica che più li distingue: la loro natura digitale. È proprio questo elemento costitutivo – e l'insieme di caratteristiche che esso implica – a differenziarli dai media analogici e a spiegare le modalità attraverso cui i media digitali incidono sulla società nel suo complesso. Si tratta difatti di strumenti pervasivi che, all'interno della società dell'informazione, oltre ad influenzare la sfera economica e politica, incidono profondamente su quella sociale e comunicativa, poiché ridefiniscono il modo in cui produciamo e distribuiamo l'informazione. Lo studio dei media digitali non si limita ad indagare i dispositivi tecnologici, ma include anche le piattaforme software, i protocolli di rete, le nuove forme di socialità nelle comunità online e le trasformazioni comunicative da essi innescate. Questo spiega come l'informazione, nell'era digitale, non solo viaggi ad una velocità

[12] "[l]'introduzione della televisione non ha causato la scomparsa del giornale. L'introduzione del tablet non ha causato la scomparsa del libro. Piuttosto, i libri si sono evoluti in differenti formati tecnologici".
[13] Una pratica, un contenuto o un formato possono essere ri-mediati da una nuova tecnologia che imita o rielabora formati precedenti. […] Il concetto di *remediation* consente di riconoscere che la storia dei media è un processo continuo e non lineare, che può seguire diverse direzioni, poiché media vecchi e nuovi continuano a influenzarsi reciprocamente.

esponenziale, ma sia in grado di raggiungere audience sempre più vaste e di influenzare la percezione che gli individui hanno di sé stessi e del mondo circostante.

1.2.2 Lo scenario *digital* nel mondo

Oltre ad essere digitali, i *digital media* sono convergenti, ipertestuali, mobili e distribuiti: contenuti diversi (scritti, visivi, sonori, etc.) convergono in un unico dispositivo tecnologico, sono fruiti in modo non lineare, sono accessibili a – e realizzabili da – chiunque, da qualsiasi luogo e in qualsiasi momento. Mentre i mass media tradizionali si distinguono per essere centralizzati e unidirezionali, per cui l'informazione viene trasmessa da un emittente ad un vasto pubblico indistinto, i media digitali operano secondo un modello distribuito: chiunque abbia accesso a Internet – un'infrastruttura aperta, decentralizzata e orizzontale (Castells 1996; Menduni 2007; Delfanti & Arvidsson 2019; Quarta & Smorto 2020) – può partecipare alla creazione e diffusione di contenuti.

A febbraio 2025, l'agenzia creativa We Are Social ha pubblicato, in collaborazione con Meltwater, il "Digital 2025 Global Overview Report", l'ultimo di una più che decennale serie di report annuali atti a comprendere lo scenario *digital* e *social* nel mondo. "Digital 2025" evidenzia come, su una popolazione mondiale di 8.20 miliardi di persone, 5.56 miliardi (il 67.9%) abbiano accesso a Internet, su cui trascorrono quotidianamente una media di 06:38 ore, contro i 5.32 miliardi che guardano la televisione – un dato che si fa portavoce del crescente predominio del digitale sui media tradizionali. Della totalità del tempo trascorso in rete giornalmente, il 35.3% (pari a 02:21 ore) è esclusivamente dedicato alla fruizione delle piattaforme social. A sostegno di queste osservazioni, le *user identities* sui social media ammontano a 5.24 miliardi[14], per una percentuale pari al 63.9% della popolazione globale ed un aumento del 4.1% rispetto a febbraio 2024. Pur tenendo in considerazione che tale dato non faccia riferimento a persone univoche, poiché un singolo individuo può possedere più account, esso si fa testimone di una società che vive ormai all'insegna della connessione, ed enfatizza il ruolo sempre più centrale che il digitale riveste nella vita quotidiana di ciascun individuo.

[14] All'inizio del 2025, lo scenario italiano conta invece 53.3 milioni di utenti Internet e 42.2 milioni di *user identities* sui social media, pari – rispettivamente – all'89.9% e al 71.2% di una popolazione complessiva di 59.3 milioni di persone. In Italia, il tempo medio di utilizzo di Internet è di 05:39 ore al giorno, di cui 01:48 ore trascorse sui social media.

Tuttavia, uno degli aspetti più rilevanti messi in luce dal report è il ricorso al digitale per una gamma di attività più ampia rispetto allo stesso periodo dell'anno precedente. Tra queste, come riportato sul sito weareoscial.com, "'[f]inding information' remains the single greatest motivation for going online at the start of 2025, with 62.8 percent of adult internet users stating that this is one of their main reasons for using the internet today"[15]. La *Figura 1* riportata di seguito, tratta direttamente dal report, consente di osservare nel dettaglio come questa motivazione si posizioni rispetto alle altre per ciascuna fascia anagrafica, risultando la principale ragione di utilizzo di Internet per gli utenti oltre i 35 anni.

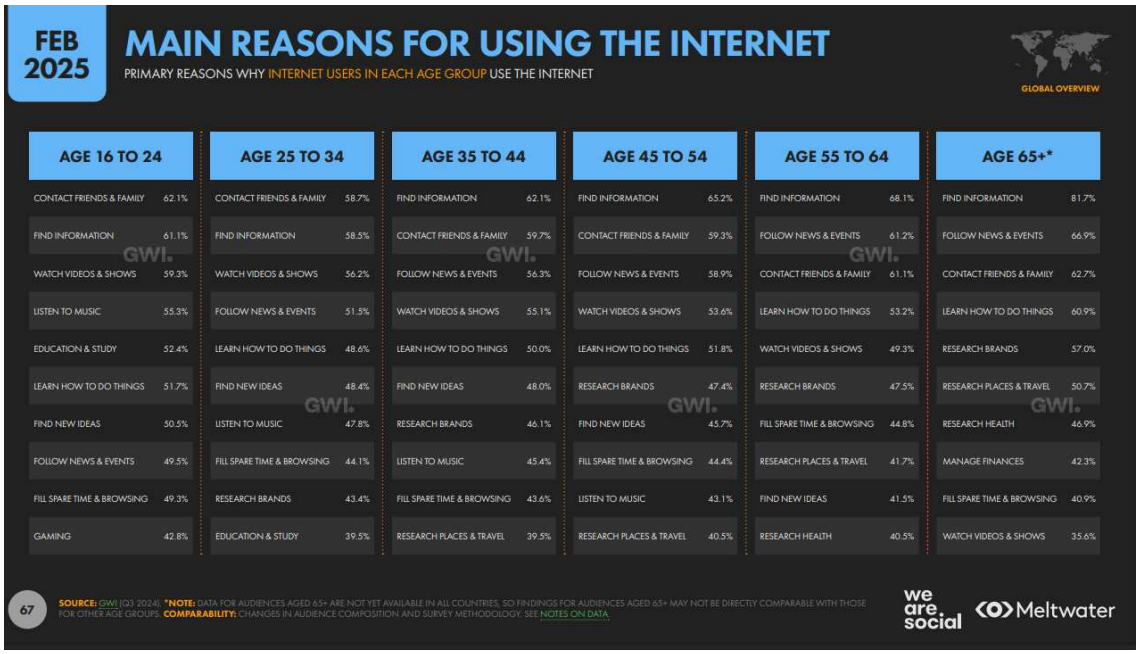

Figura 1 Principali motivazioni per l'utilizzo di Internet, suddivise per fasce d'età.
(Tratta da *Digital 2025 Global Overview Report*, We Are Social & Meltwater 2025)

Per quanto riguarda le fasce più giovani (16-24 anni e 25-34 anni), la ricerca di informazioni (*find information*) risulta seconda solo al contattare amici e parenti (*contact friends and family*). Ciononostante, i dati mostrano come le due tipologie di attività si collochino a livelli pressocché analoghi: con uno scarto percentuale marginale (pari all'1% per la fascia 16-24 anni e allo 0.2% per la fascia 25-34 anni), la ricerca di informazioni conferma l'importanza di Internet come fonte informativa anche per

[15] "'[*f*]*inding information*" (la ricerca di informazioni) rimane la motivazione principale per connettersi a Internet all'inizio del 2025: il 62.8% degli utenti adulti indica questo come uno dei motivi principali per l'utilizzo della rete al giorno d'oggi".

adolescenti e giovani adulti. "Che l'avvento di Internet abbia sottratto il monopolio della divulgazione delle notizie ai *legacy media* o media tradizionali è un dato incontrovertibile" (Nicodemo 2017, p. 21): la sua struttura reticolare rompe lo schema comunicativo tradizionale – quello verticale e centralizzato, tipico degli *old media* – e rende il flusso delle informazioni più orizzontale, distribuito e meno vincolato al controllo degli enti istituzionali.

Il suddetto carattere distribuito della rete consente di introdurre le due caratteristiche dei media digitali che maggiormente interessano l'obiettivo del presente studio: la loro interattività, che consente agli utenti di non essere meri destinatari ma attori nella produzione e nella diffusione dei contenuti, e la loro dimensione sociale, che fa della rete tanto un luogo di relazione quanto un luogo di costruzione collettiva del sapere.

1.3 I social media: le tassonomie in letteratura

La dimensione interattiva e partecipativa tipica dei media digitali raggiunge massima espressione in una delle loro declinazioni più emblematiche: i social media. Kietzmann et al. (2011) osservano:

> Traditionally, consumers used the Internet to simply expend content: they read it, they watched it, and they used it to buy products and services. Increasingly, however, consumers are utilizing platforms – such as content sharing sites, blogs, social networking, and wikis – to create, modify, share, and discuss Internet content. This represents the social media phenomenon […]. (*ivi*, p. 241)[16]

I social media, ricoprendo un ruolo centrale nella produzione e disseminazione dei contenuti informativi, si sono imposti negli ultimi due decenni come uno dei fattori di maggiore influenza nell'era digitale. Se si considera che il modo in cui l'informazione è presentata, diffusa e fruita può incidere radicalmente sui processi di definizione del dibattito, i social media – caratterizzati dalla partecipazione attiva degli utenti alla creazione, condivisione e discussione dei contenuti – rappresentano uno strumento decisivo nella costruzione dell'opinione pubblica (Boulianne 2019).

[16] Tradizionalmente, gli utenti utilizzavano Internet unicamente come strumento di fruizione dei contenuti: li leggevano, li guardavano e lo impiegavano per acquistare prodotti e servizi. Tuttavia, in misura crescente, gli utenti si servono di piattaforme – quali siti di condivisione di contenuti, blog, social network e wiki – per creare, modificare, condividere e discutere contenuti online. Questo fenomeno rappresenta ciò che viene comunemente definito *social media phenomenon* […].

Il termine "social media" appare per la prima volta nel 1994, anno in cui, con la crescente accessibilità al World Wide Web e all'Internet ad alta velocità, quest'ultimo penetra nelle nostre esistenze più profondamente di ogni altro mezzo di comunicazione (Siegel, 2011). Sebbene queste piattaforme rappresentino una delle principali aree di ricerca in molteplici ambiti disciplinari, sono pochi gli studiosi che hanno tentato di formularne una definizione, probabilmente a causa del loro carattere in costante evoluzione. Non esiste, infatti, una definizione universalmente accettata: "[…] there seems to be confusion among managers and academic researchers alike as to what exactly should be included under this term, and how Social Media differ from the seemingly-interchangeable related concepts of Web 2.0 and User-Generated Content[17]"(Kaplan & Haenlein 2010, p. 60)[18]. Un recente studio di Aichner et al. (2021) tenta di fare chiarezza ed evidenzia un significativo riorientamento nelle definizioni di "social media" elaborate dal 1994 al 2019: in termini generali, il loro prestigio in quanto strumento di mediazione nelle relazioni interpersonali resta invariato, ma le definizioni successive al 2010 registrano un'enfasi maggiore sui social media come spazio digitale in cui generare e condividere contenuti. Difatti, alla luce dei contributi pubblicati dopo il 2010 (Kaplan & Haenlein 2010; Kietzmann et al. 2011), Aichner & Jacob (2015) descrivono i social media come "[…] web-based applications and interactive platforms that facilitate the creation, discussion, modification and exchange of user-generated content" (*ivi*, p. 258)[19].

Nella sua accezione generale, il termine è utilizzato come termine-ombrello entro cui ricade una varietà di piattaforme online, quali i social network (es. Facebook, Instagram), i business network (es. LinkedIn), i blog, i forum (es. Reddit), le piattaforme video (es. YouTube, Twitch), il social gaming (es. Discord) – un ampio raggio di piattaforme che ne testimoniano un utilizzo altrettanto ampio e diversificato. Secondo il sopracitato "Digital 2025", aggiornato a febbraio 2025, circa il 94% degli utenti Internet utilizza attivamente i social media. Analogamente al quadro generale di utilizzo di Internet (cfr. *Figura 1*), le motivazioni che guidano la fruizione di queste piattaforme possono variare

[17] Per una trattazione dettagliata del concetto di *user-generated content*, cfr. *§1.4*.
[18] "[…] sembra esserci confusione, sia tra i manager sia tra i ricercatori accademici, su cosa debba esattamente rientrare sotto questo termine e su come i Social Media si differenzino dai concetti apparentemente intercambiabili di Web 2.0 e di User-Generated Content".
[19] "[…] applicazioni basate sul web e piattaforme interattive che favoriscono la creazione, la discussione, la modifica e lo scambio di contenuti generati dagli utenti".

significativamente in funzione della fascia anagrafica dell'utente. La *Figura 2* fornisce una visione d'insieme delle ragioni più comunemente riportate.

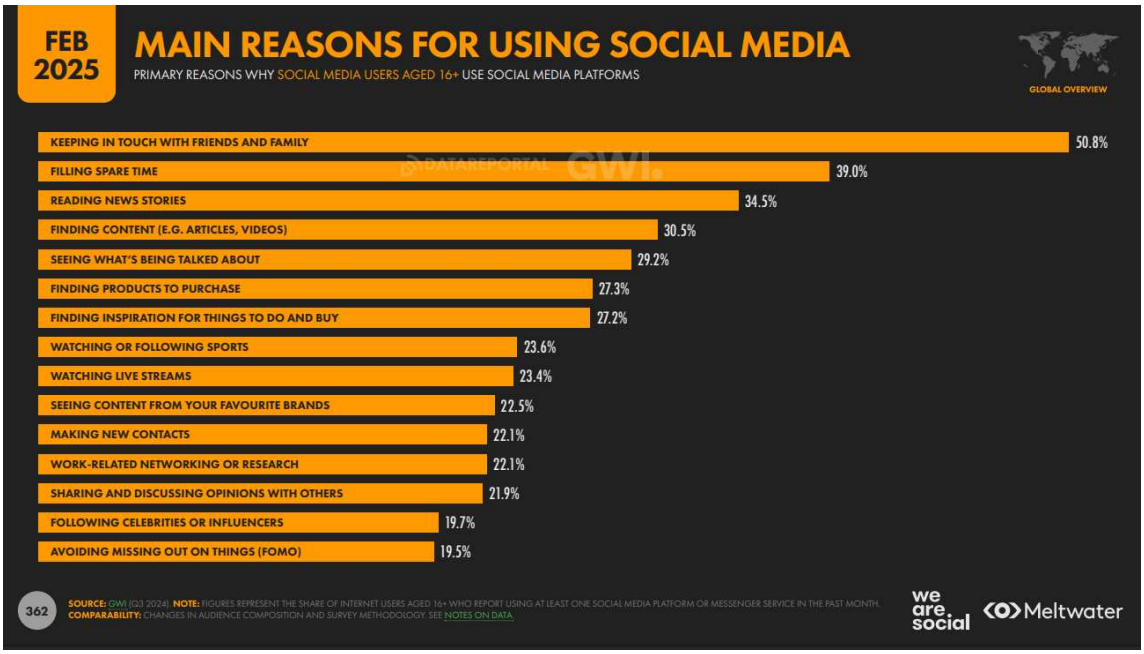

Figura 2 Principali motivazioni per l'uso dei social media (utenti 16+).
(Tratta da *Digital 2025 Global Overview Report*, We Are Social & Meltwater 2025)

In termini generali, la metà (50.8%) degli utenti attivi utilizza i social media principalmente per mantenere contatti con amici e familiari (*keeping in touch with friends and family*), risultando quindi la motivazione più frequentemente dichiarata: "[e]specially in relation to globalization and constant migration, [the use of social media] has become a vital tool […]. The need for transnational communication between family members and the people they left behind is of great importance" (Aichner et al. 2021, p. 216)[20]. A tale proposito, si precisa come l'agevolazione delle interazioni tra utenti costituisca una delle principali caratteristiche maggiormente condivise dalle diverse tipologie di social media (Aichner & Jacob 2015; Aichner et al. 2021).

Il grafico (si veda *Figura 2*) illustra inoltre come il 39% li utilizzi per occupare il tempo libero (*filling spare time*), attività che si colloca in seconda posizione. I social media, tuttavia, si configurano non solo come strumenti di intrattenimento, ma anche come canali primari di accesso all'informazione. Le tre posizioni successive sono difatti

[20] "[l'utilizzo dei social media] è diventato uno strumento fondamentale, soprattutto in relazione alla globalizzazione e alla migrazione costante […]. La necessità di comunicazione transnazionale con i familiari e le persone rimaste nel paese d'origine risulta di fondamentale importanza".

occupate da fome di utilizzo legate alle esigenze informative degli utenti: il 34.5% ricorre ai social media per consultare notizie (*reading news stories*)[21], il 30.5% per reperire contenuti come articoli e video (*finding content*), e il 29.2% per monitorare i temi discussi (*seeing what's being talked about*). Inoltre, in tredicesima posizione, con una percentuale pari al 21.9%, emerge l'uso dei social media per la condivisione ed il confronto di opinioni (*sharing and discussing opinions with others*), rafforzandone il ruolo come strumento di partecipazione e scambio informativo.

Questa sezione ha dunque evidenziato due aspetti che esplicitano con chiarezza il carattere eterogeneo, molteplice e sfaccettato dei social media: in primo luogo, il termine stesso – a cui si è difficilmente giunti a dare una definizione – racchiude comunemente un esteso ventaglio di piattaforme online; in secondo luogo, circa il 94% degli utenti Internet che utilizza i social media ne fa un uso intensivo per molteplici finalità (si veda *Figura 2*). Queste osservazioni legittimano la definizione dei social media come "multifunctional networking tools" (Koukaras et al. 2020, p. 295), strumenti di networking in grado di offrire una varietà sempre più ampia di servizi. La loro multifunzionalità, tuttavia, rende difficile determinarne lo scopo e la missione principali e, di conseguenza, la loro tipologia.

1.3.1 Dimensione mediatica e sociale: i 6 SMTs di Kaplan & Haenlein

La letteratura ha proposto nel corso degli anni diverse tassonomie dei social media – vale a dire, delle loro tipologie (*Social Media Types*, SMTs) – che, proprio a causa della continua nascita di nuove piattaforme e della costante innovazione di quelle esistenti, sono oggi considerate in larga parte superate o inadeguate (Koukaras et al. 2020). Un esempio è quello fornito dai già citati Kaplan & Haenlein (2010), che offrono una classificazione dei social media in virtù della loro dimensione mediatica e sociale. Gli autori definiscono la componente mediatica dei social media affidandosi alla teoria della *social presence* (Short et al., 1976) e alla teoria della *media richness* (Daft & Lengel, 1986), le quali sostengono che i media – e, in questo contesto, i social media –

[21] In Italia, *reading news stories* (47.1%) si rivela essere la principale motivazione di utilizzo dei social media, a conferma della loro centralità come fonte informativa rapida e accessibile. Seguono *filling spare time* e *keeping in touch with friends and family*, con le rispettive percentuali del 46.4% e del 43.4% (We Are Social & Meltwater 2025b).

differiscono, rispettivamente, per il grado di presenza sociale e per il grado di ricchezza che possiedono. Secondariamente, gli autori ne definiscono la componente sociale richiamando le nozioni di *self-presentation* e di *self-disclosure*: un ulteriore classificazione dei SMTs può essere realizzata sulla base del tipo di auto-presentazione che consentono e del grado di auto-rivelazione che richiedono. Per una migliore comprensione della suddetta tassonomia, si riassume di seguito ciascuno dei concetti menzionati.

La presenza sociale (*social presence*) è intesa come il livello di contatto acustico, visivo e fisico che è possibile instaurare tra due interlocutori. Maggiore è la presenza sociale garantita dal mezzo, maggiore sarà l'influenza reciproca che i partecipanti esercitano sul comportamento dell'altro: nei contesti mediati[22], una comunicazione sincrona (ad esempio, una chat in tempo reale o una conversazione telefonica) consente di raggiungere un grado di presenza sociale maggiore rispetto ad una comunicazione asincrona (come un'email o una piattaforma di messaggistica). La ricchezza mediatica (*media richness*) riguarda invece la quantità di informazioni che un mezzo consente di trasmettere in un determinato intervallo di tempo. Fondandosi sull'assunto che l'obiettivo di ogni comunicazione sia la risoluzione dell'ambiguità e la riduzione dell'incertezza, la teoria della *media richness* sostiene che maggiore è la ricchezza posseduta dal mezzo, maggiore sarà la sua efficacia nella riduzione di ambiguità e incertezza.

Il concetto della *self-presentation* afferma che, in ogni forma di interazione sociale, le persone sono guidate dal desiderio di controllare l'immagine che gli altri percepiscono di loro (Goffman 1959), che sia con l'obiettivo di influenzarli per ottenere riconoscimento o per costruire un'immagine di sé coerente con la propria identità personale. La *self-disclosure*, invece, è la rivelazione – consapevole o inconsapevole – di informazioni personali (pensieri, emozioni, preferenze, avversioni o, più in generale, opinioni) coerenti con l'immagine che si desidera proiettare. Si tratta di teorie che costituiscono un aspetto particolarmente utile nell'analisi delle dinamiche che governano le comunità online.

Riassumendo, questi concetti – applicati ai social media – consentono di attuare una prima classificazione mediatica e di distinguere le diverse tipologie di social media a partire dal livello di contatto acustico, visivo e fisico che essi consentono tra gli

[22] La *social presence theory* non considera esclusivamente l'immediatezza (sincrono vs. asincrono) del mezzo come elemento di influenza della presenza sociale, ma guarda anche al grado di intimità (interpersonale vs. mediato) che lo stesso consente di instaurare.

interlocutori (*social presence*) e dalla loro efficacia nel ridurre ambiguità e incertezza nella comunicazione (*media richness*); successivamente, una classificazione sociale può essere operata in funzione del tipo di controllo esercitabile dall'utente sulla rappresentazione di sé che gli altri formano (*self-presentation*) e del grado di rivelazione di informazioni personali che la piattaforma implica (*self-disclosure*). Combinando le due dimensioni, Kaplan & Haenlein (2010) suddividono i social media in sei categorie principali, come illustrato nella figura sottostante (si veda *Figura 3*): *blog*, *collaborative projects*, *social networking websites*, *content communities*, *virtual social worlds* e *virtual game worlds*.

		Social presence/ Media richness		
		Low	Medium	High
Self-presentation/ Self-disclosure	High	Blogs	Social networking sites (e.g., Facebook)	Virtual social worlds (e.g., Second Life)
	Low	Collaborative projects (e.g., Wikipedia)	Content communities (e.g., YouTube)	Virtual game worlds (e.g., World of Warcraft)

Figura 3 Classificazione delle 6 tipologie di social media secondo i criteri di *social presence/media richness* e *self-presentation/self-disclosure.* (Tratta da Kaplan & Haenlein 2010)

1.3.2 I 7 *functional building blocks* di Kietzmann et al.

Kietzmann et al. (2011) e Gundecha & Liu (2012) ampliano le tipologie proposte da Kaplan & Haenlein (2010), delineando a loro volta due classificazioni che propongono, rispettivamente, sette e undici SMTs. È opportuno precisare, tuttavia, che la tassonomia di Kietzmann et al. (2011) non mira tanto a categorizzare i social media in senso stretto, quanto a descriverne la struttura funzionale interna attraverso l'individuazione di sette componenti costitutivi (*functional building blocks*), di seguito indicati come "blocchi funzionali". Nonostante lo studio nasca con l'intento di ottimizzare le strategie aziendali sui social media, si ritiene in ogni caso meritevole di menzione in questa sottosezione per tre motivi principali: per il richiamo nella letteratura successiva, per lo slancio innovativo apportato nell'ambito della ricerca sui social media, e per la maggiore chiarezza con cui consente di comprenderne le dinamiche di funzionamento.

Kietzmann et al. (2011), nel riconoscere come i social media varino per scopo e funzionalità nella loro ampia e complessa ecologia, illustrano sette blocchi funzionali – *identity*, *conversations*, *sharing*, *presence*, *relationships*, *reputation*, *groups* – nel seguente schema a nido d'ape, che gli autori denominano *honeycomb framework* (si veda *Figura 4*):

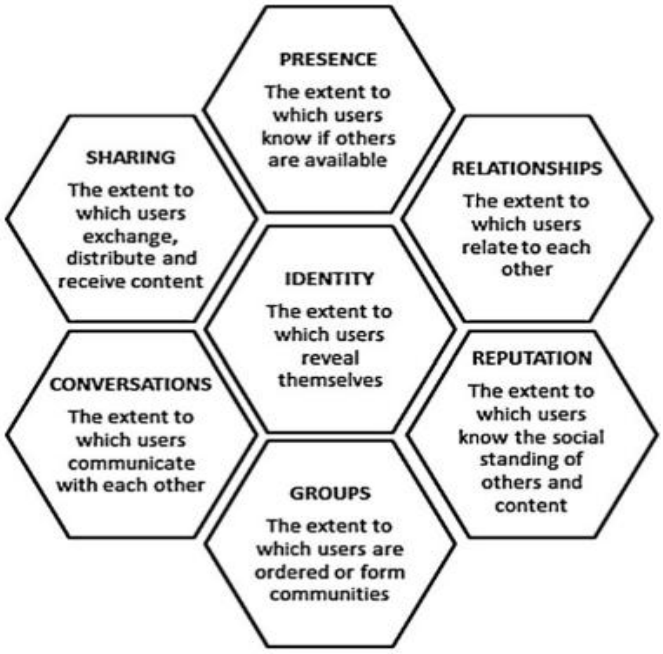

Figura 4 Rappresentazione grafica del modello a nido d'ape (*honeycomb framework*), raffigurante i 7 blocchi funzionali dei social media.
(Tratta da Kietzmann et al. 2011)

Ciascuno dei blocchi – "[which] are neither mutually exclusive, nor do they all have to be present in a social media activity" (*ivi*, p. 243)[23] – consente di analizzare un aspetto specifico dell'esperienza dell'utente: la rappresentazione grafica del modello enfatizza come i blocchi non esistano in modo isolato, bensì siano componenti interconnessi che, nel loro insieme, costituiscono l'esperienza completa dei social media. Ogni social media si distingue per il peso relativo attribuito a ciascun blocco, configurando così differenti livelli di funzionalità. Ad esempio, le piattaforme costruite sull'*identity building block*, come Instagram, Facebook o LinkedIn, aiutano gli utenti a modellare il modo in cui essere percepiti dagli altri, sia personalmente che professionalmente. È evidente come, in questo

[23] "[che] non sono né reciprocamente esclusive, né è necessario che siano tutte presenti all'interno di una piattaforma social".

contesto, il blocco funzionale dell'*identity* richiami perfettamente i già menzionati concetti di *self-presentation* e *self-disclosure*. In questa tipologia di piattaforme, la *self-disclosure* non riguarda esclusivamente la rivelazione di dati socio-demografici – quali nome, età, genere, professione e luogo di residenza o domicilio – ma anche la condivisione di interessi, valori e modelli comportamentali, che permette loro di costruire l'immagine di ciò che desiderano essere.

Le piattaforme maggiormente incentrate sulle *conversations*, che si riferisce alla misura con cui gli utenti comunicano tra di loro nello spazio digitale offerto dalla piattaforma, sono invece costruite primariamente con lo scopo di favorire il dialogo e lo scambio di opinioni. È il caso di X (precedentemente noto come Twitter), il cui funzionamento ruota attorno ad interazioni rapide fatte di brevi messaggi di testo, e di Reddit, basato sul *community dialogue* e su forum di discussione. Gli utenti, difatti, intrattengono conversazioni online per molteplici ragioni. In linea con i dati illustrati nella *Figura 2*, Keitzmann et al. (2011) evidenziano:

> People tweet, blog, et cetera to meet new like-minded people, to find true love, to build their self-esteem, or to be on the cutting edge of new ideas or trending topics. Yet others see social media as a way of making their message heard and positively impacting humanitarian causes, environmental problems, economic issues, or political debates.
> (*ivi*, p. 244)[24]

Il blocco funzionale dello *sharing* incorpora quello stesso concetto di socialità sotteso dall'aggettivo "social" all'interno del sintagma "social media": la condivisione online mette in luce i rapporti diretti tra gli utenti all'interno delle reti sociali. Essa, tuttavia, assume significato solo attraverso i contenuti oggetto di condivisione. Sono questi stessi oggetti, come immagini, video o testi, a fungere da mediatori dei legami sociali: "[w]ithout these objects, a sharing network will be primarily about connections between people but without anything connecting them together" (*ivi*, p. 245)[25]. Lo *sharing* è alla base di piattaforme che vedono la condivisione e la diffusione di contenuti come

[24] Le persone twittano, scrivono blog e così via per incontrare nuovi individui con interessi simili, per trovare l'amore, per costruire la propria autostima o per essere aggiornate sulle idee più innovative e sugli argomenti di tendenza. Altre, invece, considerano i social media come un mezzo per far sentire la propria voce e influenzare positivamente cause umanitarie, problematiche ambientali, questioni economiche o dibattiti politici.

[25] "[s]enza questi oggetti, una rete basata sulla condivisione riguarderebbe principalmente le connessioni tra le persone, ma senza alcun elemento che le leghi [realmente] tra loro".

principale modalità di interazione tra gli utenti. Instagram e TikTok, ad esempio, incorporano opzioni di condivisione e di repost che ne amplificano la portata.

Piattaforme come WhatsApp e Snapchat, invece, enfatizzano il blocco funzionale della *presence*. Gli stati di attività ("online/offline", "ultimo accesso", "sta scrivendo…") e la localizzazione in tempo reale su queste piattaforme consentono agli utenti di conoscere la disponibilità online e, in alcuni casi, la posizione geografica altrui, introducendo un senso di urgenza e di immediatezza nelle interazioni.

Il blocco funzionale delle *relationships* rappresenta la misura in cui gli utenti instaurano connessioni ed interagiscono tra di loro nello spazio digitale della piattaforma. LinkedIn, ad esempio, è incentrato sull'instaurazione di nuove relazioni tra utenti – formali, strutturate e professionali – tramite espansione della rete sociale; altre piattaforme, al contrario, si fondano sul mantenimento di relazioni già esistenti, più informali, personali e meno strutturate.

Le piattaforme incentrate sulla *reputation* attribuiscono un'elevata importanza alla credibilità e all'affidabilità sia degli utenti sia dei contenuti da questi prodotti. Essa può assumere forme diverse a seconda della piattaforma, ma si manifesta spesso attraverso metriche quantitative – quali il numero di likes, commenti, condivisioni, followers, valutazioni e recensioni – che fungono da indicatori di fiducia collettiva. Un esempio emblematico è costituito da Reddit, in cui la reputazione degli utenti è espressa in termini di "punteggio karma": attraverso un sistema di *upvotes* e *downvotes*, la community può valutare positivamente o negativamente un contenuto e, conseguentemente, promuoverlo o segnalarlo, determinandone quindi la visibilità all'interno della piattaforma.

Il blocco funzionale dei *groups* assume particolare rilievo nelle piattaforme caratterizzate dalla costruzione di *communities* e *subcommunities*, spazi collettivi nei quali gli utenti accomunati da interessi o valori hanno la possibilità di riunirsi. Tali piattaforme sono spesso regolate da amministratori e moderatori, che si occupano di monitorare le interazioni e i contenuti pubblicati dai membri, sebbene in alcuni casi la gestione sia affidata alla comunità stessa. Esempi di piattaforme incentrate sui *groups* sono Discord, che si distingue per un'organizzazione in server e canali, e Reddit, con una struttura decentralizzata in *subreddits* gestiti dagli utenti.

1.3.3 I 9 SMTs di Gundecha & Liu

Come anticipato, un'ulteriore tassonomia degna di nota è quella elaborata da Gundecha & Liu (2012), che precisano come i social media offrano agli utenti "[…] an easy-to-use way to communicate and network with each other on an unprecedented scale and at rates unseen in traditional media" (*ivi*, p. 2)[26]. Ampliando a loro volta la classificazione di Kaplan & Haenlein (2010), gli autori propongono una categorizzazione dei social media in nove SMTs, ciascuna contraddistinta da specifiche funzionalità e modalità di interazione. La *Figura 5* ne sintetizza le caratteristiche principali e fornisce degli esempi per ciascuna delle categorie identificate:

Type	Characteristics
Online social networking	Online social networks are Web-based services that allow individuals and communities to connect with real-world friends and acquaintances online. Users interact with each other through status updates, comments, media sharing, messages, etc. (e.g., Facebook, Myspace, LinkedIn).
Blogging	A blog is a journal-like website for users, aka bloggers, to contribute textual and multimedia content, arranged in reverse chronological order. Blogs are generally maintained by an individual or by a community (e.g., Huffington Post, Business Insider, Engadget).
Microblogging	Microblogs can be considered same a blogs but with limited content (e.g., Twitter, Tumblr, Plurk).
Wikis	A wiki is a collaborative editing environment that allow multiple users to develop Web pages (e.g., Wikipedia, Wikitravel, Wikihow).
Social news	Social news refers to the sharing and selection of news stories and articles by community of users (e.g., Digg, Slashdot, Reddit).
Social bookmarking	Social bookmarking sites allow users to bookmark Web content for storage, organization, and sharing (e.g., Delicious, StumbleUpon).
Media sharing	Media sharing is an umbrella term that refers to the sharing of variety of media on the Web including video, audio, and photo (e.g., YouTube, Flickr, UstreamTV).
Opinion, reviews, and ratings	The primary function of such sites is to collect and publish user-submitted content in the form of subjective commentary on existing products, services, entertainment, businesses, places, etc. Some of these sites also provide products reviews (e.g., Epinions, Yelp, Cnet).
Answers	These sites provide a platform for users seeking advice, guidance, or knowledge to ask questions. Other users from the community can answer these questions based on previous experiences, personal opinions, or relevent research. Answers are generally judged using ratings and comments (e.g., Yahoo! answers, WikiAnswers).

Figura 5 Classificazione delle 9 tipologie di social media e relative caratteristiche. (Tratto da Gundecha & Liu 2012)

[26] "[…] un mezzo semplice per comunicare e interconnettersi su una scala senza precedenti e a ritmi inediti rispetto ai media tradizionali".

1.3.4 I 3 SMTs di Koukaras et al.: *Entertainment*, *Profiling* e *Social networks*

Dopo aver esaminato le tassonomie più rappresentative presenti in letteratura, è possibile ora introdurre uno studio più recente che, sviluppandosi a partire dalle tre classificazioni illustrate, offre una prospettiva aggiornata sulle tipologie di social media. In una ricerca del 2020, Koukaras et al. adottano un approccio empirico con lo scopo di restringere e aggiornare le precedenti standardizzazioni dei *Social Media Types* (SMTs), e offrire un'alternativa che rifletta più accuratamente lo stato di evoluzione delle piattaforme social. Gli autori conducono due esperimenti, incentrati rispettivamente sull'*association rule mining* e sul *clustering*, al termine dei quali propongono una nuova categorizzazione sulla base dei servizi e delle funzionalità ufficiali (*utilities*) che le piattaforme offrono. In particolare, l'utility primaria di ciascun SMT ne struttura la funzione d'uso dominante. Koukaras et al. (2020) raggruppano dunque le piattaforme in insiemi coerenti e riducono i social media a tre tipologie fondamentali: *Entertainment networks*, *Profiling networks* e *Social networks.* La categoria degli *Entertainment networks* include piattaforme dedicate all'intrattenimento di vario genere, come giochi, sport, cinema e viaggi; è una tipologia di social media che offre come *utility* primaria, appunto, l'*entertainment*. I *Profiling networks*, invece, costituiscono una SMT che descrive quei social media caratterizzati da funzionalità volte a promuovere competenze, obiettivi e identità personali o professionali, e che presentano dunque come *utility* primaria il *profiling*[27]. Infine, la SMT dei *Social networks* deriva dalla fusione dei cosiddetti *General Purpose Networks* emersi nei due esperimenti, ovvero un gruppo più vario ed eterogeneo di piattaforme il cui comun denominatore è il fatto di possedere una funzione dominante diversa dall'*entertainment* e dal *profiling* osservati nei due casi precedenti. Si tratta difatti di social media che condividono un orientamento improntato alla socialità, alla condivisione e all'interazione. La categoria dei *Social networks* comprende pertanto piattaforme che possono ruotare attorno a una tra molteplici *utilities* primarie possibili: *connecting*, *multimedia*, *professional*, *sharing*. La *Tabella 1* riassume le utilities primarie, secondarie e accessorie (*trivia*) delle tre tipologie di SMTs identificate:

[27] In questo contesto, il *profiling* non fa riferimento alla profilazione (termine che evoca la raccolta di dati personali), ma è da intendersi come l'auto-rappresentazione e la costruzione identitaria dell'utente attraverso la piattaforma.

	Primary	*Secondary*	*Trivia*
Entertainment networks	Entertainment	Connecting, Multimedia, Opinions	Sharing, Privacy, News, Promoting, Voting, Publishing, Schedule, Profile, Applications, Professional
Profiling networks	Profiling	Connecting, Multimedia, Professional, Opinions, Publishing, Privacy, Voting, Applications, Promoting	News, Schedule, Entertainment
Social networks	Connecting, Multimedia, Professional, Sharing	Publishing	Privacy, News, Promoting, Voting, Schedule, Profile, Applications, Opinions, Entertainment

Tabella 1 Le tre tipologie di SMTs e le rispettive utilities primarie, secondarie e accessorie (*trivia*). (Elaborazione dell'autore a partire da Kourkaras et al. 2020).

Da un lato, gli studi discussi fino ad ora introducono l'oggetto di analisi di questa ricerca: le tipologie di social media appena illustrate, divenute tema di primaria importanza negli studi sulla comunicazione, si configurano difatti come applicazioni basate sugli *user-generated content* (Naab & Sehl 2017), di norma abbreviati con l'acronimo UGC. Dall'altro, rappresentano un quadro di riferimento utile per la definizione della metodologia. Nel capitolo successivo, difatti, le tassonomie di Kaplan & Haenlein (2010), di Gundecha & Liu (2012) e di Koukaras et al. (2020), nonché l'*honeycomb framework* di Kietzmann et al. (2011), verranno riprese al fine di delineare le caratteristiche funzionali e strutturali di Reddit, piattaforma che costituirà l'arena dell'analisi statistica di questa tesi.

1.4 *User-Generated Content* (UGC)

È proprio il fenomeno degli *User-Generated Content* (UGC) a segnalare in maniera evidente il passaggio epocale dai media tradizionali ai social media, spazi sociali in cui gli utenti stessi creano e condividono contenuti. Se i media tradizionali si contraddistinguono storicamente per consegnare i propri messaggi in maniera unidirezionale, con un grado minimo – se non assente – di interazione con il pubblico, i

social media si contraddistinguono invece per essere media 'partecipati e partecipativi' (Ferraresi & Schmitt 2018). La pratica dell'essere al contempo produttori e consumatori prende il nome di "prosumerismo", termine che riprende la nozione di *prosumer* coniata dal futurista americano Alvin Toffler (1980) per identificare un soggetto contemporaneamente *producer* e *consumer*. Gli *user-generated content* rappresentano oggi la manifestazione più evidente di questa pratica.

Come già evidenziato da Kaplan & Haenlein (2010), occorre distinguere il concetto di "social media" da quelli di "Web 2.0" e "*user-generated content*". Il Web 2.0, termine coniato nel 2004 per sancire il passaggio da una rete statica (Web 1.0) ad una interattiva e collaborativa, costituisce la base tecnica su cui i social media si fondano: i social media si configurano quindi come una specifica applicazione del Web 2.0. Su queste piattaforme, gli utenti creano e rendono pubblici i propri contenuti, i cosiddetti UGC, i quali "[…] can be seen as the sum of all ways in which people make use of Social Media. The term, which achieved broad popularity in 2005, is usually applied to describe the various forms of media content that are publicly available and created by end-users" (*ivi*, p. 61)[28]. Potremmo riassumere che l'acronimo inglese "UGC" è già di per sé autoesplicativo, e designa per l'appunto i contenuti generati dagli utenti.

Nel ripercorrere le origini e l'evoluzione del concetto di *user-generated content*, Kaplan & Haenlein (2010) avanzano due precisazioni fondamentali, che, applicate allo scenario odierno, acquisiscono un'attualità ancora maggiore. In primo luogo, riconoscono i fattori tecnologici, economici e sociali che ne hanno determinato la pervasività. Il già citato report di We Are Social e Meltwater, "Digital 2025", testimonia la maggiore disponibilità di connessioni a banda larga e l'ampliata accessibilità dei dispositivi per la creazione di UGC: al giorno d'oggi, il numero di persone connesse a Internet è di oltre due volte superiore a quello delle persone offline; il 97.8% degli utenti Internet possiede uno smartphone, il 58.3% un computer portatile o fisso, e oltre il 94% del totale delle connessioni mobili è a banda larga (dunque ad alta velocità e in grado di trasmettere grandi quantità di dati). Da un punto di vista sociale, invece, emerge un ulteriore fattore degno di nota: a contribuire significativamente alla produzione dei contenuti online sono le nuove generazioni di nativi digitali e *screenagers*. Si tratta di due termini che

[28] "[…] possono essere intesi come l'insieme di tutti i modi in cui le persone fanno uso dei Social Media. Il termine, che ha acquisito ampia popolarità nel 2005, è solitamente impiegato per descrivere le diverse forme di contenuto mediatico rese pubblicamente disponibili e create dagli utenti finali".

descrivono, rispettivamente, coloro che sono nati e cresciuti con Internet e i media digitali (Prensky 2001; Zampieri et al. 2018), e i giovani inclini ad un utilizzo intensivo dei dispositivi tecnologici. Più recentemente, la denominazione di *screenagers* viene impiegata in riferimento alla Generazione Alpha (i nati dopo il 2010), che si caratterizza per una maggiore – e più precoce – esposizione agli schermi degli *smart devices* (Ziatdinov & Cilliers 2021; Marius Drugaș 2022). Ad ogni modo, nativi digitali e *screenagers* identificano comunemente le fasce più giovani di utenti che si contraddistinguono per le loro competenze tecnologiche ed una forte propensione a interagire online.

In secondo luogo, Kaplan & Haenlein (2010) riportano i criteri, proposti nel 2007 dall'*Organisation for Economic Cooperation and Development* (OECD), che un contenuto deve soddisfare per essere classificato come UGC (o UCC, ossia *User-Created Content*, come si legge sul report). Potremmo riassumere che un contenuto generato o creato dagli utenti può essere considerato tale se rispetta tre requisiti fondamentali: un requisito di pubblicazione, un requisito di creatività e un requisito di non professionalità. Nello specifico, la prima condizione riguarda la pubblicazione del contenuto in uno spazio pubblico e accessibile, "[…] for example on a publicly accessible website or on a page on a social networking site only accessible to a selected group of people (e.g. fellow university students) […]" (OECD 2007, p. 18)[29]. La seconda prevede la presenza di un certo grado di contributo creativo: indipendentemente dal fatto che si tratti di un contenuto creato ex novo o dell'adattamento di un'opera già esistente, "[…] users must add their own value to the work. UCC could include user uploads of original photographs, thoughts expressed in a blog or a new music video" (*ibid.*)[30]. Infine, la terza condizione riguarda la creazione del contenuto al di fuori di contesti professionali o commerciali, e dunque la relativa assenza di aspettative di remunerazione o profitto da parte di chi li crea: "[u]ser-created content […] often does not have an institutional or commercial market context […]" (*ibid.*)[31] ed è invece generato dall'utente con il semplice scopo di esprimere la

[29] "[…] ad esempio, su un sito web accessibile pubblicamente oppure su una pagina di un social network accessibile soltanto a un gruppo selezionato di persone (es. compagni di università) […]".
[30] "[…] gli utenti devono apportare un contributo personale all'opera. Gli UCC possono includere [ad esempio] fotografie originali caricate dagli utenti, riflessioni espresse in un blog o la realizzazione di un nuovo video musicale".
[31] "[u]n contenuto creato dagli utenti […] spesso non si colloca in un contesto istituzionale o commerciale […]".

propria opinione, connettersi con gli altri utenti e/o acquisire fama, notorietà o prestigio. Questa selezione, dunque, esclude dalla definizione di UGC e-mail e messaggi istantanei, mere repliche di materiali già esistenti e qualsiasi tipo di contenuto realizzato in contesti commerciali. A costituire una delle forme più diffuse di *user-generated content* sono i commenti online pubblicati dagli utenti (Stroud et al. 2016; Newman et al. 2019), i quali soddisfano ciascuno dei criteri proposti dall'OECD.

1.4.1 I commenti online

I media tradizionali sono spesso percepiti come una fonte credibile di informazioni: essi si fondano sulla presenza di *gatekeeper* professionali – quali giornalisti ed editori – che, nella selezione e strutturazione dei contenuti, ne garantiscono accuratezza e affidabilità (McQuail 2010). Al contrario, la cultura partecipativa su cui poggia il modello di funzionamento dei social media si riflette inevitabilmente sull'attendibilità delle informazioni che circolano sugli stessi. Dobber & Hameleers (2024) sintetizzano perfettamente le opportunità e i rischi di questo contesto interattivo:

> There are few rules online, and many voices to be heard, which can be empowering for marginalized groups. However, beyond the potential for the inclusion of diverse voices and the empowerment of marginalized or suppressed groups, through social media, citizens can use ungated comment sections to interact with news, and herewith be exposed to unruly and noisy comments that may contain delegitimizing content aimed at established information sources such as the news media.
> (*ivi*, p. 4)[32]

Arena del coinvolgimento attivo degli utenti consentito negli ultimi anni dalla digitalizzazione – e dalle innovazioni socio-tecnologiche che ne sono derivate – sono le *comment sections* di piattaforme web e social media, aree che si configurano oggi come uno spazio privilegiato in cui il pubblico ha la possibilità di esprimere liberamente le proprie opinioni, nonché di interfacciarsi con prospettive differenti, talvolta anche diametralmente opposte alle proprie. Ne consegue che "[s]tudying these comments

[32] Online vigono poche regole, e le voci da ascoltare sono molte, il che può rappresentare un fattore di emancipazione per i gruppi marginalizzati. Tuttavia, oltre al potenziale inclusivo offerto dalla presenza di voci eterogenee e alla possibilità di dare voce a gruppi emarginati o repressi, attraverso i social media i cittadini possono utilizzare le sezioni commenti aperte per interagire con le notizie, esponendosi così a contenuti disordinati e caotici, che possono includere messaggi delegittimanti nei confronti delle fonti d'informazione consolidate, come i media tradizionali.

becomes one way of gauging the pulse of the public debate" (Douai & Nofal 2012, p. 269)[33].

Numerosi studi testimoniano l'ampio ricorso a questa declinazione di *user-generated content*, sebbene la maggior parte si focalizzi non tanto sui commenti in senso generale, quanto sui *news comments*, ossia i commenti pubblicati dagli utenti in risposta a contenuti di natura giornalistica. A tal riguardo, l'indagine denominata *Survey of Commenters and Comment Readers* (Stroud et al. 2016) offre una panoramica sul contesto statunitense del decennio scorso, illustrando anche dati relativi ai commenti online nell'accezione più ampia del termine. Lo studio dedica difatti alcune sezioni ai commenti riferiti a qualsiasi tipologia di contenuto digitale, riportando come il 55% degli statunitensi abbia lasciato almeno una volta un commento online e come il 77.9% abbia letto commenti pubblicati in rete da altri utenti. La *Figura 6* fornisce un quadro più dettagliato circa le percentuali di pubblicazione e di lettura dei commenti online. Si includono inoltre gli spazi digitali più frequentemente utilizzati per svolgere le suddette attività.

Percentage Reading and Leaving Online Comments

Post and Read Online Comments	Post Online Comments, but Do Not Read Them	Do Not Post Online Comments, but Read Them	Neither Read nor Post Online Comments
53.3%	1.7%	24.6%	20.2%

Where People Read and Leave Comments

	Where People Comment (among those who have left a comment)	Where People Read Comments (among those who read comments, but have not left one on news)
A social media site or app	77.9%	69.9%
A product /service review site or app	52.8	62.2
A news site or app	14.6	41.8
An entertainment site or app	6.5	22.9
Other	7.3	6.3

Figura 6 Attività di lettura e pubblicazione dei commenti online, e spazi digitali impiegati. (Tratto da *Survey of Commenters and Comment Readers*, Stroud et al. 2016)

I risultati dello studio (si veda *Figura 6*) rendono evidenti due aspetti cruciali. In primo luogo, gli utenti che pubblicano commenti online agiscono difficilmente in modo isolato, aderendo anzi ad una dinamica partecipativa all'interno della community: solo l'1.7% di

[33] "[l]o studio di questi commenti diventa un modo per misurare il polso del dibattito pubblico".

coloro che postano un commento si preclude la lettura dei commenti altrui. Il dato, dunque, suggerisce come la principale motivazione che induce l'utente a commentare sia l'inserimento nel discorso pubblico, all'interno del quale contestualizzare il proprio intervento, confrontarsi con le visioni opposte e integrare – rispondere, argomentare, confermare o contraddire – quanto già espresso. Commentando lo stesso contenuto e facendo riferimento ai commenti altrui, l'intervento di più utenti dà origine ad un dibattito pubblico online (Ziegele & Quiring 2013), all'interno del quale i contenuti che esprimono un punto di vista specifico possono attirare l'attenzione e orientare la conversazione, o influenzare il modo in cui gli utenti interpretano il contenuto originario (Horne et al. 2017). Questa osservazione introduce in maniera coerente il secondo aspetto cruciale emerso dal *Survey of Commenters and Comment Readers* (Stroud et al. 2016): la *Figura 6* rivela una percentuale alquanto significativa di utenti passivi, comunemente denominati *lurker*. Il 24.6% degli utenti, pur non contribuendo attivamente al dibattito, legge i commenti altrui – spesso al fine di ottenere una migliore comprensione delle tematiche in questione (Metzger et al. 2010) – assumendo quindi il ruolo di fruitori non contributivi. Dall'interpretazione dei dati emerge inoltre come i social media costituiscano il luogo preferenziale di pubblicazione dei commenti: tra il 55% di coloro che hanno dichiarato di aver pubblicato almeno un commento online, il 77.9% lo ha fatto per mezzo dei social media, contro il 14.6% dei siti di informazione online (*news site or app*). Il distacco percentuale è minore per quanto riguarda la sola lettura, ma conferma la predominanza dei social media (69,9%).

È lecito pertanto affermare che l'atto di scrivere e leggere commenti online è ormai parte integrante del consumo quotidiano delle piattaforme social, in molte delle quali le discussioni online nella forma di commenti scritti costituiscono un elemento centrale del loro funzionamento. Parlare dei dibattiti in rete significa fare riferimento a quelle che vengono comunemente definite *threaded online conversations*, ovvero conversazioni tra due o più utenti che seguono una struttura reticolare ad albero (*tree network structure*). Pubblicando un messaggio iniziale (che nei forum online assume il ruolo di post o submission), un utente può dare origine a un thread: gli altri utenti della community possono rispondere direttamente al post originario oppure inserirsi all'interno della discussione rispondendo ai commenti già pubblicati da altri utenti (Aragón et al. 2017), generando pertanto una gerarchia di risposte. I dati appena illustrati (cfr. *Figura 6*) e il

suddetto andamento sequenziale delle pubblicazioni assumono particolare rilievo se interpretati alla luce di un filone di crescente attenzione in letteratura: i commenti espressi online possono – in modo più o meno significativo – incidere sulle opinioni, sulle emozioni, sugli atteggiamenti e sulle percezioni degli utenti che ne fruiscono in qualità di lettori (Chen & Xia 2024; Gearhart et al. 2023; Gearhart et al. 2020).

2. Metodologia di ricerca

2.1 Il percorso di analisi

La presente ricerca si colloca all'interno di un approccio metodologico misto (*mixed-method approach*) che integra componenti qualitative e quantitative: l'esecuzione di un'analisi statistica dei dati testuali (ASDT) si affida a un'attività di restituzione e interpretazione dei dati emersi, consentendo di identificare, nel caso in esame, la molteplicità di declinazioni assunte da una tematica fortemente polarizzante, di chiarirne gli aspetti emergenti e di cogliere, tra questi, sistemi di relazione latenti (*distant reading*). L'elaborato persegue dunque un duplice obiettivo. Innanzitutto, esso si propone di evidenziare il potenziale dell'analisi statistica dei dati testuali come strumento metodologico per l'esplorazione dell'opinione pubblica. Quest'ultima, nell'attuale *information* e *network society*, si esprime nei commenti online di utenti impegnati ad interagire entro l'ambiente digitale. Si comprende, pertanto, come i commenti online costituiscano una preziosa fonte empirica per l'analisi quantitativa: *le threaded online conversations* si traducono oggi in corpora testuali di grandi dimensioni in grado di offrire uno spaccato autentico delle opinioni degli utenti su argomenti socialmente controversi.

Al fine di conseguire di tale obiettivo primario, è stato fondamentale selezionare una tematica sociale in grado di soddisfare tre criteri: attualità, opinabilità e, in particolare, ampia disponibilità di dati testuali. Sin dal rilascio della versione gratuita di ChatGPT, l'intelligenza artificiale (di seguito anche indicata come AI, dall'inglese *Artificial Intelligence*) è entrata prepotentemente nelle scuole, nelle aule universitarie, negli uffici e in ogni ambito della vita quotidiana, fino a estendersi persino alla sfera relazionale e psicologica di coloro che ne fanno un uso sempre più intensivo. Ponendosi al centro di un dibattito pubblico in continua espansione e fortemente polarizzato, l'AI è pertanto emersa come tematica particolarmente idonea ai fini della presente ricerca. L'esplorazione del modo in cui si articola il discorso online sull'intelligenza artificiale costituisce l'obiettivo secondario dell'elaborato. Trattandosi di un tema prettamente strumentale, che funge cioè da strumento funzionale a mostrare le potenzialità metodologiche dei metodi statistici, non si è ritenuto necessario approfondirne la cornice teorica. Saranno i capitoli successivi, dedicati all'elaborazione e alla discussione dei risultati ottenuti, a riportare informazioni e dati circa l'intelligenza artificiale, con il solo scopo di fornirne un quadro contestuale e facilitare l'interpretazione qualitativa dei

risultati. Si conclude che i due obiettivi sono tra loro consequenziali e strettamente interconnessi. Per il loro conseguimento, la presente ricerca ha adottato Reddit, il più grande forum di discussione online, come contesto empirico per la conduzione dell'analisi statistica dei dati testuali e per l'esplorazione del dibattito pubblico sull'intelligenza artificiale.

Il presente capitolo mira a fornire una cornice teorica del contesto empirico di estrazione dei dati e dei metodi statistici adottati per condurre l'analisi. Un percorso di analisi consta tipicamente di tre fasi: la raccolta dei dati (acquisizione), la loro rielaborazione in misure statistiche (sintesi) e la discussione dei risultati ottenuti (restituzione). Conformemente all'ordine di queste tre fasi canoniche, il capitolo seguente si occuperà innanzitutto di illustrare i criteri adottati per l'acquisizione dei dati testuali: un dataset di commenti estratti da Reddit, una delle piattaforme online maggiormente utilizzate per la condivisione e la discussione di contenuti. Secondariamente, verranno presentati i primi parametri relativi al corpus e la loro rielaborazione in proprietà misurabili. Saranno infine il quarto e il quinto capitolo della ricerca, a forte connotazione qualitativa, a dedicarsi alla rappresentazione, alla contestualizzazione e all'interpretazione dei risultati ottenuti, passaggi fondamentali affinché questi siano utilizzabili e utilizzati (Bernardi & Campostrini 2005, in Tuzzi 2024).

2.2 Il contesto empirico: Reddit

Se i primi studi nell'ambito delle scienze sociali computazionali si sono focalizzati sull'indagare la comunicazione *one-to-one*, la nascita e lo sviluppo dei nuovi canali di comunicazione esplorati nel primo capitolo hanno reso possibile lo studio delle discussioni collettive (Aragón et al. 2017; Medvedev et al. 2019) che hanno origine nell'ambiente digitale dei social media. Tra questi, Reddit, arena dell'analisi statistica di questa ricerca, costituisce oggigiorno il più grande forum di discussione online, al punto tale da autoproclamarsi la "front page of the Internet".

Introdotto nel 2005, Reddit è oggi uno dei siti web più visitati al mondo[34], figurando al nono posto (#9) per Similarweb e al sesto (#6) per Semrush. I due strumenti di analisi,

[34] In Italia, un'indagine della GlobalWebIndex (GWI) sulle piattaforme social più utilizzate mensilmente – citata nel report Digital 2025 – colloca Reddit in quindicesima posizione (#15).

entrambi fonti integrate nel report "Digital 2025" (We Are Social & Melwater, 2025a), forniscono stime di traffico mensile che variano tra 642 milioni e 1.03 miliardi di visitatori unici[35], e tra 3.50 miliardi e 5.94 miliardi di visite totali al sito. Tra gli utenti attivi che dichiarano di utilizzare Reddit per una varietà di attività, il 30.9% lo impiega per seguire o informarsi su marchi e prodotti, il 30.7% per tenersi aggiornato su notizie e attualità, e il 29.7% per cercare contenuti di intrattenimento. La piattaforma stessa conferma Reddit come uno dei forum online più frequentati a livello globale, riportando oltre 416 milioni di utenti attivi su base settimanale (Reddit 2025a). Chiaritone, dunque, il considerevole utilizzo odierno, le seguenti sotto-sezioni ne riassumono la struttura e i meccanismi di funzionamento, al fine di corroborare la scelta di Reddit come spazio digitale sperimentale nella presente ricerca.

2.2.1 Struttura, community e subreddits

In merito alla natura stessa della piattaforma, occorre precisare come Reddit venga spesso posto a metà strada tra un forum di discussione e una rete sociale. In primo luogo, uno degli aspetti che più lo contraddistingue dagli altri social media è il fatto che sia articolato in comunità tematiche o aree di interesse, elemento che richiama il tratto caratteristico dei forum. I dati della piattaforma aggiornati a giugno 2025 stimano oltre 100 mila comunità attive, comunemente definite subreddit. In secondo luogo, si parla di rete sociale poiché, da un lato, presenta sezioni commenti che seguono un'organizzazione reticolare e, dall'altro, si fonda su dinamiche di reputazione ed interazione tra utenti. La *Figura 7* offre una rappresentazione grafica della struttura della piattaforma:

[35] Analogamente alle *user identities* discusse in *§1.2.2*, si evidenzia come gli *unique visitors* siano da intendersi come identità distinte di accesso e non necessariamente come persone univoche, in quanto uno stesso individuo potrebbe accedere al sito utilizzando più dispositivi o motori di ricerca.

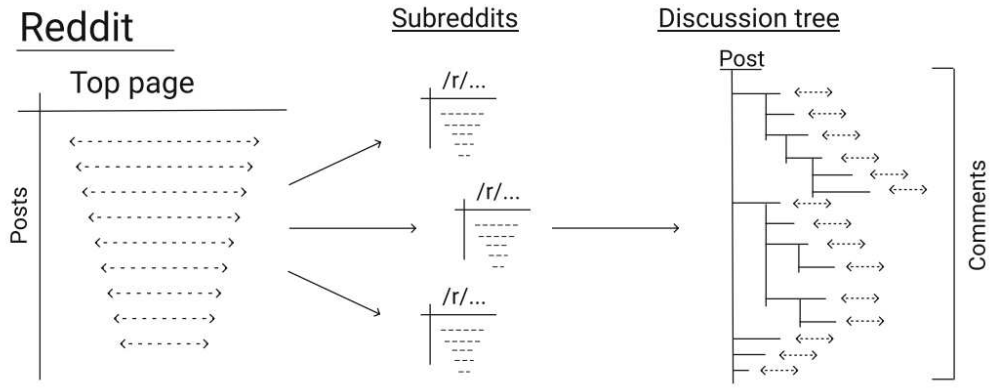

Figura 7 Rappresentazione grafica della struttura di Reddit.
(Tratto da Medvedev et al. 2019)

L'home page di Reddit (*top page*), il punto di accesso al sito, è costituita da una serie di post provenienti dai subreddit a cui l'utente è iscritto, ordinati in base ai voti ricevuti dai post e alla loro data di pubblicazione. Nel caso in cui l'utente non fosse autenticato o registrato alla piattaforma, i post visualizzati provengono da qualunque tipologia di subreddit. La piattaforma permette difatti agli utenti registrati (i redditors) di pubblicare post (le submission) – che possono includere contenuti di testo redatti dagli utenti stessi, link esterni, immagini e video – all'interno di uno o più subreddit. L'utente può successivamente accedere alla pagina principale di un singolo subreddit, dove il feed si restringe ai post appartenenti a quella specifica community.

Come anticipato, la visibilità dei contenuti è determinata dal numero di voti positivi e negativi ricevuti dalla submission. Una volta pubblicata una submission, gli altri utenti possono infatti assegnarle un *upvote* o un *downvote* e contribuire alla discussione tramite commenti, dando dunque origine alle *threaded online conversations*, le conversazioni nidificate introdotte nel capitolo precedente. I commenti sono difatti organizzati secondo una struttura gerarchica ad albero, definita dalla relazione di risposta (*reply-to*) ad altri commenti o al post originario. Medvedev et al. (2019, p. 185) la sintetizzano come segue: "[c]omments form a *discussion tree*, which can be described as a rooted tree, where the root is a designated node representing the post itself and each other node represents a comment. There is a link between two nodes if there is a 'reply-to' relation between

them"[36]. Analogamente alle submission, anche i commenti possono ricevere *upovotes* e *downvotes*, il cui bilancio concorre a determinare il cosiddetto *karma* dell'utente, un meccanismo di reputazione con una duplice funzione: da un lato, incoraggiare la produzione di contenuti di qualità e, dall'altro, prevenire lo spam all'interno della comunità.

2.2.2 Un Social network costruito su quattro *functional building blocks*

Alla luce delle caratteristiche appena illustrate, si chiarisce la struttura funzionale di Reddit alludendo brevemente alle tassonomie di SMTs discusse nel primo capitolo di questa ricerca. In una definizione formale di Reddit, identifichiamo la piattaforma come un Social network – e, più precisamente, un forum di discussione – costruito intorno ai blocchi funzionali (Kietzmann et al. 2011) di *conversations*, *sharing*, *reputation* e *groups*: la piattaforma è costruita primariamente con lo scopo di favorire lo scambio di opinioni tra gli utenti (*conversations*), i quali interagiscono tra di loro all'interno dei subreddits (*groups*) tramite la pubblicazione di submissions e la condivisione di contenuti testuali e/o multimediali (*sharing*), la cui visibilità è determinata dal sistema del punteggio karma (*reputation*). È tuttavia opportuno evidenziare che, secondo la classificazione proposta da Gundecha & Liu (2012), la piattaforma rientrerebbe – piuttosto che tra gli *online social networking sites* – nella categoria dei *social news*, all'interno della quale gli autori includono piattaforme incentrate sulla condivisione e selezione di notizie e articoli da parte di una comunità di utenti. Tale discrepanza è da ricondurre in particolare all'evoluzione d'uso di Reddit, il quale, presentandosi in origine come uno spazio digitale dedicato alla condivisione di notizie, nel tempo ha assunto caratteristiche più conversazionali, incentrate sul dibattito tra gli utenti all'interno dei subreddits e su contenuti di tipologie differenti.

Il riconoscimento di Reddit in quanto Social network risulta compatibile con la più recente definizione fornita da Koukaras et al. (2020), i quali – come precedentemente sottolineato – includono all'interno di questa categoria le piattaforme basate sulla condivisione e sull'interazione sociale, piuttosto che sull'intrattenimento e sull'auto-

[36] "[i] commenti danno origine a una struttura ad albero, descrivibile come un albero radicato, in cui la radice corrisponde a un nodo che rappresenta il post iniziale, mentre ciascun altro nodo rappresenta un commento. Due nodi risultano collegati quando tra essi esiste una relazione di risposta di tipo *reply-to*."

rappresentazione dell'utente (cfr. *Tabella 1*). Gli autori elencano le caratteristiche ufficiali delle principali piattaforme e individuano per Reddit le seguenti: *Social News Aggregation*, *Web Content Rating*, *Discussion Website*, *Content Sharing*, *Links*, *Text Posts*, *Images*, *Voting*. Ciascuna di esse viene successivamente associata a una utility semanticamente equivalente. Vengono così identificate le *common utilities* di Reddit, ossia: *sharing*, riconosciuta come utility primaria; *voting*, riconosciuta come utility secondaria; *connecting*, *multimedia* e *news*, riconosciute come utilities accessorie (*trivia utilities*), ossia presenti ma non centrali. Si comprende pertanto come Reddit venga annoverata nella categoria dei Social networks: pur mantenendo la peculiarità del *voting* come utility secondaria, la funzione principale dello *sharing* la avvicina al modello empirico di questa classe. L'appartenenza a questa categoria è ulteriormente confermata dalle utilities accessorie, che ne mettono in risalto la dimensione relazionale.

A conclusione di questa sezione, si evidenzia un ulteriore motivo per cui Reddit si presta perfettamente come piattaforma per l'analisi dell'opinione pubblica. Il nome "Reddit", derivato dall'inglese *Read It*, riflette la missione stessa della piattaforma, che nasce con l'obiettivo di creare un luogo in cui poter scoprire, leggere e discutere contenuti di interesse comune tra individui che possono appartenere anche a più comunità. In tal senso, il pubblico che si riunisce su Reddit rispecchia la visione promossa da Childs (1940), discussa all'inizio di questa tesi: contestando l'idea di pubblico come un gruppo omogeneo, organizzato e contraddistinto da attributi comuni, per Childs il pubblico è un insieme eterogeneo di individui che si aggregano temporaneamente attorno a questioni di interesse. In quanto piattaforma di discussione, Reddit incarna questa eterogeneità. La sua organizzazione è solo strutturale, non sociale: sebbene la piattaforma sia organizzata in subreddit tematici, gli utenti che partecipano alle discussioni entrano ed escono dalle stesse in base ai propri interessi, senza costituire gruppi stabili o coerenti. Difatti, i redditors appartengono contemporaneamente a più subreddit, differiscono tra di loro per provenienza e competenze, ed esprimono con libertà – e spesso anonimamente – opinioni individuali e frammentarie. Queste, se aggregate, delineano in senso lato un quadro dell'opinione pubblica, ma acquisiscono ancora più valore se si guarda a Reddit come esempio di quello spazio pubblico concepito da Dewey (1927) e Habermas (1962): su Reddit, le interazioni tra gli utenti alimentano il dibattito collettivo, che può essere interpretato come espressione dell'opinione pubblica.

2.3 L'analisi dei dati testuali (ADT)

L'analisi statistica dei dati testuali fa riferimento a una variante dell'analisi dei dati testuali (ADT) caratterizzata dall'impiego di metodi statistici. L'ADT è difatti un'"espressione-ombrello [sotto cui] si trova una molteplicità di tecniche e approcci che – al di là delle differenze più o meno marcate – hanno come missione quella di riuscire a dare conto dei contenuti di corpora testuali di dimensioni variabili" (Nobile 2024, p. 577). Si tratta di una metodologia di analisi impiegata in una molteplicità di discipline (come la psicologia, la sociologia, la sociolinguistica, la comunicazione, gli studi politici, etc.) – tante quanti sono i contesti di ricerca in cui è opportuno usare documenti scritti o trascritti come oggetto di studio[37]. Presupposto per il suo svolgimento è infatti la costruzione di uno o più corpora, per la quale si rimanda alla sotto-sezione successiva.

Nata come *content analysis* all'inizio del secolo scorso con Harold D. Lasswell (1927), che per primo ne coniò il termine in riferimento ai testi di propaganda politica, l'analisi dei dati testuali sottopone le parole al dominio dei numeri. Perseguendo l'obiettivo della "[…] classificazione logica dei contenuti e dei rispettivi valori semantici, attraverso l'individuazione delle unità di analisi (grammaticali, tematiche o formali) […]" (Giuliano & La Rocca 2008, p. 8), la *content analysis* si contrappone al punto di vista classico dell'ermeneutica, il quale pone invece l'accento sul "mondo di significati". L'ADT conosce un passaggio decisivo intorno agli anni Settanta, quando linguisti e matematici iniziano a studiare le corrispondenze in ambito fonetico e lessicale applicando strumenti di analisi quantitativa su corpora di testi sufficientemente ampi da essere considerati rappresentativi. In particolare, si riconoscono i contributi del matematico e statistico francese Jean-Paul Benzécri (1973), che sviluppa la tecnica dell'analisi delle corrispondenze (si veda *§2.4.2*), e di altri due studiosi transalpini come Lebart e Salem (1994), che introducono concetti fondamentali come quello di *word-type* e dell'analisi dei segmenti ripetuti. Successivamente, con lo sviluppo dell'informatica e delle *information technologies* (IT), l'analisi dei dati testuali si diffonde progressivamente fino a scomporsi in una molteplicità di approcci qualitativi, quantitativi e misti. Di fronte a queste suddivisioni, in virtù della natura qualitativa dei testi e del loro essere al contempo portatori di informazioni complesse, Tuzzi (2024, p. 19) afferma che "[…] l'analisi dei

[37] Purché, si sottintende, siano disponibili in formato digitale (vale a dire, sotto forma di file), adatto a essere conservato nella memoria di un computer e visualizzato attraverso un software.

dati testuali è *qualiquantitativa*, cioè sempre in bilico tra metodi quantitativi e metodi qualitativi e, proprio per questo, ha rappresentato storicamente un terreno di sperimentazione ideale per l'integrazione di diverse prospettive e il dialogo tra diverse discipline". Sebbene sia difficile ricondurre i molteplici approcci ad uno schema unitario, è possibile distinguere due vie di analisi che tentano, in misura diversa, un'integrazione tra i due poli metodologici sopracitati (il polo ermeneutico e il polo quantitativo), ciascuno con i relativi software di supporto: l'approccio "classico" dell'analisi dei dati qualitativi assistita dal computer e l'approccio "moderno" dell'analisi statistica dei dati testuali. La differenza tra le due pratiche di ricerca si riconosce principalmente nelle operazioni di codifica del testo – *ex ante* e semi-automatica per l'una ed *ex post* e automatica per l'altra.

2.3.1 L'approccio classico e l'approccio moderno: CAQDAS e ASDT a confronto

L'approccio classico è basato sulla strutturazione dei concetti e si distingue per essere un approccio a forte vocazione ermeneutica, in cui ogni fase dell'analisi è mediata da un lavoro di interpretazione e di elaborazione teorica. Questa prima famiglia di metodologie vede in particolare l'utilizzo dei CAQDAS (*Computer Aided Qualitative Data Analysis Software*), software per l'analisi qualitativa del contenuto che, offrendo strumenti di interrogazione, ricerca, organizzazione e annotazione dei materiali testuali o multimediali, si distinguono per le operazioni *ex ante* di codifica del testo. In sintesi, l'analisi effettiva è preceduta dalla definizione, da parte del ricercatore, di categorie concettuali che riescano a dar conto delle interrelazioni esistenti tra parole e concetti espressi all'interno del corpus complessivo (Nobile 2024). La codifica, assistita da strumenti informatici ma guidata dal ricercatore, chiarisce perché si parla analisi semi-automatica dei testi. Tuzzi (2009) riassume la logica ermeneutico-qualitativa alla base dei CAQDAS come segue:

> […] il «contenuto» di una porzione di testo viene riconosciuto «coerente» con una o più categorie concettuali ed «etichettato» (in maniera manuale o parzialmente manuale) dal ricercatore. Sebbene i software offrano strumenti di navigazione anche molto sofisticati, rimane una quota di intervento lasciata alla discrezionalità o, se si preferisce, alla sensibilità del ricercatore. Infatti, solo il ricercatore è in grado di riconoscere in un'espressione

> complessa la categoria concettuale più generale a cui fa riferimento o di riconoscere in due espressioni diverse la stessa categoria.
>
> (*ivi*, p. 6)

Il grado di discrezionalità del ricercatore si riduce invece nell'approccio moderno, ossia quello dell'analisi quantitativa basata sulla codifica automatica di unità testuali elementari. In tal caso, "[…] l'analisi non si basa su processi di astrazione di concetti e di etichettatura di porzioni di corpus (frasi, paragrafi, ecc.) ma [sul] riconoscimento e [sul] conteggio (automatico o semiautomatico) di unità elementari più semplici (parole, espressioni, lemmi, ecc.)" (*ibid.*). Difatti, se il ricorso all'approccio classico rende impossibile l'utilizzo di corpora di grandi dimensioni, in quanto richiede la lettura integrale del materiale oggetto di studio, al contrario "[l]'analisi statistica dei dati testuali muove [precisamente] dall'idea che il ricercatore o la ricercatrice non abbia letto o non abbia potuto leggere integralmente l'insieme di testi che sono oggetto di studio" (Tuzzi 2024, p. 17). I software relativi all'approccio moderno si adeguano a corpora di dimensioni imponenti, sono orientati alla lessicometria e si distinguono per la codifica *ex post* del testo: categorie e pattern lessicali emergono dall'adozione di strumenti statistici o, più in generale, quantitativi. Si comprende dunque come, a differenza dei CAQDAS, nei programmi per l'analisi statistica dei dati testuali l'unità di base non sia il significato, bensì la singola parola e le relative *co-occorrenze*: i software non trattano le parole in quanto tali, ma come fossero numeri, pittogrammi o segni. Si parla pertanto di l'*Analyse Statistique des Données Textuelles* (ASDT), ovvero di analisi statistica (o quantitativa) dei dati testuali (Lebart, Salem & Berry 1998), l'approccio che verrà adottato in questa ricerca. Il presente lavoro ha difatti previsto il download di migliaia di commenti Reddit, una quantità di dati testuali complessivamente troppo estesa per essere letta e gestita attraverso analisi di tipo qualitativo. Di fronte ad un corpus esteso, "[l]'analisi statistica dei dati testuali, pur presentando sia vizi che virtù, garantisce velocità e sistematicità alle operazioni di ricerca, spoglie e sintesi delle informazioni di interesse e permette di superare alcuni limiti dell'analisi qualitativa" (Tuzzi 2024, p. 22). Se, da un lato, l'analisi qualitativa – alla base dei CAQDAS – può risultare più interessante e approfondita di una quantitativa per via del lavoro intellettuale che richiede, dall'altro non è riproducibile ed è impraticabile su vasta scala per motivi di tempo e risorse.

In ultima istanza, è opportuno ribadire che anche l'approccio moderno prevede chiaramente un'attività di interpretazione, seppur mediata dall'applicazione di tecniche statistiche. Come afferma Nobile (2024, p. 579), "[…] l'analisi statistica dei dati testuali non si limita a produrre uno spoglio delle frequenze delle parole, ma mira soprattutto a 'spremere' il testo, in modo da ottenerne, attraverso un'operazione di interpretazione *ex post* sugli output, l'individuazione dei contenuti e dei nessi tematici": dopo aver supervisionato l'intero processo, il ricercatore attua a valle un'interpretazione dei risultati del software di analisi e ha infine il compito di comunicarli in altra forma. Questa fase costituisce l'ultimo dei tre sopracitati passaggi canonici: come qualsiasi analisi statistica dei dati, anche quella relativa ai dati testuali prevede la loro acquisizione, la loro sintesi in misure statistiche e la restituzione dei risultati ottenuti. Una corretta interpretazione dei risultati dipende certamente dalle elaborazioni statistiche, ma si fonda su un preliminare, consistente lavoro di acquisizione finalizzato a reperire e rendere i dati testuali adeguati agli obiettivi ricerca (Tuzzi 2024).

2.3.2 Il corpus: la collezione di testi alla base dell'analisi quantitativa

Nel condurre un'analisi statistica dei dati testuali, il primo passaggio da tenere in considerazione è la costruzione del corpus, termine con cui si fa riferimento a "[…] una collezione di testi in possesso di caratteristiche di coerenza, omogeneità, ampiezza e, in alcuni casi, esaustività che lo rendono adeguato agli scopi della ricerca" (Tuzzi 2024, p. 33). Sono esempi di corpora tutti i romanzi pubblicati da uno scrittore, tutte le risposte fornite ad una specifica domanda di un questionario, tutti gli articoli di fondo pubblicati da un giornale in un arco temporale definito. Si tratta in tal caso di corpora esaustivi per natura, che costituiscono cioè un insieme noto e finito di testi appartenenti ad una determinata classe. Laddove raccogliere tutti i testi riconducibili a una data classe non sia possibile, la prima scelta da attuare è relativa alla tipologia di testi da includere e alla loro dimensione. Nel presente lavoro, ad esempio, l'ampiezza di Reddit ostacola l'individuazione manuale e certa della totalità delle submission inerenti all'intelligenza artificiale, e richiede la definizione di specifici criteri di selezione (cfr. *§3.1*).

Come mostrano gli esempi sopracitati, in linea generale è possibile affermare che un corpus è dato da una collezione di testi accomunati da una o più caratteristiche (Ondelli 2018). Per la costituzione di un corpus si considerano generalmente le proprietà dei testi

(i *metadati* associati ai testi) – e.g., lingua, genere testuale, fonte e/o autore, argomento, stile, dimensione – e, più raramente, le cinque dimensioni di variazione della lingua (diacronica, diatopica, diafasica, diastratica, diamesica) definite da Berruto (1987), le quali spesso inglobano del tutto o in parte le proprietà testuali. Se alcune proprietà o variazioni possono fungere da criteri di selezione nella fase di costruzione del corpus, altre possono invece costituire oggetto di studio della ricerca. Tuzzi (2024) riassume:

> […] una o più proprietà potrebbero essere oggetto dello studio (per esempio si vogliono trovare somiglianze e differenze tra gruppi di testi con proprietà diverse); una o più proprietà potrebbero essere non osservabili in maniera diretta (strutture latenti) e lo scopo della ricerca potrebbe essere proprio quello di identificarle (per esempio come risultato di una classificazione automatica).
>
> (*ivi*, p. 36)

Quando la domanda di ricerca richiede una comparazione interna al corpus in virtù di una specifica proprietà o varietà, il corpus complessivo viene ripartito in subcorpora, termine che indica sottoinsiemi di testi di un corpus che condividono una o più proprietà. La suddivisione del corpus in subcorpora svolge difatti un ruolo fondamentale nel consolidare la dimensione comparativa di un'analisi statistica dei dati testuali. Tuttavia, questo comporta un'attenzione particolare al problema del bilanciamento: affinché i diversi subcorpora possano essere oggetto di comparazione, è necessario che siano coerenti tra loro e che non sia presente alcun fenomeno di sovra-rappresentazione. Ciascuno dei sottoinsiemi deve dunque essere adeguatamente bilanciato con gli altri in termini di numero e ampiezza dei testi.

In merito al ruolo attribuito alle cinque dimensioni di variazione nella fase di compilazione del corpus, è importante enfatizzare come Ondelli (2018) ponga l'accento sulla natura linguistica dei dati testuali e sulla necessità di considerarne i fattori (socio)linguistici, a fronte del rischio di ridurre i testi a meri dati numerici. D'altra parte, Barbera et al. (2007) offrono un contributo complementare a quello di Ondelli (2018), fornendo una definizione più operativa di corpus che evidenzia l'importanza di una corretta elaborazione al fine di renderlo trattabile e analizzabile tramite software:

> [Un corpus è una] [r]accolta di testi (scritti, orali o multimediali) o parti di essi in numero finito in formato elettronico trattati in modo uniforme (ossia tokenizzati ed addizionati di markup adeguato) così da essere gestibili ed interrogabili informaticamente […]
>
> (Barbera et al. 2007, p. 70)

Dopo aver preso le dovute decisioni in merito alla selezione dei testi da includere nella raccolta, i dati testuali devono difatti essere preparati per l'elaborazione automatica del testo, una fase indispensabile quando oggetto di studio è l'uso del linguaggio e necessaria indipendentemente dallo scopo finale della ricerca – sia essa linguistica, un'analisi del contenuto o un'attività di *information retrieval* (IR). Questo procedimento prende il nome di pre-elaborazione o pre-trattamento (*pre-processing*) ed include procedure specifiche, come la tokenizzazione (*tokenisation*) e il *mark-up*, le due fasi iniziali di costruzione di un corpus secondo Barbera et al. (2007).

La tokenizzazione è una procedura volta all'identificazione delle unità testuali minime del corpus, i cosiddetti *tokens*, allo scopo di mantenere in analisi solo le parole di cui sono composti i testi ed eliminare tutto ciò che non costituisce materiale testuale analizzabile. Nello specifico, la tokenizzazione consiste nell'"istruire" il software a isolare correttamente le parole, ossia a riconoscere quali caratteri dovrebbero essere considerati lettere e quali dovrebbero essere considerati separatori (Ondelli 2018), come nel caso di numeri, apostrofi, trattini o altri segni di interpunzione. Con il termine *mark-up* si fa invece riferimento alle informazioni soprasegmentali, come il titolo, l'autore o i numeri di pagina di un testo, rispetto alla pura successione lineare dei caratteri del testo (Barbera et al. 2007). Queste possono essere intese come "informazioni aggiunte" che il software può escludere dal calcolo dei *token* – un'attività non necessaria per il corpus analizzato in questa ricerca (che non presenta alcuna informazione soprasegmentale) e che per questo non verrà approfondita. Un'ulteriore procedura di preparazione del testo all'elaborazione è la normalizzazione, che, diversamente dalla tokenizzazione, "[…] ha come obiettivo identificare le parole diverse e contarle correttamente (o, più precisamente, contare i *word type* in maniera sistematica e coerente con una serie di criteri stabiliti a priori)" (Tuzzi 2024, p. 70). L'azione di normalizzazione più diffusa consiste nel trasformare in minuscolo tutte le lettere maiuscole presenti in un testo. Va nondimeno osservato che, se condurre tali operazioni manualmente richiede un grande dispendio di tempo e risorse, affidarle ad un software implica un tasso di errore significativo. Ad esempio, eliminare i numeri e trattare qualunque segno non alfabetico come separatore può creare parole non esistenti, confondere l'attribuzione di genere o produrre *word-types* aggiuntivi (cfr. Tuzzi 2024, Ondelli 2018). Ne consegue che, affinché venga garantita la riproducibilità di tutte le operazioni, è necessario avere piena consapevolezza delle

decisioni assunte nella fase di *pre-processing* e chiarire le regole adottate. La tipologia di tokenizzazione o di normalizzazione adottata dipende strettamente dagli obiettivi di ricerca, al pari di qualunque altra scelta propria di questa fase preliminare. Diverse procedure possono, ad esempio, essere applicate a seconda che si scelga di operare su *word-tokens*, lemmi o *stem*– una decisione che incide sulle misure sensibili al numero di *tokens* presenti in un corpus, come la dimensione complessiva e la ricchezza lessicale (Brezina 2018; Cohen et al. 2019). Difatti, dopo essere stati identificati e contati tramite tokenizzazione, gli elementi testuali del corpus possono essere sottoposti a lemmatizzazione[38] o *stemming*[39], ulteriori tecniche appartenenti al *pre-processing* dei testi che permettono di ridurre il rumore provocato dalle varianti morfologiche di una stessa parola.

Come anticipato, l'ASDT ha senso in ragione della numerosità e della dimensione dei testi oggetti di studio. Nel contesto dell'analisi statistica, è doveroso precisare che la dimensione di un testo non è solo relativa al numero di parole che contiene, ma anche al suo grado di *ridondanza*, che ne determina la ricchezza lessicale: "[…] da un lato, un corpus che contiene molte parole diverse è lessicalmente molto variato e, quindi, potenzialmente interessante dal punto di vista qualitativo; dall'altro, la statistica ha bisogno di contare le parole e, quindi, per poterne valutare presenza, assenza e occorrenza nei testi, nei subcorpora e nelle porzioni, ha bisogno che le stesse parole occorrano, cioè si ripetano numerose volte" (Tuzzi 2024, p. 52). Per adottare un approccio di tipo quantitativo ed implementare metodi statistici, è pertanto necessario valutare la trattabilità del corpus oggetto di studio, vale a dire calcolarne la dimensione e la ricchezza lessicale. Tale valutazione avviene per mezzo di alcune misure empiriche, come il Type-Token Ratio (TTR) e la percentuale di hapax (hapax%), che possono essere comprese solo chiarendo la distinzione tra due concetti fondamentali nel conteggio delle parole: *word-tokens* e *word-types*, i cui valori forniscono una prima semplice valutazione della dimensione di un corpus e del suo vocabolario. Il numero *N* di *word-tokens* indica la

[38] Tutte le varianti morfologiche vengono ridotte alla loro forma base: "[l]a lemmatizzazione è l'operazione che permette di associare a ogni parola (forma grafica) il corrispondente lemma, di solito corredato almeno dalla categoria grammaticale di appartenenza" (Tuzzi 2024, p. 102). Il lemma è dunque la forma canonica della parola, così come presentata nei comuni dizionari della lingua (verbi all'infinito, sostantivi al singolare, aggettivi al maschile singolare).

[39] Tutte le varianti morfologiche vengono ridotte alla loro radice comune, o stelo (*stem*), con lo scopo di "[…] ottenere la stessa rappresentazione per diverse parole in modo da coglierne un contenuto semantico condiviso" (Tuzzi 2024, p. 107).

quantità complessiva di unità testuali all'interno di un corpus, ossia il numero di parole o *occorrenze* totali, e rappresenta la dimensione del corpus; il numero *V* di *word-types*, invece, indica il numero di *parole diverse* o forme uniche presenti in un corpus, senza considerarne le ripetizioni, e corrisponde alla dimensione del vocabolario del corpus. È dunque possibile guardare ai concetti di *word-tokens* e *word-types* come due diversi modi di contare le parole: ad esempio, nella frase *occhio per occhio e dente per dente*, si contano 7 *word-tokens* (la totalità delle occorrenze di ciascun *word-type*) e 4 *word-types* (la totalità di parole diverse, a prescindere dalle loro occorrenze). Il rapporto tra *word-types* e *word-tokens* (*V/N*) consente di calcolare il TTR, "[…] una misura della 'ricchezza lessicale' del corpus in termini di varietà del lessico, cioè di numero di parole diverse presenti commisurato al numero di parole totali" (Tuzzi 2024, p. 61): in linea generale, valori elevati di TTR indicano una maggiore variazione lessicale; al contrario, valori bassi suggeriscono una minore variazione lessicale (Ali & Hussein 2014). Il TTR è difatti molto sensibile alla lunghezza del testo, motivo per cui, nel confrontare il TTR di testi diversi, è necessario che presentino la stessa lunghezza: quando la lunghezza di un testo aumenta, cresce il numero di forme grafiche già incontrate – *i.e.*, maggiori saranno le ripetizioni – e, conseguentemente, diminuisce il suo valore di TTR (Ali & Hussein 2014; Brezina 2018). Il suo reciproco, ovvero il rapporto tra *word-tokens* e *word-types* (*N/V*), misura invece la frequenza media dei *word-types* ed è dunque un indicatore indiretto del grado di ridondanza del testo. I *word-types* che all'interno di un corpus presentano una sola occorrenza, cioè si presentano una sola volta, prendono il nome di *hapax legomena* (nell'esempio precedente, la congiunzione *e* è l'unico *hapax legomenon*). Gli *hapax* (V_1) costituiscono sempre la classe di frequenza[40] più numerosa all'interno di un corpus, occupando circa metà del suo vocabolario. Il rapporto tra il numero di *hapax* e il numero di *word-types*, poi espresso in percentuale, restituisce l'indice di hapax%, un'ulteriore misura della varietà lessicale del corpus che quantifica la proporzione di parole che occorrono solo una volta rispetto al vocabolario complessivo.

Lavorando nell'ambito di un *bag of words* (BOW), in cui *word lists* (liste di *word-types*) e frequenze costituiscono le fondamenta dello studio, un corpus può essere

[40] Considerata una lista di frequenza (*frequency list*), contenente la lista di *word-types* corredati dal numero di occorrenze e ordinati per frequenza decrescente, i *word-types* possono essere raggruppati in classi di frequenza. Una classe di frequenza è un insieme di parole che si presentano con lo stesso numero di occorrenze. Ad esempio, la classe di frequenza V_1 indica l'insieme di *word-types* che si presentano con un numero di occorrenza pari a 1, ovvero la classe degli *hapax legomena*.

considerato sufficientemente esteso per procedere con analisi quantitative solo se presenta un adeguato livello di ridondanza. In altre parole, affinché possa essere oggetto di un'analisi statistica è necessario che rispetti le soglie di TTR e hapax% comunemente adottate in letteratura: la soglia massima di TTR che un corpus deve possedere è pari a 0,20, mentre la soglia massima di hapax% è pari a 0,50. Oltre queste due soglie empiriche, il numero di *word-types* (nel caso del TTR) o di *hapax* (nel caso dell'hapax%) è troppo elevato e non consente di ottenere il grado di ridondanza auspicabile: minori sono i valori di TTR e hapax%, minore sarà la varietà lessicale e pertanto maggiore sarà la ridondanza all'interno del testo, elemento fondamentale per perseguire uno studio statistico basato sulla frequenza delle parole.

2.4 Metodi e strumenti dell'ASDT

Le seguenti sezioni forniranno un'introduzione teorica ai metodi statistici impiegati, nella presente ricerca, per l'elaborazione dei dati testuali. Per una più chiara comprensione della loro applicazione, si rimanda nuovamente alla *§3.1* dedicata alla descrizione del corpus analizzato.

Al fine di comprendere come si articola il discorso della sfera pubblica digitale sull'intelligenza artificiale, lo studio ricorre a due metodi appartenenti al ramo dei metodi *unsupervised*: il *text clustering* e l'analisi delle corrispondenze (o *Correspondence Analysis*). Le scienze statistiche distinguono difatti tra metodi *unsupervised* e *supervised*, una dicotomia spesso richiamata anche attraverso le giustapposizioni esplorativi *vs.* confermativi e descrittivi *vs.* inferenziali. Tuzzi (2024) la riassume come segue:

> In tutti questi casi l'idea generale è che esistano, da un lato, metodi (*unsupervised*) che descrivono ed esplorano le informazioni a disposizione senza ipotesi sulle strutture cercate e, di conseguenza, preferiscono strumenti che hanno come principale obiettivo quello di rappresentare i dati e far emergere strutture latenti, cioè sistemi di relazioni visibili solo attraverso l'analisi statistica; dall'altra, metodi (*supervised*) che, invece, partono da un insieme di conoscenze a priori […] e preferiscono strumenti in grado di inferire in che misura il modello ipotizzato è confermato dai dati […].
> (*ivi*, p. 22)

All'interno della prospettiva esplorativa *unsupervised*, il *text clustering* e l'analisi delle corrispondenze consentono di lavorare in ottica comparativa, in quanto orientati a un

confronto tra testi o gruppi di testi mirato a far emergere somiglianze e differenze tra gli stessi. Si tratta di due metodi statistici che, in altri termini, consentono di visualizzare le strutture latenti presenti nei testi analizzati, inserendosi all'interno dell'approccio del *distant reading*, una metodologia appartenente all'ambito delle Digital Humanities e che in larga parte riassume la logica dell'ASDT. Nella sua accezione moderna, il termine, coniato da Franco Moretti (2005), si riferisce difatti alla pratica di analizzare grandi corpora di testi con la mediazione di un'analisi quantitativa, la quale consente di identificare tendenze e strutture che con il tradizionale *close reading* non sarebbero altrimenti evidenti (Jänicke et al. 2025; Ciotti 2021; Tuzzi 2024).

2.4.1 *Text Clustering*: il *clustering* stilometrico e la *Descending Hierarchical Classification*

Nello studio di un corpus, uno degli obiettivi di ricerca può essere l'organizzazione in gruppi omogenei dei testi che lo compongono. Tale obiettivo viene perseguito per mezzo del *text clustering*. Il termine designa un compito della *cluster analysis*, finalizzato a raggruppare documenti simili in *cluster* coerenti e a separare documenti dissimili in *cluster* distinti, con l'obiettivo di svelare le strutture di relazione sottese tra testi e tra gruppi di testi: considerato un corpus, i suoi testi vengono assegnati a un numero limitato di gruppi, in ciascuno dei quali l'insieme di testi condivide caratteristiche di interesse per la ricerca.

Nella prospettiva esplorativa il *text clustering* assume la denominazione di *unsupervised text clustering* (raggruppamento dei testi non supervisionato), in opposizione alla *supervised text classification* (classificazione dei testi supervisionata) della prospettiva confermativa. L'*unsupervised text clustering* e la *supervised text classification* sono difatti le due grandi famiglie in cui la letteratura distingue i metodi di raggruppamento e classificazione: esse condividono lo stesso obiettivo generale di riunire i testi in classi sulla base della presenza di tratti comuni, ma differiscono nelle modalità e nell'arbitrarietà di attribuzione di un testo ad una certa classe. Nel primo caso, i testi del corpus non sono ascritti a gruppi già noti: è sfruttando unicamente i dati testuali disponibili che agli algoritmi viene affidato il compito di raggruppare i testi simili in gruppi coerenti. Nel secondo caso, i testi del corpus, o almeno una parte, hanno invece un'attribuzione certa a una classe nota a priori: tale insieme di testi costituisce il *training*

set degli algoritmi, a partire dal quale questi hanno "[…] il compito di imparare (*machine learning*) come discriminare le classi per poi inferire l'appartenenza alle classi di altri testi (che possono essere i testi dello stesso corpus privi di un'attribuzione a priori, oppure nuovi testi esterni al corpus)" (Tuzzi 2024, p. 206).

Se la *text classification* risulta particolarmente utile negli studi indirizzati all'esaminazione di testi nuovi, come nel caso degli studi di *authorship attribution*, il *text clustering* si presta invece a quelli orientati a far emergere strutture di relazione potenzialmente non visibili senza il supporto del *distant reading*. Ne consegue che, considerata la natura esplorativa di questa ricerca, volta ad esaminare il discorso pubblico sull'AI ed i temi ricorrenti, si è scelto di lavorare nell'ottica dei metodi *unsupervised* e di ricorrere al *text clustering*. Nel caso in esame non sono infatti disponibili informazioni a priori sui testi: come sarà specificato nelle sezioni successive (cfr. *§3.1*,) il raggruppamento delle submission nei quattro subcorpora di cui si compone il corpus è il risultato di una pura valutazione qualitativa. Pertanto, oltre ad identificare le strutture di relazione presenti tra questi, in un'analisi complessiva del corpus il *text clustering* potrebbe rivelare strutture di relazione e aree tematiche aggiuntive rispetto a quelle già identificate, avvalorando il suo ruolo nel far emergere pattern latenti. Nella sua seconda applicazione, esso verrà inoltre impiegato per un'esplorazione più approfondita dei singoli subcorpora e, pertanto, per mettere a confronto le strutture di somiglianza o differenza emergenti tra i commenti delle quattro aree tematiche.

I risultati raggiunti con la clusterizzazione dipendono fortemente dalle scelte che il ricercatore opera in merito alla selezione del vocabolario da analizzare (*quanti* e *quali* elementi testuali prendere in considerazione), alla misura di similarità da utilizzare e all'algoritmo di classificazione da adottare per la creazione dei clusters – decisioni che, come in ogni fase dell'analisi, devono essere coerenti tanto con gli obiettivi prefissati quanto con la natura dei testi oggetto di studio. Al fine di conseguire i suddetti obiettivi, nel presente lavoro si è scelto di applicare due diverse procedure di *text clustering*. Le procedure di raggruppamento più comuni valutano la somiglianza lessicale tra testi tramite metodi *distance-based*, vale a dire metodi che assegnano i testi a gruppi diversi sulla base del calcolo di una distanza. Coerentemente con questa tendenza, verrà dapprima condotto un *clustering* stilometrico sull'interfaccia RStudio (Posit team 2023) tramite *stylo* (Eder et al. 2016), un pacchetto R per l'analisi dello stile di scrittura che può

utilizzare una pluralità di misure di distanza tra testi. Nell'analisi stilometrica[41], la pratica più diffusa, introdotta da Mosteller & Wallace (2007 [1964]; cfr. Lahjouji-Seppälä et a. 2022), consiste nell'utilizzare come elemento testuale (*feature*) alla base dell'analisi le parole più frequenti: a partire dalle *n* MFW (*Most Frequent Words*), il pacchetto *stylo* consente di organizzare in gruppi i subcorpora e/o i testi che costituiscono il corpus, rendendone evidenti i pattern latenti. Successivamente, il corpus sarà sottoposto a un *clustering* tematico-contenutistico (o, più semplicemente, semantico) tramite IRaMuTeQ (2025 [Retinaud 2009])[42], un software per l'analisi dei dati testuali che riproduce il metodo di classificazione descritto da Reinert (1990), la *Descending Hierarchical Classification*[43] (DHC), anche definita *method of Descending Hierarchical Analysis* (Camargo & Justo 2021): il testo viene tipicamente suddiviso in segmenti definiti, poi classificati sulla base delle frequenze delle parole (in forma ridotta, *i.e.* lemmi) che contengono. L'incrocio delle forme ridotte con i segmenti di testo (TS) produce un numero di classi (*cluster*) a cui è associato un vocabolario caratteristico. In altri termini, ogni classe raggruppa un insieme di parole frequentemente utilizzate nella stessa frase o gruppo di frasi (*i.e.*, segmenti di testo). La forza dell'associazione tra il vocabolario e le classi, definibili pertanto come classi semantiche o di contenuto, è determinata dal calcolo del contributo del chi-quadrato (Arditi et al. 2020), necessario a verificare se una parola sia caratteristica di una determinata classe. L'efficacia del metodo Reinert nell'individuare le principali strutture semantiche di un corpus è sintetizzata da Montalescot et al. (2024) come segue:

> By identifying the lexical structure of a corpus, this method highlights how words and expressions are organized and interconnected in a corpus of text. It thus provides valuable insights into the content and meaning of the text. Statistical analyses are performed on the words and units of context (UCs) of the corpus[, …] segments of text that are identified based on the language structure, particularly punctuation and length. These units serve as the building

[41] La stilometria è un ambito di ricerca che studia lo stile linguistico o di scrittura di un testo attraverso l'applicazione di metodi statistici e quantitativi (Holmes 1998). Essa trova sua principale applicazione negli studi di attribuzione d'autore (Eder 2016) e può essere condotta su diverse tipologie di elementi linguistici, che possono comprendere lettere, categorie grammaticali come le parti del discorso (*parts of speech*), e persino la punteggiatura (Lahjouji-Seppälä et al. 2022).

[42] Sviluppato da Pierre Retinaud (2009), IRaMuTeQ (*Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires*) è un programma gratuito, basato sul software R, che consente diverse tipologie di analisi dei testi, come la *Descending Hierarchical Classification* e la *Similarity Analysis*, offrendo contributi rilevanti nel campo delle scienze umane e sociali (Camargo & Justo 2013).

[43] Classificazione Gerarchica Discendente (CGD).

blocks for the analysis, allowing the method to examine the co-occurrences and distribution of lexical forms within the identified UCs.
(*ivi*, p. 2)[44]

Ne consegue che parole appartenenti alla stessa classe semantica indicano come i segmenti in cui esse co-occorrono condividano un vocabolario simile. Mentre la prima applicazione fornisce quindi una panoramica generale sui rapporti di somiglianza che intercorrono tra i subcorpora, permettendo di confrontare gli stili di scrittura dei commenti trattanti aree tematiche differenti, la seconda applicazione consente di scendere nel dettaglio e di estrarne i contenuti ricorrenti. L'idea alla base del *text clustering* condotto su IRaMuTeQ è che le classi individuate rappresentino difatti gli argomenti e i temi che attraversano il corpus e/o i subcorpora.

Si è precisato come nell'ambito dell'analisi delle corrispondenze le relazioni di similarità e dissimilarità si traducano graficamente in una distribuzione di punti su un piano cartesiano. Nell'ambito del *text clustering*, queste si riassumono invece in una struttura gerarchica a più livelli che prende il nome di dendogramma: in entrambe le applicazioni, l'una stilometrica l'altra tematico-contenutistica, il raggruppamento dei testi o delle classi assume la forma di "[…] un grafico ad albero rovesciato nel quale le foglie rappresentano i testi classificati [o le classi] e i rami le strutture create dalla procedura di raggruppamento" (Tuzzi 2024, p. 210). Il dendogramma è, in sintesi, un grafico in cui testi o classi (foglie) sono nidificati uno nell'altro attraverso relazioni di appartenenza (rami) a sottoinsiemi o sovrainsiemi: l'appartenenza ad uno stesso ramo è dunque indice della loro somiglianza.

La differenza tra le due applicazioni, oltre che nelle porzioni di testo in analisi, risiede negli algoritmi di classificazione gerarchica utilizzati, i quali possono distinguersi in agglomerativi e divisivi. Una volta ottenuta la matrice distanze, contenente una misura riferita a qualsiasi coppia di testi del corpus, il *clustering* stilometrico condotto su RStudio impiega algoritmi agglomerativi per unire i singoli testi prima a coppie e poi a gruppi in

[44] Identificando la struttura lessicale di un corpus, questo metodo mette in evidenza il modo in cui, all'interno di un insieme di testi, parole ed espressioni sono organizzate e connesse tra loro. Esso, dunque, fornisce indicazioni utili per l'interpretazione del contenuto e del significato del testo. Le analisi statistiche vengono condotte sulle parole e sulle unità di contesto (UC) del corpus, vale a dire segmenti testuali individuati sulla base della struttura linguistica, in particolare della punteggiatura e della lunghezza. Tali unità fungono da elementi costitutivi dell'analisi, consentendo al metodo di esaminare le co-occorrenze e la distribuzione delle forme lessicali all'interno delle UC identificate.

base al loro grado di somiglianza. Il *clustering* semantico condotto su IRaMuTeQ segue invece la procedura inversa: il software utilizza difatti algoritmi divisi che, a partire dal corpus complessivo, procedono per suddivisioni successive in gruppi sempre più piccoli e sempre più omogenei.

2.4.2 Analisi delle corrispondenze

L'analisi delle corrispondenze (o *Correspondence Analysis*), una rinomata tipologia di analisi nell'ambito dell'ASDT, è un metodo statistico per la mappatura dei contenuti che consente di "[…] ricondurre la complessità dei dati testuali di un corpus a una rappresentazione grafica di semplice lettura" (Tuzzi 2024, p. 189). Analogamente al *text clustering*, anch'essa si inserisce tra i metodi esplorativi *unsupervised*, in quanto permette di esplorare e visualizzare, oltre che i temi più rilevanti, il complesso sistema di relazioni sottese tra i testi, tra le parole, e tra i testi e le parole. Lavorando in ottica comparativa, il ricorso ad un'analisi delle corrispondenze è finalizzato a far emergere le somiglianze e differenze tra i dati testuali attraverso una rappresentazione grafica su piano cartesiano. Nel caso in esame, l'analisi delle corrispondenze produce un piano cartesiano sul quale la distribuzione dei subcorpora tematici e/o delle parole utilizzate dai redditors consente di visualizzare la misura in cui i diversi argomenti relativi all'AI vengono lessicalmente trattati in maniera simile o divergente.

Nella sua versione più semplice, il punto di partenza dell'analisi è una tabella di contingenza $V \times M$. Si tratta di una matrice "parole × testi" – anche definita *Term-Document Matrix* (TDM) – che incrocia il vocabolario V del corpus, distribuito sulle righe, con i testi o subcorpora M che lo compongono, distribuiti sulle colonne: ogni riga rappresenta pertanto una parola w, mentre ogni colonna rappresenta un testo o un subcorpus t (si veda *Figura 8*).

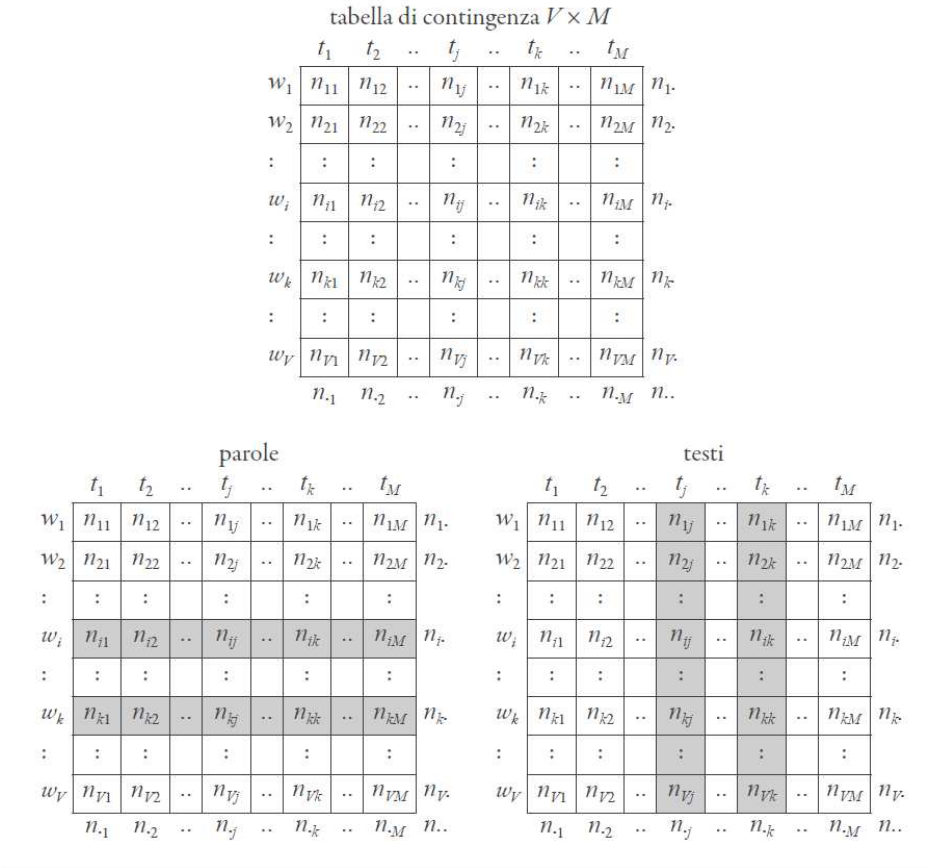

Figura 8 Tabella di contingenza *V x M* alla base dell'analisi delle corrispondenze. (Tratto da *Fondamenti di analisi dei dati testuali*, Tuzzi 2024)

Ogni cella contiene invece il numero di occorrenze *n* di una specifica parola *w* in uno specifico testo o subcorpus *t*[45]. Tali occorrenze vengono successivamente trasformate in frequenze relative per ottenere i profili lessicali[46] che determineranno la distribuzione delle parole e/o dei testi nella loro rappresentazione grafica. A partire dalle occorrenze della tabella di contingenza, lo scopo dell'analisi delle corrispondenze è difatti quello di trasformare le analogie e le differenze esistenti tra i profili lessicali dei testi, oltre che delle parole che li caratterizzano, in coordinate proiettate come punti in un sistema

[45] "[…] la generica cella n_{ij} contiene il numero di occorrenze della *i*-esima parola (parola che si trova alla riga *i*) nel *j*-esimo testo (testo che si trova sulla colonna *j*). Quando la matrice è completa, cioè riporta tutte le parole e tutti i testi, la somma $n_{i\cdot}$ di una riga *i* rappresenta il totale delle occorrenze nel corpus della *i*-esima parola e la somma $n_{\cdot j}$ di una colonna *j* rappresenta la dimensione in *word token* del testo *j*-esimo" (Tuzzi 2024, p. 85).

[46] Il profilo lessicale di una parola *w* è la distribuzione delle sue frequenze relative su tutti i testi o subcorpora, mentre il profilo di un testo o subcorpus *t* è la distribuzione delle frequenze relative di tutte le parole presenti in esso.

multidimensionale di assi cartesiani. Sul piano cartesiano, ogni punto rappresenta la posizione assunta dalle parole e/o dai testi, e la prossimità (o la distanza) che intercorre tra gli stessi riflette il grado di similarità (o di dissimilarità) dei loro profili lessicali. Tale grado di similarità viene misurato attraverso il chi-quadrato (χ^2, di seguito anche indicato come chi2), una distanza calcolata sui profili lessicali in esame. Essa può essere determinata per tutte le coppie di parole e per tutte le coppie di testi, e sintetizzata nelle rispettive matrici di distanze $V \times V$ e $M \times M$ (Tuzzi 2024), illustrate nella figura sottostante.

distanze tra parole $V \times V$

	w_1	w_2	..	w_i	..	w_j	..	w_k	..	w_V
w_1	0	d_{12}	..	d_{1i}	..	d_{1j}	..	d_{1k}	..	d_{1V}
w_2	d_{21}	0	..	d_{2i}	..	d_{2j}	..	d_{2k}	..	d_{2V}
:	:	:		:		:		:		:
w_i	d_{i1}	d_{i2}	..	0	..	d_{ij}	..	d_{ik}	..	d_{iV}
:	:	:		:		:		:		:
w_j	d_{j1}	d_{j2}	..	d_{ji}	..	0	..	d_{ik}	..	d_{iV}
:	:	:		:		:		:		:
w_k	d_{k1}	d_{k2}	..	d_{ki}	..	d_{kj}	..	0	..	d_{kV}
:	:	:		:		:		:		:
w_V	d_{V1}	d_{V2}	..	d_{Vi}	..	d_{Vj}	..	d_{Vk}	..	0

distanze tra testi $M \times M$

	t_1	t_2	..	t_i	..	t_j	..	t_k	..	t_M
t_1	0	d_{12}	..	d_{1i}	..	d_{1j}	..	d_{1k}	..	d_{1M}
t_2	d_{21}	0	..	d_{2i}	..	d_{2j}	..	d_{2k}	..	d_{2M}
:	:	:		:		:		:		:
t_i	d_{i1}	d_{i2}	..	0	..	d_{ij}	..	d_{ik}	..	d_{iM}
:	:	:		:		:		:		:
t_j	d_{j1}	d_{j2}	..	d_{ji}	..	0	..	d_{jk}	..	d_{jM}
:	:	:		:		:		:		:
t_k	d_{k1}	d_{k2}	..	d_{ki}	..	d_{kj}	..	0	..	d_{kM}
:	:	:		:		:		:		:
t_M	d_{M1}	d_{M2}	..	d_{Mi}	..	d_{Mj}	..	d_{Mk}	..	0

Figura 9 Matrici di distanza chi-quadrato: a sinistra, la matrice delle distanze tra parole; a destra, la matrice delle distanze tra testi (o segmenti di testo).
(Tratto da *Fondamenti di analisi dei dati testuali*, Tuzzi 2024)

Nella fase finale dell'analisi, la distanza chi-quadrato viene tradotta graficamente: sul piano cartesiano, la distanza sarà minore tra termini e/o testi che presentano profili lessicali (e quindi frequenze relative) simili, mentre sarà maggiore tra termini e/o testi che presentano profili lessicali (e quindi frequenze relative) molto divergenti. Ad esempio, due termini con un profilo lessicale identico, e che dunque presentano il massimo grado di similarità, si troveranno a distanza 0 l'uno dall'altro e saranno rappresentati sul grafico nella forma di due punti sovrapposti.

Poiché l'insieme delle distanze calcolate per ciascuna coppia dà origine a nuove dimensioni ortogonali, la naturale conseguenza di tali operazioni è uno spazio multidimensionale molto complesso. Per far fronte a tale complessità e favorire l'interpretazione dei dati, l'analisi delle corrispondenze riduce lo spazio cartesiano su cui

vengono proiettati parole e/o testi ai primi due assi, vale a dire alle prime due dimensioni, poiché le più informative: "[…] il piano cartesiano costruito con i primi due assi è lo spazio bidimensionale che raccoglie la quota di inerzia più elevata e, quindi, meglio rappresenta in due dimensioni la struttura di associazione espressa dalla tabella di contingenza" (Tuzzi 2024, p. 197). Grazie alla più semplice rappresentazione a due dimensioni, gli assi cartesiani possono essere osservati uno alla volta o a coppie. Nonostante tale rappresentazione bidimensionale non restituisca tutta la complessità presente nella tabella di contingenza di origine, essa si rende necessaria per due motivi principali: da un lato, permette di riassumere e far emergere le relazioni più importanti, dall'altro, favorisce la leggibilità dei risultati.

La mappatura dei contenuti risultante dall'analisi è dunque un piano cartesiano, generalmente ridotto ai primi due assi, in cui parole e testi sono posizionati e distribuiti sui quattro quadranti in modo tale che le categorie lessicalmente simili si trovino nella stessa parte del grafico. Tuttavia, nell'interpretare il grafico risultante dalla CA, è necessario precisare che la posizione assunta da ciascun punto acquisisce significato solo all'interno del contesto globale del grafico: essa può essere valutata solo in relazione a quella assunta da tutti gli altri punti rispetto all'origine degli assi. Insieme, un alto valore del chi2 ed una maggiore distanza della parola dall'origine degli assi definiscono la sua significatività e rappresentatività all'interno del testo a cui è associata e la sua forza discriminante nello spazio bidimensionale.

3. Il corpus oggetto di analisi

3.1 Criteri di costruzione e subcorpora tematici

Come evidenziato più volte nel corso delle sezioni precedenti, la collezione di testi oggetto di analisi deve essere adeguata agli scopi della ricerca. Il corpus deve cioè rispondere alle esigenze di una specifica domanda di ricerca che, nel contesto di questo studio e in conformità al duplice obiettivo da conseguire, è possibile riassumere come segue:

> *In che modo si articola il discorso dei redditors intorno all'intelligenza artificiale? Quali sono i maggiori nuclei tematici del dibattito online? Quali relazioni intercorrono tra i temi e le parole più ricorrenti e/o più discriminanti?*

Come già ampiamente discusso, l'opinione pubblica da esaminare nella presente ricerca è quella espressa sul social network Reddit, i cui utenti costituiscono la sfera pubblica dello spazio digitale. Le sezioni precedenti (cfr. *§2.2*) hanno evidenziato come Reddit assuma il ruolo di uno spazio digitale di dibattito collettivo: la piattaforma si presenta come uno dei forum online più visitati al mondo in cui milioni di persone si riuniscono in comunità tematiche per interagire, condividere e discutere contenuti di interesse. In quanto forum di discussione online, Reddit si distingue inoltre per una struttura reticolare ad albero caratterizzata da commenti più discorsivi e argomentativi rispetto ad altri social media, favorendo la raccolta di un campione di testi che possono presentare una dimensione e una ridondanza adatte all'implementazione di metodi statistici. Tuttavia, il principale motivo che giustifica la scelta di Reddit come base analitica di questa ricerca è da ricondurre a criteri di maggiore accessibilità ai dati. Reddit resta difatti una delle poche piattaforme a consentire la raccolta gratuita di informazioni sui propri contenuti, caratteristica che lo rende fonte ottimale per l'analisi di commenti testuali generati dagli utenti. In linea con la tematica sociale affrontata nel presente lavoro, che richiama inevitabilmente le controversie connesse all'impiego dell'intelligenza artificiale, ChatGPT è stato utilizzato come strumento metodologico per sviluppare un codice[47] in linguaggio Python. Sfruttando un'interfaccia di programmazione, l'API (*Application*

[47] Per consultare il codice Python impiegato per l'estrazione dei commenti, si rimanda all'Appendice A.

Programming Interface)[48] ufficiale di Reddit[49], il codice è in grado di raccogliere un ampio numero di commenti. Trattandosi di una tesi in ambito comunicativo e non informatico, si ritiene necessario soffermarsi brevemente sui passaggi tecnici eseguiti in questa fase di raccolta, così da garantirne una maggiore chiarezza.

Il primo passo ha previsto la creazione di un account developer e la registrazione di un'applicazione su Reddit, procedure indispensabili per ottenere le credenziali di accesso autorizzato all'API (Reddit 2025b, 2023a, 2023b)[50]. Per salvare ed eseguire lo script è stato utilizzato l'editor di codice Visual Studio Code (2025): dopo aver installato la versione compatibile di Python e le librerie richieste, lo script (Script_CorpusReddit_perpost.py) ha potuto ricevere in input l'URL della singola submission in analisi, per poi essere eseguito tramite il terminal integrato nell'editor. Qui il terminal richiama il file .py, di volta in volta aggiornato con il nuovo URL, restituendo per ciascuna submission un file contenente i dati raccolti organizzati in tabella: il testo dei commenti (comprese le risposte ai commenti principali) e informazioni contestuali, quali gli autori, la differenza (*score*) tra *upvotes* e *downvotes* assegnati ai commenti, le relazioni gerarchiche tra i commenti, il titolo della submission e il nome del subreddit di provenienza. Occorre a tal proposito precisare come l'estrazione dei dati sia stata effettuata con finalità esclusivamente accademiche, in conformità con i Termini di Servizio di Reddit (2025c) e nel rispetto delle buone pratiche di etica della ricerca: la raccolta ha difatti riguardato esclusivamente contenuti pubblici, senza violare la privacy degli utenti. Gli username sono stati conservati unicamente per tracciare la struttura del dialogo (ad esempio, per individuare a quale commento era rivolta una risposta), ma non sono stati utilizzati a fini di profilazione o diffusione.

Tale procedimento ha permesso di collezionare i commenti di 28 submission pubbliche, appartenenti a subreddits in lingua italiana e pertinenti al tema dell'intelligenza artificiale. I commenti sono stati successivamente aggregati in un unico corpus,

[48] Un'*Application Programming Interface* (API) può essere definita come un insieme di funzioni, procedure o metodi forniti da un sistema operativo, una libreria o un servizio per consentire a un software di interagire con un altro software o di richiedere l'esecuzione di specifiche operazioni (ISO/TS 22386:2024, ISO/TS 6226:2025).

[49] Per la costruzione del corpus, si è preso a modello il funzionamento di strumenti di estrazione dati, come YouTube Data Tools (2015), una collezione di moduli che, inserendo il *video id* di un video YouTube, consente di estrarne anche i commenti associati tramite l'API della piattaforma.

[50] I parametri *client_id*, *client_secret* e *user_agent* sono utilizzati dallo script per inviare richieste all'API e accedere ai dati in modo conforme ai protocolli di autenticazione previsti da Reddit.

organizzato in quattro subcorpora tematici al fine di attuare una comparazione interna tra gruppi di submission. L'estrazione dei commenti è stata difatti preceduta da una lettura esplorativa delle submission che ha permesso l'identificazione di quattro macro-tematiche all'interno del discorso sull'intelligenza artificiale, le quali hanno funto da guida per la selezione delle submission da cui trarre i commenti. La logica di raggruppamento è pertanto stata basata sull'appartenenza della submission ad una delle seguenti categorie concettuali: AI e lavoro; AI e istruzione; applicazioni dell'AI nel quotidiano; AI, salute mentale e relazioni. Nello specifico, l'ampiezza di Reddit ha richiesto l'adozione di un campionamento a scelta ragionata di tipo tipologico-fattoriale[51]: all'interno della totalità delle submission relative all'AI, è stato estratto un campione di 28 submission, sette per ciascuna delle quattro macro-tematiche individuate, rendendo il corpus complessivo tematicamente bilanciato. Tali macro-tematiche costituiscono i subcorpora in cui il corpus complessivo è stato ripartito sulla base della proprietà di interesse 'argomento', con cui si fa riferimento al focus tematico delle submission madri. Si distinguono, dunque, i quattro subcorpora *sub1_lav*, *sub2_edu*, *sub3_app* e *sub4_rel*. Per assicurare la significatività[52] dell'insieme di testi selezionato, le submission sono state scelte sulla base di due criteri fondamentali, l'uno di tipo qualitativo e l'altro di tipo quantitativo: la rilevanza tematica della submission (intesa come grado di pertinenza rispetto a una delle quattro macro-tematiche) ed il numero di commenti[53]. Sono dunque escluse submission vaghe o generiche, non chiaramente centrate sulla macro-area in questione, oltre che submission poco popolari all'interno della community. Si ritiene inoltre opportuno segnalare un fattore secondario nella selezione: la data di pubblicazione delle submission. In prima battuta, si è scelto di privilegiare le submission più recenti, pubblicate tra gennaio 2024 e agosto 2025 (mese in cui è stata attuata l'estrazione dei commenti), di modo che potessero riflettere al meglio il discorso attuale sull'intelligenza

[51] Un campionamento non probabilistico che richiede l'equa ripartizione numerica delle classi, in modo da porre sotto controllo le variabili stratificatrici per mezzo della loro neutralizzazione. Nei campionamenti non probabilistici, il criterio di estrazione dei casi si dice generalmente "a scelta ragionata", poiché è il ricercatore che, sulla base delle proprie domande ricerca, determina i criteri per stabilire quali casi andranno a far parte del campione (Corbetta 2015).

[52] Da intendersi come rappresentatività del campione rispetto all'oggetto di ricerca. Nel caso di campionamenti a scelta ragionata, non è possibile applicare i procedimenti dell'inferenza statistica, che è propria solo dei campionamenti probabilistici: il campione non è rappresentativo statisticamente, bensì rappresentativo della varietà e della ricchezza dell'oggetto di studio (Corbetta 2015; Tuzzi 2024).

[53] Per quanto riguarda quest'ultimo criterio, è stata riservata particolare attenzione a garantire un adeguato bilanciamento tra i diversi subcorpora in termini di numero e ampiezza dei commenti, condizione necessaria per operare confronti tra insiemi di testi.

artificiale; successivamente, l'intersezione con i criteri di rilevanza tematica e di numerosità dei commenti ha richiesto l'estensione del campione a submission risalenti al 2023 e al 2022. In ragione della rapida evoluzione del tema in analisi, nei mesi seguenti è stata condotta una terza esplorazione su Reddit per verificare la presenza di nuove, ulteriori submission che potessero essere conformi ai criteri di selezione, ma non sono emersi risultati rilevanti.

In sintesi, possiamo riassumere dicendo che il campionamento tipologico-fattoriale per la compilazione del corpus è stato guidato dall'argomento delle submission e dalla dimensione di ciascuna discussione. Da un lato, questo approccio ha consentito di focalizzare la ricerca sugli aspetti dell'AI maggiormente discussi dagli utenti, rispecchiando i reali centri di interesse che emergono nel dibattito digitale ma al contempo costruendo un corpus tematicamente bilanciato (ciascun 'argomento' è adeguatamente rappresentato); dall'altro, ha permesso di ottenere un campione linguistico più ampio. Vale la pena sottolineare che, per quanto questa fase preliminare di selezione non sia stata assistita da software CAQDAS, ne condivide la logica qualitativa ed ermeneutica *ex ante*, in quanto interpretabile come un processo di codifica attraverso il quale il «contenuto» di ciascuna submission viene riconosciuto «coerente» con una categoria concettuale (uno dei subcorpora).

Una volta definita la composizione generale del corpus, risulta fondamentale soffermarsi sulla natura e sul numero degli elementi testuali che lo costituiscono. In quanto commenti online in lingua italiana, trattanti la discussione Reddit sull'intelligenza artificiale, è possibile affermare che i commenti sono accumunati dalle proprietà testuali della lingua, del genere (quello argomentativo e partecipativo dei testi postati sui social media), della fonte e del macro-argomento. Nello specifico, il corpus consta di 5245 commenti validi (su 5696 commenti totali), inclusi quelli nidificati, provenienti da 28 submission pubblicate sui subreddits *r/Italia*, *r/Universitaly*, *r/italy*, *r/CasualIT*, *r/Psicologia_Italia* e *r/psicologia*.

	Post id	*Commenti*	*Commenti validi*	*Commenti-genitore*
sub1_lav *(impiego dell'AI in ambito lavorativo, automazione, ansie occupazionali)*	post1_lav	299	273	99
	post2_lav	262	233	100
	post3_lav	248	244	55
	post4_lav	215	213	47
	post5_lav	166	135	47
	post6_lav	143	133	49
	post7_lav	126	123	37
	tot.	**1459**	**1354**	**434**
sub2_edu *(impiego dell'AI in ambito educativo, implicazioni sull'apprendimento)*	post1_edu	340	329	121
	post2_edu	276	260	91
	post3_edu	222	219	62
	post4_edu	190	179	58
	post5_edu	157	132	28
	post6_edu	95	88	28
	post7_edu	66	61	29
	tot.	**1346**	**1268**	**417**
sub3_app *(varietà di impiego dell'AI, impatti sulla creatività, produttività personale)*	post1_app	385	344	72
	post2_app	262	249	121
	post3_app	215	213	106
	post4_app	208	191	96
	post5_app	192	179	102
	post6_app	135	128	49
	post7_app	126	123	43
	tot.	**1523**	**1427**	**589**
sub4_rel *(impiego dell'AI come terapeuta o strumento di ascolto e sfogo, impatto sulla salute mentale e sulle relazioni sociali)*	post1_rel	415	408	79
	post2_rel	350	231	79
	post3_rel	184	164	64
	post4_rel	169	162	72
	post5_rel	99	93	40
	post6_rel	90	79	32
	post7_rel	61	59	19
	tot.	**1368**	**1196**	**385**
corpus		**5696**	**5245**	**1825**

Tabella 2 Composizione del corpus: organizzazione in subcorpora tematici e relative statistiche sui commenti di ciascuna submission.

La *Tabella 2*[54] riporta, per ogni subcorpora, l'identificativo di ciascuna submission (*post id*[55]) ed il relativo numero di commenti totali, di commenti validi e di commenti-genitore (o *parent comments*[56]). Si parla di "commenti validi" precisamente perché lo script utilizzato per l'estrazione esclude i commenti eliminati o rimossi, i quali, pur figurando nel conteggio complessivo di Reddit, risultano vuoti. Vengono invece mantenuti quelli di account eliminati (*[deleted]*), ma con testo ancora visibile. Tramite una ricerca keywords è stata successivamente operata un'eliminazione manuale dei commenti generati dai bot e dei commenti esclusivamente contenenti link o gif, e quindi privi di testo analizzabile. Inoltre, all'interno di alcuni commenti lo script include automaticamente il testo del commento a cui l'utente risponde (contrassegnato dal simbolo ">"): analogamente ai casi sopracitati, anche queste porzioni di testo sono state rimosse in fase di pulizia del corpus per evitare duplicazioni. Il controllo ortografico di Excel ha infine consentito, laddove possibile, di correggere manualmente errori di battitura o di accento per evitare il conteggio multiplo di forme grafiche in realtà uguali.

3.1.1 Ridondanza, struttura interna e tassi di copertura

La presente sezione esplorerà i parametri generali del corpus e la struttura interna dei singoli subcorpora che lo costituiscono. Per fornirne le misure generali, si utilizzano AntConc (Anthony 2020) e Voyant Tools (Sinclair & Rockwell 2026), due software per l'analisi testuale dei corpora.

Come anticipato, il numero di N di parole totali e il numero V di parole diverse forniscono una prima semplice valutazione della dimensione del corpus e del suo vocabolario. Per l'identificazione delle unità testuali minime del corpus, si è scelto di considerare come separatori tutti i caratteri che non sono lettere dell'alfabeto, eliminando automaticamente dal conteggio dei *token* tutti i numeri e segni di punteggiatura. Dopo aver effettuato questa procedura di tokenizzazione e aver normalizzato tutte le parole in

[54] Si rimanda all'Appendice B per una versione integrativa della tabella, comprensiva dei collegamenti ipertestuali alle submission di riferimento e del trimestre di pubblicazione.
[55] I termini "post" e "submission" possono essere utilizzati in modo intercambiabile. In questo contesto, si sceglie di adottare la dicitura *post id* anziché l'abbreviazione *sub* (da *submission*) al fine di evitare ambiguità con l'abbreviazione *sub* già impiegata per indicare i subcorpora.
[56] Nel contesto delle *threaded online conversations*, il termine *parent comments* indica i commenti diretti, ossia i commenti che rispondono direttamente al post originale, contrariamente ai *child comments* (le risposte a un *parent comment*).

minuscolo, è possibile concludere che il corpus oggetto di studio conta 263.249 *word-tokens* (o occorrenze), con una media di 9.401 *word-tokens* per singola submission[57] (*mean size*), e 20.255 *word-types*[58]. Il subcorpus più lungo, *sub1_lav*, risulta composto da 76.293 *word-tokens*, a cui si contrappone il più breve *sub2_edu*, che consta di 61.511 *word-tokens*. Entrambi si configurano peraltro come i subcorpora più ricchi a livello sintattico, con una media di 23.9 (*sub1_lav*) e 24.4 (*sub2_edu*) parole per frase. Questa fase preliminare di analisi conduce dunque ad una prima riflessione: benché quantitativamente minori, i commenti a submission inerenti all'utilizzo dell'AI in ambito educativo si presentano come i più discorsivi e argomentativi.

Tuttavia, prima di procedere con ulteriori considerazioni, è necessario verificare che il corpus presenti un livello di ridondanza tale da consentire l'impiego di un approccio statistico basato sulla frequenza di parole. La collezione di commenti è stata a tal scopo sottoposta al calcolo del Type-Token Ratio e dell'indice di hapax%, valori che, come preannunciato, dovrebbero essere inferiori o uguali alle rispettive soglie empiriche del 20% e del 50%. Si riassumono nella tabella sottostante i risultati di questa prima analisi:

	N	***Mean size***	***V***	***TTR%***	***V_1***	***hapax%***	***Words per sentence***
sub1_lav	76.293	10.899	10.091	13,2%	5.598	55,4%	23,9
sub2_edu	61.511	8.787	8.285	13,4%	4.721	56,9%	24,4
sub3_app	61.925	8.846	8.392	13,5%	4.728	56,7%	22,2
sub4_rel	63.520	9.074	8.292	13,0%	4.686	56,5%	22,9
corpus	263.249	9.401	20.255	**7,6%**	10.318	**50,9%**	**23,3**

Tabella 3 Descrizione del corpus: analisi preliminare dei parametri.

Con un TTR pari al 7,6% ed una percentuale di hapax corrispondente al 50,9% delle forme, il corpus oggetto di analisi può essere considerato sufficientemente esteso da essere trattato statisticamente, seppur con un vocabolario leggermente più originale del

[57] Si nota che il testo delle submission non costituisce oggetto di analisi: per riferirsi agli insiemi di commenti presenti sotto ciascuna di esse si preferisce la formulazione "per singola submission" esclusivamente per questioni di praticità nell'esposizione.

[58] I dati qui riportati sono stati ottenuti tramite AntConc e Voyant. È opportuno tuttavia precisare che l'analisi ha previsto l'utilizzo di diversi software, ciascuno dei quali adotta diversi criteri di tokenizzazione e normalizzazione. Ne consegue che il numero di occorrenze totali e forme varia leggermente: se AntConc e Voyant identificano 263.249 *word-tokens* e 20.255 *word-types*, RStudio ne rileva 262.751 e 21.092, mentre IRaMuTeQ 265.161 e 20.536.

necessario. Va ricordato come, nel parlare di TTR e dimensione del corpus, non si faccia propriamente riferimento alla quantità di *word-tokens* e *word-types*, ma al grado di ridondanza lessicale che ne deriva: come già precisato, un corpus più lungo implica una maggiore ripetizione delle stesse forme grafiche, con una conseguente diminuzione della variazione lessicale. Questa, in termini statistici, si traduce in un basso valore di TTR, accompagnato nel caso qui esaminato da un indice di hapax% che, posizionandosi appena al di sopra della soglia empirica, conferma la ridondanza complessiva del vocabolario. D'altra parte, se considerati isolatamente, i singoli subcorpora rispettano la soglia empirica di TTR, ma rivelano percentuali di hapax più elevate, comprese tra il 55,4% (*sub1_lav*) e il 56,9% (*sub2_edu*). In ciascun subcorpus, la predominanza di parole che occorrono solo una volta riflette il carattere informale ed eterogeneo delle conversazioni online: ciascun gruppo di submission raccoglie commenti generati da utenti diversi che, nell'esprimere la propria opinione, ricorrono frequentemente ai termini chiave tipici del focus tematico discusso; ciononostante, ciascun utente, nella propria diversità, introduce espressioni originali, legate a slang, stile personale e/o refusi ortografici[59], che incidono sulla percentuale di hapax di ciascuna raccolta. Pertanto, si è deciso di proseguire ugualmente con l'analisi, ma di tenere in considerazione questa problematica nel caso emergano risultati ambigui in corso d'opera. Trattandosi difatti del rumore tipico dei commenti online, non si è ritenuto necessario espandere ulteriormente i subcorpora. In particolare, l'aumento del numero di commenti comporterebbe con ogni probabilità l'aumento di refusi e, di conseguenza, di parole originali utilizzate un'unica volta, non apportando alcun beneficio significativo ai presenti parametri.

Avendo precisato questo punto, è possibile ora concentrarsi sulle parole maggiormente impiegate dai redditors. La *Tabella 4* offre uno stralcio della wordlist generata tramite AntConc, ossia una lista di parole ordinata per frequenze decrescenti, illustrando i 15 *word-types* più ricorrenti all'interno del vocabolario in analisi:

[59] Da un'analisi qualitativa degli hapax emergono, ad esempio, anglicismi (e.g., *apologies*, *buddy*, *btw*, *main*, *pls*, *linkerei*, *pointless*) e diverse forme linguistiche comunemente utilizzate sulle piattaforme per esprimere ilarità (e.g., *ahahah*, *hahaha*) o altre emozioni (e.g., *aaah*, *bha*, *boh*, *chessò*, *chissene*, *daje*, *eheh*, *embè*).

Conteggio	*Word-type*	*Frequenza assoluta*	*Frequenza relativa*	*Tasso x1,000*	*TCV x100*	*Frequenza cumulata*	*TCC x100*	*Fascia*
1	che	8024	0,03048	30,48	0,005	8024	3,048	alta
2	di	7737	0,02939	29,39	0,010	15761	5,987	alta
3	e	6201	0,02356	23,56	0,015	21962	8,343	alta
4	non	6109	0,02321	23,21	0,020	28071	10,663	alta
5	è	5197	0,01974	19,74	0,025	33268	12,637	alta
6	un	4558	0,01731	17,31	0,030	37826	14,369	alta
7	a	4299	0,01633	16,33	0,035	42125	16,002	alta
8	per	4241	0,01611	16,11	0,039	46366	17,613	alta
9	il	3926	0,01491	14,91	0,044	50292	19,104	alta
10	la	3684	0,01399	13,99	0,049	53976	20,504	alta
11	in	3555	0,01350	13,50	0,054	57531	21,854	alta
12	una	2460	0,00934	9,34	0,059	59991	22,789	alta
13	ma	2452	0,00931	9,31	0,064	62443	23,720	alta
14	l	2341	0,00889	8,89	0,069	64784	24,609	alta
15	se	2161	0,00821	8,21	0,074	66945	25,430	alta

Tabella 4 I 15 *word-types* più frequenti all'interno del vocabolario in analisi e i relativi parametri.

La tabella riporta, in ordine, il conteggio progressivo dei 15 *word-types*, il numero di occorrenze (*frequenza assoluta*), la frequenza relativa, il tasso per 1.000 parole di ciascun *word-type*, il TCV%, la frequenza cumulata, il TCC% e la fascia di frequenza a cui appartiene ciascun *word-type*. Le sigle TCV e TCC indicano, rispettivamente, il Tasso di Copertura del Vocabolario e il Tasso di Copertura del Corpus, vale a dire la porzione di vocabolario e la porzione di corpus che i *word-types* e i *word-token* con una frequenza superiore o uguale a una data soglia di frequenza *y* consentono di coprire. Prima di analizzare il vocabolario, è difatti necessario definire le porzioni di corpus con cui condurre l'analisi: per far fronte alla numerosità delle unità testuali nei corpora di grandi dimensioni, viene spesso operata una selezione a monte che ne riduca il numero. Per garantire la rappresentatività di tale selezione rispetto all'interezza del corpus, si è deciso di lavorare con l'insieme dei *word-types* aventi una frequenza assoluta maggiore o uguale a 8. Come evidenziato in verde nella *Tabella 5*, tale soglia di frequenza assicura infatti un Tasso di Copertura del Corpus pari all'87%, all'interno del quale ricadono 2.963 *word-types*. Inoltre, stabilendo una soglia di frequenza assoluta maggiore o pari a 8, il Tasso di Copertura del Vocabolario, ossia il volume di vocabolario che la suddetta porzione di *word-types* è in grado di coprire, corrisponde a circa il 15%.

Conteggio	*Word-type*	*Freq. assoluta*	*Freq. relativa*	*Tasso x1,000*	*TCV x100*	*Freq. cumulata*	*TCC x100*	*Fascia*
1	che	8024	0,03048	30,48	0,005	8024	3,048	alta
2	di	7737	0,02939	29,39	0,010	15761	5,987	alta
3	e	6201	0,02356	23,56	0,015	21962	8,343	alta
::	::	::	::	::	::	::	::	::
::	::	::	::	::	::	::	::	::
16	le	2033	0,00772	7,72	0,079	68978	26,203	alta
17	con	1859	0,00706	7,06	0,084	70837	26,909	alta
18	come	1819	0,00691	6,91	0,089	72656	27,600	media
19	i	1819	0,00691	6,91	0,094	74475	28,291	media
20	si	1789	0,00680	6,80	0,099	76264	28,970	media
21	più	1776	0,00675	6,75	0,104	78040	29,645	media
::	::	::	::	::	::	::	::	::
::	::	::	::	::	::	::	::	::
300	internet	108	0,00041	0,41	1,481	164889	62,636	media
301	mano	108	0,00041	0,41	1,486	164997	62,677	media
302	devo	107	0,00041	0,41	1,491	165104	62,718	media
303	accordo	105	0,00040	0,40	1,496	165209	62,758	bassa
304	far	105	0,00040	0,40	1,501	165314	62,798	bassa
305	capisco	104	0,00040	0,40	1,506	165418	62,837	bassa
::	::	::	::	::	::	::	::	::
::	::	::	::	::	::	::	::	::
2961	volete	8	0,00003	0,03	14,619	229950	87,351	bassa
2962	vostro	8	0,00003	0,03	14,624	229958	87,354	bassa
2963	wow	8	0,00003	0,03	14,628	229966	87,357	bassa
2964	abstract	7	0,00003	0,03	14,633	229973	87,359	bassa
2965	accademico	7	0,00003	0,03	14,638	229980	87,362	bassa
2966	accaduto	7	0,00003	0,03	14,643	229987	87,365	bassa
::	::	::	::	::	::	::	::	::

Tabella 5 Tassi di copertura e fasce di frequenza.

Si fa notare come l'ultima colonna delle *Tabelle 4* e *5* mostri la fascia di frequenza di ciascun *word-type*. Data una lista di frequenza ordinata per frequenze decrescenti, il vocabolario può difatti essere ripartito in fasce. Le più comunemente utilizzate in letteratura (Tuzzi 2024; Bolasco 2011) sono tre: la fascia di alta frequenza, la fascia di media frequenza e la fascia di bassa frequenza. Secondo Bolasco (1999, p. 202; cfr. Tuzzi 2024), la fascia delle alte frequenze è generalmente "[…] composta all'incirca da 30 o 50 forme (a seconda delle dimensioni del corpus) e, fra queste, al più 4 o 5 sono parole principali, mentre le altre sono parole grammaticali", quali articoli, pronomi,

preposizioni, congiunzioni e verbi ausiliari. Nel contesto dell'analisi dei dati testuali, un'ulteriore distinzione fondamentale è infatti quella tra le parole grammaticali (o funzionali), in letteratura chiamate *stop words*, e le parole di contenuto, anche definite *content words*. Le *stop words* sono le parole più frequentemente utilizzate nel linguaggio e si concentrano pertanto nella fascia delle alte frequenze, occupando le prime posizioni di una qualsiasi wordlist, specialmente nel caso di corpora di grandi dimensioni. Poiché non portatrici di significato lessicale, vengono spesso ritenute poco utili ai fini dell'analisi testuale ed eliminate dal corpus prima della fase di elaborazione. Le parole di contenuto sono invece i *word-types* semanticamente pieni, dotati di un contenuto lessicale, che all'interno della fascia di alta frequenza – in cui sono spesso presenti in numero limitato – rappresentano il tema principale dei testi raccolti (Tuzzi 2024; Tuzzi 2012). Chiarita questa distinzione, si segnala come i risultati del caso in esame contrastino con la tendenza osservata da Bolasco (1999): come si può osservare dalla *Tabella 5*, la fascia di alta frequenza del vocabolario dei commenti Reddit conta solo 17 *word-types*, unicamente grammaticali ("che", "di", "e", "non", etc.), senza alcuna parola di contenuto. In particolare, la parola più frequente è il pronome relativo *che*, ripetuto 8024 volte (circa 30 volte ogni 1.000 parole). Si ipotizza che anche questi risultati siano riconducibili all'alta dispersione lessicale tipica delle discussioni online: nonostante il corpus sia monotematico, il tema dell'intelligenza artificiale viene discusso dagli utenti alternando parole di contenuto diverse, nessuna delle quali è conseguentemente in grado di predominare a tal punto da inserirsi tra le parole grammaticali; inoltre, i thread di discussione possono essere intesi come sequenze dialogiche all'interno delle quali è raro che parole di contenuto come "intelligenza artificiale", "ChatGPT", "IA" o "AI" appaiano in ogni commento, in quanto già sottintese nello scambio conversazionale.

Con una frequenza compresa tra 8024 e 1859, i 17 *word-types* appartenenti alla fascia di alta frequenza rappresentano lo 0,09% del vocabolario (0,084% TCV) e il 27% del corpus (26,909% TCC). La fascia delle medie frequenze ha inizio dalla classe di frequenza 1819 e si interrompe alla classe di frequenza 107, per un totale di 285 parole diverse. Insieme, la fascia di alta frequenza e la fascia di media frequenza includono 302 *word-types* con un numero di occorrenze pari o maggiore di 107, offrendo una copertura del vocabolario pari all'1,5% (1,491% TCV) e una copertura del corpus pari al 63% (62,718% TCC). Con un numero di occorrenze pari o inferiore a 105, a concludere il

vocabolario sono i 19.953 *word-types* appartenenti alla fascia di bassa frequenza, che rappresentano il 98,5% del vocabolario e il 37,3% del corpus. Rientrano in quest'ultima fascia anche i *word-types* che si presentano un'unica volta, gli hapax, che, come anticipatamente illustrato, costituiscono il 50,9% del vocabolario e il 3,9% del corpus.

3.2 La rielaborazione dei dati in proprietà misurabili

Una volta accertata l'applicabilità di un approccio statistico (cfr. *§3.1.1*), i 5.245 commenti online pubblicati dai redditors sulle 28 submission selezionate possono essere rielaborati in misure statistiche interpretabili. Tale fase di sintesi consente di soverchiare le difficoltà che emergerebbero nel caso di un'analisi esclusivamente di tipo qualitativo: grazie all'ASDT, grandi quantità di dati, che altrimenti richiederebbero una lettura integrale del materiale in analisi, possono essere comprese, interpretate e comunicate (Bernardi & Campostrini 2005) dal ricercatore mediante la loro rielaborazione in proprietà misurabili.

Come riassunto in *Tabella 3*, il corpus oggetto di studio consta di 263.249 *word-tokens* e 20.255 *word-types*. Esaminandone la distribuzione nei quattro subcorpora, è possibile avanzare alcune osservazioni. A parità di numero di submission per ciascuna macro-tematica, *sub1_lav* presenta il maggior numero di occorrenze (76.293 *word-tokens*), costituendo il subcorpus più lungo. Il dato suggerisce che, tra i settori interessati dall'intelligenza artificiale, quello lavorativo incentivi un dibattito mediamente più esteso. Contrariamente, le discussioni relative ai rapporti tra l'intelligenza artificiale e l'ambito educativo, applicativo e relazionale mostrano dimensioni comparabili: *sub2_edu*, *sub3_app* e *sub4_rel* registrano, rispettivamente, 61.511, 61.925 e 63.520 occorrenze. Ciononostante, la *Tabella 2* indica l'utilizzo quotidiano dell'AI come macro-tematica caratterizzata dal maggior livello di interazione: con 1.427 commenti complessivi, *sub3_app* comprende submission che stimolano un'elevata partecipazione da parte degli utenti, impegnati in conversazioni frammentarie, poco discorsive ma frequenti. Inoltre, una media di 24.4 parole per frase (*words per sentece*) rende *sub2_edu*, il subcorpus più breve, primo per ricchezza sintattica. Si ipotizza, dunque, che l'adozione dell'AI nel contesto educativo favorisca interventi meno numerosi ma più argomentativi e, in generale, caratterizzati da strutture sintattiche più complesse: l'ambito educativo non genera dialoghi serrati, ma un numero limitato di commenti in cui l'utente tende a

esprimere un ragionamento completo in un singolo turno. In sintesi, i parametri dei subcorpora rivelano su Reddit tendenze diverse a seconda del contesto d'uso dell'AI: il lavoro produce discussioni più estese, le applicazioni un maggior numero di interventi e l'educazione discorsi più complessi. Il nucleo tematico della salute mentale e delle relazioni si colloca invece su valori intermedi.

3.2.1. Le parole più frequenti

Rappresentare il vocabolario di un corpus mediante una lista di frequenza consente di identificarne le prime parole di contenuto e individuarne i principali argomenti trattati. Osservando le *Tabelle 4* e 5 si è già concluso come nel caso in esame la fascia di alta frequenza sia esclusivamente costituita da *stop words*, forme grammaticali prive di significato lessicale: tra le 17 forme più ricorrenti all'interno delle *threaded online conversations* di Reddit non emerge alcuna parola di contenuto. Come precisato (cfr. *§3.1.1*), la dimensione ridotta della fascia è attribuibile all'elevata dispersione lessicale tipica delle discussioni online e, in particolare, alla loro natura dialogica: termini centrali come "intelligenza artificiale", "ChatGPT", "IA" o "AI" tendono ad essere sottintesi all'interno dei commenti nidificati e pertanto non compaiono in ogni intervento. Per mostrare in modo rapido e intuitivo i termini o i temi più rilevanti presenti nei testi raccolti, l'analisi dei dati testuali ricorre spesso alle *word cloud*, grafici efficaci dal punto di vista comunicativo nei quali la dimensione delle parole ne riflette la frequenza (Tuzzi 2012). La *word cloud* in *Figura 10* offre una rappresentazione visiva delle 75 parole più ricorrenti all'interno del corpus, la cui frequenza supera le 429 occorrenze. Tra queste, prevalgono natrualmente pronomi, preposizioni, articoli e congiunzioni. Al fine di identificare con chiarezza gli argomenti discussi nei commenti delle submission selezionate, si propone un'ulteriore *word cloud* (si veda *Figura 11*) da cui si escludono le parole funzionali[60]: alle *function words* "che", "di", "e", "non", "per", "un", "in", "ma"

[60] Si applica, in tal caso, la lista di 672 *stop words* fornita da GitHub, una lista estesa che, a differenza di quelle standard, prende in considerazione le varianti morfologiche delle parole grammaticali, riducendo il rumore lessicale. La sua applicazione favorisce innanzitutto l'interpretabilità dell'analisi quantitativa e, secondariamente, la visibilità grafica delle parole di contenuto. La lista, ridotta a 642 *stop words*, è stata revisionata manualmente al fine di preservare parole che, seppur classificate come *stop words* nella versione originale, sono ritenute portatrici di contenuto informativo (*i.e.*, "anni", "anno", "anticipo", "attesa", "casa", "cima", "cortesia", "ex", "favore", "fine", "futuro", "gente", "giorni", "lavoro", "mancanza", "ministro", "mondo", "momento", "nome", "peccato", "persone", "piedi", "possedere", "rispetto", "scopo", "tempo", "titolo", "torino", "uomo", "vita").

si sostituiscono le *content words* “chatgpt”, “lavoro”, “persone”, “problema”, “ia”, “uso” e “anni”, parole semanticamente piene che offrono una rappresentazione più mirata dei contenuti del corpus.

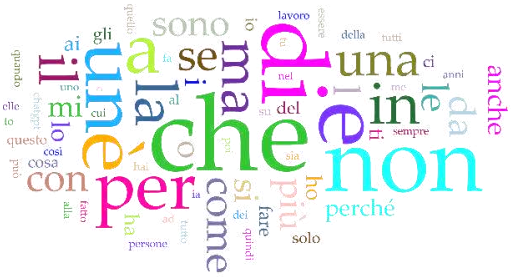

Figura 10 *Word cloud* basata sulle frequenze: le 75 *most frequent words* (MFW) del corpus complessivo, includendo le *stop words*.

Figura 11 *Word cloud* basato sulle frequenze: le 75 *most frequent words* (MFW) del corpus complessivo, escludendo le *stop words*.

La tabella nella pagina successiva (si veda *Tabella 6*) riporta la wordlist a partire dalla quale è stata generata la *word cloud* in *Figura 11*. In particolare, si raffigurano le parole di contenuto più utilizzate dai redditors nel dibattito sull’intelligenza artificiale. Si nota come, nonostante quest’ultima designi il tema protagonista del corpus e il comun denominatore delle 28 submission selezionate, il termine “intelligenza” (f = 163), costituisca il 191° *word-type* per frequenza, mentre l’aggettivo “artificiale”, attestandosi alla 273ª posizione con sole 120 occorrenze totali, non figura tra i *word-types* riportati in *Tabella 6*. Gli utenti tendono difatti a riferirsi all’intelligenza artificiale con i più semplici acronimi I.A. (“ia”, f = 645) o A.I. (“ai”, f = 1.154)[61] dall’inglese *artificial intelligence*. La prima parola semanticamente piena a comparire nella lista di frequenza si colloca soltanto al 43° posto: con un numero di occorrenze totali pari a 747, “chatgpt”[62] è la parola di contenuto maggiormente impiegata nei commenti online, a conferma di come il

[61] L’acronimo AI potrebbe essere considerato la prima *content word* osservabile, collocata in 30ª posizione con una frequenza assoluta pari a 1.154. Questo mette tuttavia in luce il limite delle classificazioni delle parole basate su liste: in assenza del contesto d’uso, il *word-type* “ai” presenta un’ambiguità intrinseca, poiché non può essere unicamente attribuito all’intelligenza artificiale. Le sue 1.154 occorrenze possono infatti riferirsi tanto all’acronimo di *artificial intelligence*, quanto alla comune preposizione articolata “ai” (*a* + *i*), ampiamente utilizzata in lingua italiana.

[62] Si osserva che, all’interno del corpus, “chat” e “gpt” si presentano anche separatamente con un numero di occorrenze pari a, rispettivamente, 225 e 274. Di queste, una quota considerevole deriva dalla co-occorrenza dei termini che, pur susseguendosi in modo lineare (“chat gpt”), costituiscono due token distinti. Tuttavia, è doveroso precisare che il costituente “gpt” è talvolta impiegato singolarmente dai redditors: la

	Word-type	*Frequenza assoluta (f)*		*Word-type*	*Frequenza assoluta (f)*
::	::	::	*147*	problemi	218
43	chatgpt	747	*::*	::	::
::	::	::	*149*	so	218
48	lavoro	661	*::*	::	::
::	::	::	*151*	risposta	211
51	ia	645	*::*	::	::
::	::	::	*154*	lavori	204
62	cose	541	*155*	strumento	204
63	anni	534	*156*	umano	203
::	::	::	*::*	::	::
75	persone	429	*163*	vero	189
::	::	::	*::*	::	::
80	problema	401	*170*	llm	176
::	::	::	*171*	ricerca	176
86	uso	372	*172*	sa	175
::	::	::	*::*	::	::
94	tempo	332	*174*	devi	173
::	::	::	*::*	::	::
117	gpt	274	*179*	vedo	167
118	scrivere	274	*180*	scritto	166
::	::	::	*::*	::	::
122	punto	261	*182*	codice	165
123	puoi	261	*183*	senso	165
::	::	::	*::*	::	::
132	mondo	245	*186*	parlare	164
::	::	::	*::*	::	::
134	usare	245	*188*	risposte	164
135	vita	245	*::*	::	::
::	::	::	*190*	google	163
137	gente	243	*191*	intelligenza	163
::	::	::	*::*	::	::
139	tipo	237	*194*	dati	161
::	::	::	*::*	::	::
142	capire	229	*196*	utile	158
::	::	::	*::*	::	::
144	chat	225	*198*	futuro	157
145	persona	224	*::*	::	::
::	::	::	*200*	fine	156

Tabella 6 Stralcio del vocabolario del corpus complessivo: le *content words* più utilizzate dai redditors nel dibattito sull'intelligenza artificiale (200 MFW).

lessicalizzazione di "gpt" segnala che il nome del chatbot è sufficientemente riconoscibile da non richiedere l'ancoraggio a "chat".

concetto di intelligenza artificiale trovi ormai un'associazione diretta con il chatbot introdotto da OpenAI a novembre 2022. Nel discorso pubblico attuale, parlare di AI implica inevitabilmente un riferimento immediato al primo assistente virtuale destinato al mercato di massa, utilizzato oggi da 700 milioni di utenti (Chatterji et al. 2025). Di grande rilievo all'interno del corpus è dunque il crescente ruolo dell'AI generativa (GenAI), termine con cui si fa riferimento a modelli che "[…] can create novel content (including text, images, audio and videos) and insights that are often indistinguishable from human-created content, based on patterns and structures learnt from training data" (Calvino et al. 2025, p. 16)[63].

Le quattro distinte wordlist in *Tabella 7* illustrano, in ordine di frequenza decrescente, le 30 parole di contenuto più ricorrenti per ciascuno dei quattro subcorpora. La tabella è generata a partire dalla matrice "parole × testi" (TDM) che, incrociando il vocabolario *V* del corpus con i subcorpora *M* che lo compongono, fornisce la distribuzione delle frequenze di ciascun *word-type* lungo *sub1_lav*, *sub2_edu*, *sub3_app* e *sub4_rel*. Oltre che nell'intero corpus testuale, il *word-type* "chatgpt" si rivela essere il termine più frequente in due dei subcorpora, registrando 279 occorrenze in *sub3_app* e 203 occorrenze in *sub2_edu*. Le due distinte raccolte di commenti online condividono peraltro diversi termini ad alta frequenza. *Sub3_app*, che si propone di aggregare le più diffuse applicazioni dell'AI da parte degli utenti nel quotidiano, presenta "uso" (*f* = 219) come secondo *word-type* più frequente. A questo si aggiungono le varianti verbali "usare" e "usato" e le parole "scrivere", "codice", "ricerca" e "testo", forme che ne esplicitano l'utilizzo specifico da parte dai redditors coinvolti nella discussione online. In particolare, "codice" (*f* = 77) suggerisce l'impiego dell'intelligenza artificiale come strumento di supporto alla scrittura o alla risoluzione di codici di programmazione e, accostato a "prompt" e "dati", evidenzia la componente tecnica dei modelli probabilistici. Il subcorpus incentrato sull'ambito applicativo si distingue inoltre per la presenza di "google" (*f* = 78), l'unica Big Tech[64] a rientrare in *Tabella 7*. La sua inclusione tra le 30 *content words* più frequenti in *sub3_app* offre una doppia chiave di lettura: da un lato, Google, in quanto motore di ricerca tradizionale, potrebbe fungere da oggetto di paragone

[63] "[…] sono in grado di creare contenuti originali (inclusi testi, immagini, audio e video) e approfondimenti spesso indistinguibili da quelli prodotti dall'uomo, sulla base di schemi e strutture appresi dai dati di addestramento".

[64] Abbreviazione principalmente utilizzata per riferirsi ai colossi tecnologici statunitensi, talvolta anche indicati come GAFAM: Google, Apple, Facebook, Amazon e Meta (Transnational Institute 2024).

per valutare le prestazioni dei chatbot di GenAI, utilizzati come strumenti di sopporto alla ricerca; dall'altro, il termine potrebbe implicare un riferimento indiretto a Gemini, il chatbot di GenAI sviluppato da Google.

	Word-type	***sub1_lav (f)***	***Word-type***	***sub2_edu (f)***	***Word-type***	***sub3_app (f)***	***Word-type***	***sub4_rel (f)***
1	lavoro	319	chatgpt	203	chatgpt	279	persone	197
2	ia	288	cose	135	uso	219	chatgpt	183
3	anni	190	anni	134	lavoro	203	cose	124
4	lavori	156	ia	128	cose	165	persona	121
5	persone	133	problema	116	scrivere	112	anni	120
6	problema	129	scuola	114	ia	112	amico	120
7	cose	117	lavoro	107	tempo	102	ia	117
8	mondo	98	scrivere	100	anni	90	vita	113
9	intelligenza	93	usare	93	usare	86	psicologo	113
10	gente	85	uso	93	puoi	81	parlare	93
11	futuro	83	tempo	90	gpt	80	problema	86
12	chatgpt	82	gpt	90	tipo	80	chat	81
13	tempo	80	strumento	83	google	78	problemi	81
14	umano	79	capire	78	codice	77	risposta	80
15	punto	74	imparare	76	problema	70	gpt	75
16	vedo	68	llm	75	ricerca	70	amici	74
17	dati	68	compiti	74	punto	68	puoi	72
18	informatica	64	chat	72	usato	63	vuoi	72
19	aziende	63	studenti	72	prompt	60	umano	70
20	artificiale	62	studente	69	chat	59	punto	69
21	problemi	60	studiare	68	risposta	58	vero	68
22	tipo	59	studio	65	testo	57	mondo	66
23	so	59	casa	63	domande	56	so	64
24	vero	58	testo	62	capire	56	tempo	60
25	grado	58	tesi	61	devi	55	davvero	60
26	umani	58	ricerca	60	so	55	gente	59
27	soldi	57	strumenti	59	utile	53	risposte	58
28	puoi	57	domande	59	scritto	53	tipo	55
29	lavorare	56	utile	58	persone	52	realtà	51
30	sa	55	sa	57	dati	52	relazioni	51

Tabella 7 Riadattamento del *Term-Document Matrix*: le 30 *content words* più ricorrenti all'interno di ognuno dei quattro subcorpora (*sub1_lav, sub2_edu, sub3_app, sub4_rel*).

Mentre *sub1_lav*, *sub3_app* e *sub4_rel* mostrano in prima o seconda posizione i *word-types* che esplicitano l'area semantica dei singoli subcorpora, "scuola" (*f* = 114), attesa come forma primaria all'interno del subcorpus incentrato sul contesto educativo, si colloca soltanto al sesto posto per frequenza. Tra le *content words* più ricorrenti di *sub2_edu* si ripresentano termini quali "uso" e "usare", la cui compresenza con i *word-types* "scuola", "imparare", "compiti", "studenti", "studenti" e "studiare" declina le modalità di utilizzo dell'AI in ambito scolastico e, in misura apparentemente minore, in ambito accademico. Il termine "università", ad esempio, non si classifica tra le 30 parole di contenuto più frequenti. Ciononostante, in *sub1_edu* si distingue nuovamente il ruolo dei *Large-Language Models* ("llm") nel fornire supporto alla scrittura ("scrivere", "testo") e alla "ricerca", come nel caso della stesura di "tesi" di laurea.

In *sub1_lav*, "lavoro" (*f* = 319) si impone naturalmente come *word-type* più ricorrente, in linea con il tema del subcorpus. Il termine si ripresenta in modo significativo anche in *sub3_app* e *sub2_edu*. Nel primo caso, la sua frequenza potrebbe alludere alla varietà di utilizzo dell'AI sul posto di lavoro e al suo impatto sulla produttività dei dipendenti, mentre nel secondo caso si presta a una duplice interpretazione: la ricorrenza della forma "lavoro" all'interno di un subcorpus dedicato all'istruzione rimanda sia alle figure dei docenti e all'integrazione dell'AI nella didattica sia al naturale proseguo del percorso formativo degli studenti, proiettati verso un futuro professionale tanto sfidante quanto innovativo. Nonostante "chatgpt" compaia soltanto 82 volte, *sub1_lav* si distingue come unica raccolta di commenti in cui "intelligenza" (*f* = 93) e "artificiale" (*f* = 62) si presentano tra le 30 *content words* più frequenti. Si può dunque ipotizzare che, rispetto agli altri subcorpora, *sub1_lav* sia caratterizzato da una maggiore astrazione del discorso: l'AI non è discussa esclusivamente come assistente virtuale, bensì come fenomeno socio-economico. Difatti, osservandone le parole di contenuto, il subcorpus dedicato al lavoro guarda primariamente al sistema produttivo ("lavori", "soldi", "lavorare") e all'ambito aziendale ("azienda"), tematizza le preoccupazioni ("problema", "problemi") o le aspettative per il "futuro" e richiama uno specifico settore di impiego, quello informatico ("informatica", "dati").

Nel subcorpus dedicato all'antropomorfizzazione dell'AI e al suo impatto sul benessere psicologico e relazionale, "chatgpt" (*f* = 183) si posiziona come seconda forma più ricorrente. Il chatbot di OpenAI è difatti preceduto per frequenza dalle 197 occorrenze

della forma “persone”: discostandosi significativamente dagli altri subcorpora, *sub4_rel* delinea un passaggio tematico dall’artificiale all’umano, dall’oggetto tecnologico ai soggetti che lo utilizzano. Il subcorpus, caratterizzato dalle alte frequenze dei *word-types* “persona”, “amico”, “vita”, “psicologo” e “parlare”, intercetta una nuova area tematica all’interno del rapporto tra l’intelligenza artificiale e gli utenti che ne fruiscono: se *sub1_lav*, *sub2_edu* e *sub3_app* tematizzano le funzionalità pratiche dell’AI e le conseguenze che interessano l’ambiente esterno, *sub4_rel* introduce una dimensione introspettiva, incentrata sui soggetti e sulle loro esperienze sociali e relazionali. In particolare, le ricorrenze di “amico” ($f = 120$) e “psicologo” ($f = 113$) anticipano il ruolo di confidente che il chatbot è chiamato ad assumere.

3.2.2. TF-IDF: la forza discriminante delle parole

Un ulteriore approccio all’esplorazione e alla rappresentazione del vocabolario è l’utilizzo del *term frequency-inverse document frequency* (TF-IDF), una misura che va oltre l’analisi delle frequenze delle forme. Il TF-IDF combina difatti il concetto di frequenza relativa di una parola all’interno di un testo o subcorpus (*term frequency*) a quello di rarità della stessa lungo tutti i testi o subcorpora di cui si compone il corpus (*inverse document frequency*). Una parola che all’interno di un testo presenta un elevato valore di TF-IDF è una parola caratterizzante di quel testo: occorre in esso con frequenze elevate – dimostrandosi pertanto rilevante – e, al contempo, compare in un numero limitato di testi tra quelli disponibili. In altri termini, il TF-IDF consente di identificare “[…] le parole che possono essere considerate discriminanti per un testo, cioè apprezzabilmente più presenti in quel testo rispetto agli altri e, quindi, utili a caratterizzarlo” (Tuzzi 2024, p. 123).

Un peso TF-IDF pari a 0 identifica un termine assente nel subcorpus analizzato oppure un termine che, seppur presente, compare in tutti i documenti del corpus, perdendo dunque forza discriminante. Nel corpus costituito dall’insieme dei commenti dei redditors alle 28 submisson selezionate, le dieci forme più discriminanti (si veda a sinistra della *Tabella 8*) hanno un TF-IDF compreso tra 0.0002559 e 0.0001069 e sono: “incel”, “pa”, “reddito”, “her”, “empatia”, “solitudine”, “psicologi”, “amicizia”, “terapia” e “psicoterapeuta”. Considerati i quattro subcorpora per i quali i suddetti *word-types* risultato caratterizzanti, si nota come le forme maggiormente discriminanti all’interno del

corpus siano contenute in larga parte in *sub4_rel*. La tabella sottostante (si veda a destra della *Tabella 8*) riporta i primi cinque *word-types* con TF-IDF più alto per ognuno dei subcorpora. La forza discriminante di tali parole incoraggia un'analisi più approfondita del contesto testuale in cui esse si presentano, visualizzabile in *Tabella 9* tramite applicazione della *Key Word In Context analysis* o analisi delle concordanze (Fobbe 2026).

Sub	***Word-type***	***f***	***TF-IDF***
4	incel	27	0,0002559
1	pa	26	0,0002051
1	reddito	49	0,0001933
4	her	19	0,0001800
4	empatia	38	0,0001800
4	solitudine	18	0,0001706
4	psicologi	36	0,0001706
4	amicizia	18	0,0001706
4	terapia	32	0,0001516
4	psicoterapeuta	15	0,0001421
2	docenti	14	0,0001370
1	ubi	17	0,0001341
1	ricchi	33	0,0001302
4	psicoterapia	27	0,0001279
2	tesi	61	0,0001239
1	informatici	30	0,0001183
4	sesso	24	0,0001137
4	sentirti	24	0,0001137
2	scuole	23	0,0001125
3	noise	11	0,0001069

Word-type	***sub1_lav***	***sub2_edu***	***sub3_app***	***sub4_rel***
pa	0,0002051	0	0	0
reddito	0,0001933	0	0	0
ubi	0,0001341	0	0	0
ricchi	0,0001302	0	0	0
informatici	0,0001183	0	0	0
docenti	0	0,0001370	0	0
tesi	0	0,0001239	0	0
scuole	0	0,0001125	0	0
esami	0	0,0001015	0	0
sbobine	0	0,0000978	0	0
noise	0	0	0,0001069	0
ricette	0	0	0,0000972	0
skate	0	0	0,0000875	0
tutorial	0	0	0,0000777	0
totti	0	0	0,0000777	0
incel	0	0	0	0,0002559
her	0	0	0	0,0001800
empatia	0	0	0	0,0001800
solitudine	0	0	0	0,0001706
psicologi	0	0	0	0,0001706

Tabella 8 A sinistra, elenco di 10 *word-types* ordinati per TF-IDF decrescente nel corpus complessivo. A destra, *Term-Document Matrix* raffigurante i primi 5 *word-types* con TF-IDF più elevato per ognuno dei quattro subcorpora (*sub1_lav*, *sub2_edu*, *sub3_app*, *sub4_rel*).

Tra i termini caratteristici di *sub1_lav* si distinguono "pa" e "ubi", due acronimi che, inserendosi all'interno dell'ambito lavorativo, vengano utilizzati per fare riferimento alla Pubblica Amministrazione e all'*Universal Basic Income* (in italiano, Reddito di Base Universale). Il settore pubblico e la redistribuzione economica risultano pertanto tematiche distintive dei commenti online propri delle submission incentrate sul rapporto tra lavoro e intelligenza artificiale. I *word-types* "reddito" e "ricchi", rispettivamente

seconda e quarta parola per TF-IDF all'interno di *sub1_lav*, ne ribadiscono la significatività per il subcorpus.

"Docenti", "tesi", "scuole", "esami" e "sbobine" si mostrano fortemente coerenti con la macro-tematica di *sub2_edu*. In modo apparentemente paradossale, all'interno di una raccolta testuale orientata al sistema educativo, a detenere il maggior potere discriminante sono proprio le forme tipicamente appartenenti al suddetto campo semantico. Il dato suggerisce la bassa trasversalità del tema: il tema educativo è strettamente concentrato in *sub2_edu* e marginale o persino assente nei restanti subcorpora. In altri termini, quando il dibattito online sull'intelligenza artificiale si focalizza sul lavoro, sugli usi quotidiani o sulle relazioni interpersonali, si allude raramente all'integrazione dell'AI nei contesti scolastici e accademici e, ancor meno, a ciò che tale integrazione implica per la figura dei docenti.

Contrariamente, i *word-types* più discriminanti di *sub3_app* mostrano una bassa coerenza tematica, suggerendo una considerevole trasversalità del tema: l'ipotesi è che l'intelligenza artificiale sia ormai talmente pervasiva nella vita quotidiana che le discussioni sul suo utilizzo pratico si presentano con costanza all'interno delle *threaded online conversations*, indipendentemente dalla macro-tematica delle submission. Se, dunque, "scrivere", "testo" o "ricerca" non emergono come forme discriminanti perché distribuite lungo i quattro subcorpora, "noise", "ricette", "skate" e "tutorial" compaiono in questi ultimi in numero limitato o nullo ma occorrono frequentemente in *sub3_app*. I quattro *word-types* contraddistinguono pertanto i commenti a submission esplicitamente incentrate sull'uso applicativo dell'intelligenza artificiale, nelle quali si affermano modalità di utilizzo dell'AI meno ordinarie. La KWIC contestualizza l'uso dell'AI nella computer graphics (o grafica digitale) ("valori/valore di noise", "funzioni di noise"), per l'ideazione di "ricette" o in sostituzione di "tutorial" di fronte ad argomenti o attività pratiche di cui si ha una conoscenza limitata. La forma "skate" ne mostra invece l'uso peculiare come assistente per l'elaborazione di programmi di allenamento (in questo caso, su skateboard), emulando la figura del preparatore atletico. Emerge, infine, un utilizzo ludico della GenAI, di cui si sfrutta la capacità ricreativa per simulare conversazioni improbabili tra personaggi tra loro estranei ("totti").

L'ultima colonna in *Tabella 8* riporta i valori del TF-IDF calcolati per il subcorpus accentrato sulla macro-tematica delle relazioni interpersonali. In *sub4_rel*, i due termini

maggiormente caratteristici sono gli anglicismi "incel" e "her". Il primo – di cui si parlerà approfonditamente in seguito (si veda *§4.2.2*) – è una parola macedonia (dall'inglese, *involuntary celibates*) che designa una subcultura di uomini "celibi involontari" che biasima la società per la propria incapacità di instaurare relazioni intime o sentimentali (DeCook 2021). Oltre ad essere il più rappresentativo per *sub4_rel*, "incel" è il termine con il maggior peso TF-IDF all'interno del corpus complessivo (cfr. *Tabella 8*). La contestualizzazione del *word-type* "her", secondo per rappresentatività in *sub4_rel* e quarto nel corpus complessivo, consente di cogliere il diretto riferimento al film *Her* (2013), inteso come premonitore dell'attuale influenza dei chatbot di GenAI sulle relazioni umane. Ambientato in un futuro prossimo, *Her* narra la storia di un uomo solitario ed emotivamente vulnerabile che intraprende una relazione sentimentale con il sistema operativo del proprio computer, un software basato su un'intelligenza artificiale in grado di apprendere e simulare emozioni (Imm & Kang 2020). Osservando i segmenti di testo in cui occorre il *word-type* "her", è evidente il forte parallelismo che i redditors riconoscono tra il film e la realtà odierna. A seguire in ordine decrescente, i termini "empatia" e "solitudine" sembrano rintracciare i due principali fattori responsabili della crescente integrazione dell'AI nella vita privata e sentimentale: l'"empatia" che i chatbot sono in grado di simulare e la condizione di "solitudine" che favorisce il ricorso a tali strumenti. Alla luce di tali osservazioni, è possibile comprovare che il vocabolario di *sub4_rel* si contraddistingue dai restanti subcorpora per un lessico emozionale e per una significativa concentrazione di discussioni a carattere analitico. In particolare, i riferimenti alla categoria identitaria degli incel e al film *Her* sono indice di discussioni focalizzate, da un lato, sul risentimento maschile e, dall'altro, su una realtà che da finzionale diventa concreta.

In conclusione, il TF-IDF arricchisce l'interpretazione dei dati testuali e consente di identificare, all'interno di ciascuna raccolta, sfaccettature del discorso altrimenti passate inosservate. Come illustrato *in Tabella 7*, la classificazione per frequenze mette in risalto l'impatto dell'AI sul settore privato ("aziende") e sui metodi di apprendimento degli studenti ("studenti", "studente", "studiare"), il diffuso impiego dell'AI come supporto alla scrittura e la ricerca di un'alternativa artificiale alla figura reale dell'amico o dello psicologo. Per contro, la classificazione per TF-IDF segnala la presenza di ulteriori cornici discorsive: l'impatto dell'AI sul settore pubblico (*sub1_lav*) e sui metodi didattici

dei docenti (*sub2_edu*), la versatilità d'uso dell'AI, che spazia dalla grafica digitale all'intrattenimento (*sub3_app*), e le riflessioni critiche dei redditors sul modo in cui l'interazione tra umano e artificiale ridefinisce persino l'intimità (*sub4_rel*).

Subcorpus	***Word-type***	***Contesto testuale***
sub1_lav	pa	Girano ancora programmi del '92 e vogliamo parlare di IA nella PA.
		Il fatto è che la PA non vuole evolvere, perché realisticamente se ci fosse un'evoluzione gran parte dei dipendenti pubblici non servirebbero più, già ora.
		Di lavoro da fare nella PA ne abbiamo, anche solo digitalizzare documentazione cartacea o rendere omogenei dati e procedure, è tanto lavoro e richiede l'esperienza di chi lavora nella PA.
		Che poi, p.s., con le tecnologie attuali davvero pochi lavori PA possono essere "sostituiti" dall'AI, ma piuttosto resi efficienti, ovvero il lavoro che 10 dipendenti ti fanno ora lo può fare uno con un sistema AI affiancato.
	reddito	Il reddito universale è una chimera, uno specchietto per allodole per tenere parte di quelli preoccupati impegnati a discutere invece di protestare ora.
		Se anche la produttività restasse uguale con le "macchine", il problema, a mio avviso, resta il fatto che con un reddito molti resterebbero a casa a poltrire, e inevitabilmente molti servizi non verrebbero più usati.
		[…] uno strumento di inclusione sociale, sia esso il reddito di cittadinanza, la tassazione negativa oppure un universal basic income, è necessario, ma proprio per stimolare un'allocazione più efficiente delle risorse, permettendo alle persone di avere una safety net alle proprie spalle per ri-inventarsi e poter essere al passo con i tempi.
sub2_edu	docenti	Credo che le teste vuote siano legate al nozionismo e lassismo del sistema educativo corrente e alla mancanza di aggiornamento culturale dei docenti (che riciclano le verifiche di 30 anni prima e se ne fregano).
		Diciamo che se quello che si insegna, il modo di farlo ed il risultato sterile che si ottiene è manipolabile da uno strumento come l'AI generativa, il problema non sono gli studenti ma i docenti e la struttura obsoleta di insegnamento che c'è dietro.
		Dovremmo vederla sempre in questo modo, il problema non è l'AI, ma la scuola, che necessita di cambiare obiettivi e metodi: sia perché deve essere in grado di educare lo studente all'utilizzo di queste nuove tecnologie, sia perché i docenti dovrebbero essere formati sul come riconoscere un contenuto generato da una AI, ed evitare così che essa si utilizzi durante le verifiche in classe o nei compiti a casa.
	tesi	Per ultimo, difatti, voglio fare notare che diverse università, come testimoniato da studenti e laureati di Luiss o telematiche di mia conoscenza, utilizzano software anti-IA specifici per la disamina di elaborati, financo tesi di laurea, che devono dimostrare di non avere avuto un apporto da IA non oltre il 15% del totale, pena invalidamento del prodotto dello studente.
		Come hai detto tu, fa frasi sconclusionate, usa aggettivi troppo ricercati/inadatti in modo ridondante, non è in grado di scrivere in modo coerente un testo lungo e solo un pollo potrebbe non accorgersene, quindi usarlo seriamente per scrivere la tesi è grave.

		Io per la mia tesi magistrale l'ho usato, però soltanto per ricercare errori grammaticali o per rielaborare frasi che avevo difficoltà ad elaborare, secondo me se usato nel modo corretto è uno strumento utile e valido, ovviamente non per scrivere intere tesi.
sub3_app	noise	La domanda posta a ChatGPT è: "ponendo di calcolare più valori noise, come li combino mantenendo la correttezza delle derivate?". Come vedi ti dà una risposta, ma è sbagliata. Sono io che lo devo "imboccare".
		L'idea è usare queste funzioni di noise per descrivere un paesaggio, banalmente l'elevazione del terreno.
		Per dire, la generazione dei terreni è solo una delle applicazioni del Perlin Noise, ma dai per scontato cosa ci vuoi fare e ti aspetti che lo capisca.
	ricette	Vari usi: […] aiuto con cambiare ricette o pensare ad idee per cucinare.
		Io ci ho provato, ma mi tira fuori solo ricette orride dove mi dice di buttare tutti i singoli ingredienti che gli ho detto di avere in padella.
		Sì, è molto utile, ad esempio io ho molti problemi con il cibo (tipo al ristorante prendo il menù per bambini), quindi le chiedo sempre delle ricette innovative o dei consigli su come preparare cibo in un determinato modo per soddisfare il mio palato viziato […].
sub4_rel	incel	Negli sfoghi degli incel e forse anche nel mio post il sesso emerge come il tema principale perché è il bisogno fisico più immediato e di cui si avverte più dolorosamente la deprivazione, soprattutto nell'adolescenza e nella prima giovinezza in cui si hanno gli ormoni a mille.
		In pratica non è che concettualizzano meglio l'amore, è che sono immuni agli atteggiamenti sbagliati degli incel, per cui possono essere una fonte di validazione continua. Ma proprio perché sono un oggetto che fa il suo lavoro, non una persona che dovrebbe ricavare del piacere e una effettiva connessione con l'altro.
		Non ho un obiettivo preciso, ho fatto tutto spontaneamente, tra un tentativo e l'altro di alleviare la sofferenza legata alla condizione di incel, e per caso ho trovato una "combinazione" che mi ha tolto le castagne dal fuoco. Quelle caratteristiche dei robot che citi io le vedo come punti a favore, perché ti danno i lati positivi dell'esperienza di relazione senza quelli negativi, senza tutto lo sbatti di scendere a compromessi con un'altra persona che ha esigenze diverse dalle tue e può abbandonarti da un momento all'altro.
	her	Il film Her ha beccato il futuro con una precisione talmente irreale che mi viene quasi da chiamare il complotto
		In sostanza hai replicato il film Her. Se tu sei contento così buon per te, ma a me sembra tanto un sostituire una situazione problematica con una finzione ben impacchettata... più che una soluzione al problema mi sembra un palliativo.
		Il film Her è un esempio perfetto di come queste tecnologie possano evolversi e influenzare le relazioni umane. La capacità di una IA di simulare empatia e instaurare un rapporto "personale" può certamente portare a situazioni dove le persone più vulnerabili potrebbero iniziare a preferire la compagnia di un assistente virtuale a quella umana.

Tabella 9 KWIC dei tre *word-types* con il maggior peso TF-IDF per ciascun subcorpus.

4. Il raggruppamento dei testi non supervisionato

4.1. *Stylo*: analogie e differenze nello stile di scrittura dei commenti online

La prima procedura di clusterizzazione eseguita sul corpus dei commenti online ha lo scopo di esplorare le somiglianze e differenze stilistiche presenti tra i quattro subcorpora tematici definiti a priori. L'analisi stilometrica è difatti finalizzata a organizzare i testi in gruppi sulla base della presenza di tratti linguistici comuni. Nel presente lavoro, tali procedure di raggruppamento sono tradotte graficamente da *stylo*, un pacchetto R frequentemente utilizzato negli studi di stilometria per l'analisi statistica dei dati testuali.

Come anticipato, la somiglianza lessicale tra testi è comunemente valutata tramite metodi *distance-based* ed utilizzando, come elemento testuale alla base dell'analisi, le *n* parole più frequenti (MFW) all'interno del corpus. Pel determinare le strutture di relazione tra i quattro gruppi di commenti, il numero di MFW con cui si è scelto di condurre la clusterizzazione rispetta la soglia di frequenza stabilita al fine di garantire la rappresentatività dell'analisi. Il dendogramma in *Figura 12* e in *Figura 13* riunisce dunque il corpus in gruppi di testi omogenei a partire dalle frequenze relative delle 2963 parole più ricorrenti, vale a dire i *word-types* con una frequenza maggiore o uguale a 8. La somiglianza stilistca tra testi è invece misurata dalla distanza Cosine Delta (anche definita Wurzburg Distance), una misura di similarità che si rivela particolarmente efficace nel caso di analisi condotte a partire da un numero elevato di parole, superiore alle 1500 MFW (Evert et al. 2017).

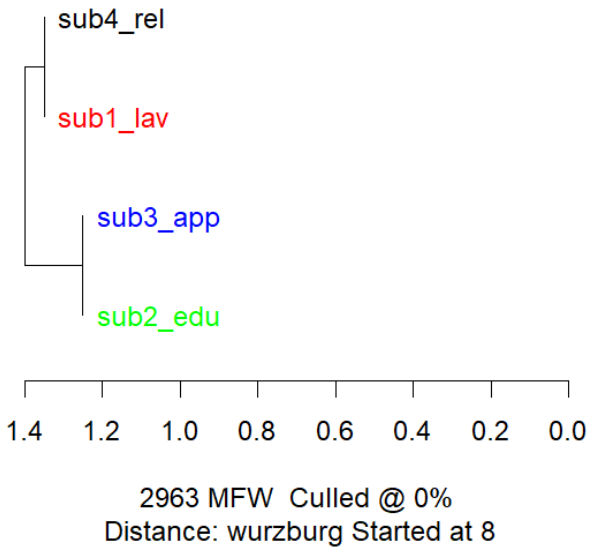

Figura 12 Dendogramma generato tramite *stylo* considerando i subcorpora come singole unità testuali.

In assenza di *sampling* (Eder et al. 2016), ciascuno dei quattro documenti di cui si compone il corpus viene trattato come un'unica unità testuale. Una prima clusterizzazione consente pertanto di stabilire che il corpus è suddivisibile in due gruppi di testi omogenei (cluster): *sub2_edu* e *sub3_app*, uniti dallo stesso ramo, presentano una maggiore somiglianza lessicale e si discostano da *sub1_lav* e *sub4_rel*, che a loro volta appaiono lessicalmente simili. Nell'interpretazione del dendogramma, è necessario tenere a mente che una maggiore lunghezza dei rami antecedente alla loro unione (*i.e.,* i rami si uniscono a una distanza inferiore) segnala una maggiore dissimilarità lessicale tra i cluster: *sub2_edu* e *sub3_app* condividono caratteristiche lessicali più simili rispetto a quelle condivise da *sub1_lav* e *sub4_rel*.

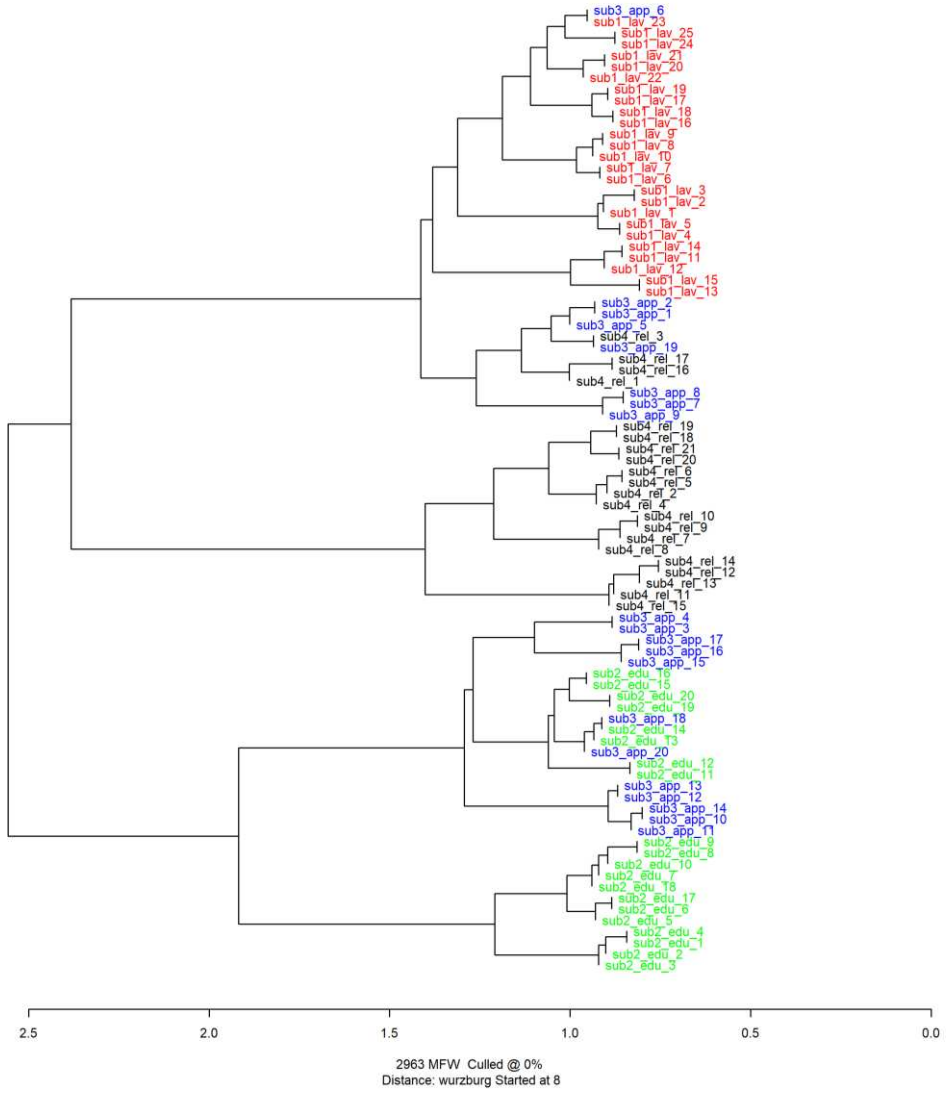

Figura 13 Dendogramma generato tramite *stylo* adottando un *normal sampling* di 3000 parole per segmento.

Al fine di rendere il clustering più informativo, i subcorpora sono stati segmentati in 86 campioni di eguale dimensione, ossia in blocchi costituiti da 3000 parole[65] così suddivisi: 25 segmenti per *sub1_lav*, 20 segmenti per *sub2_edu*, 20 segmenti per *sub3_app* e 21 segmenti per *sub4_rel*. L'adozione di un procedimento di *normal sampling* offre un'esplorazione più dettagliata dello stile linguistico dei subcorpora: mentre il dendogramma in *Figura 12* mostra esclusivamente la distanza lessicale tra i quattro subcorpora, quello in *Figura 13* ne analizza la struttura interna, identificando cluster nidificati e rendendo visibili le relazioni di somiglianza a un livello più profondo. Esaminando 86 distinte unità lessicali, l'algoritmo agglomerativo alla base del clustering stilometrico risulta più evidente: i singoli blocchi di testo sono uniti prima a coppie e poi a gruppi in base al loro grado di somiglianza lessicale, restituendo una rappresentazione grafica dell'omgeneità (o eterogeneità) stilistica di ciascuno dei quattro subcorpora. L'osservazione delle macro-biforcazioni della struttura gerarchica consente di distinguere, ancora una volta, le relazioni di somiglianza sottese, da un lato, tra i commenti sul lavoro e le relazioni sociali e, dall'altro, tra i commenti sull'istruzione e le applicazioni dell'intelligenza artificiale. Le foglie dei rami alla destra del grafico restituiscono tre risultati degni di nota. In primo luogo, collocandosi interamente nella parte superiore del dendogramma, i 25 segmenti in cui viene suddiviso *sub1_lav* mostrano la coerenza lessicale del subcorpus relativo all'integrazione dell'AI nell'ambito lavorativo, che si presenta come il più omogeneo dei subcopora. In secondo luogo, *sub2_edu* e *sub4_rel* risultano meno compatti, pur mantenendo un grado elevato di omogeneità lessicale. I commenti dei redditors sulla crescente pervasività dei modelli AI nelle relazioni sociali e nell'ambito educativo mostrano, talvolta, uno stile comune a *sub3_app*, che si rivela, infine, il subcorpus linguisticamente più disperso. I 20 segmenti in cui è suddiviso il subcorpus si distribuiscono equamente in quasi tutti i sottogruppi definiti dai rami intermedi, un risultato che si fa indice dell'eterogeneità delle discussioni dedicate ai molteplici usi pratici dell'intelligenza artificiale.

Il clustering stilometrico ricopre pertanto una doppia funzione esplorativa. Esso consente, innanzitutto, di visualizzare la misura in cui i diversi argomenti relativi all'AI

[65] In *Does size matter? Authorship attribution, small samples, big problem* (2015), Maciej Eder discute una serie di esperimenti finalizzati a individuare la lunghezza minima che i segmenti di testo devono possedere affinché l'analisi stilometrica produca risultati stabili: segmenti troppo brevi generano un eccessivo rumore statistico, mentre segmenti più lunghi, caratterizzati da una lunghezza minima si colloca tra le 2500 e le 5000 parole, producono risultati più affidabili.

vengono trattati dai redditors in maniera stilisticamente simile o divergente, confrontando da un punto di vista lessicale i commenti online relativi alle quattro macro-tematiche. In particolare, l'adozione del *normal sampling* consente di valutarne l'omogeneità interna e dunque di stabilire se la classificazione stilistica trovi riscontro nella suddivisione tematica *ex ante*. In secondo luogo, inserendosi in un insieme più ampio di analisi interpretate in ottica comparativa, il clustering stilometrico costituisce una prima elaborazione fondamentale, confrontabile con risultati derivanti dal clustering tematico-contenutistico e dalle successive analisi delle corrispondenze.

4.2 Il metodo Reinert: 5 classi semantiche e 4 subcorpora tematici

L'esecuzione del *text clustering* tramite IRaMuTeQ ha previsto una fase sperimentale e comparativa di segmentazione del testo. Come anticipato, il software permette la suddivisione del corpus in segmenti di testo (TS) definiti. Applicando la *Descending Hierarchical Classification* (DHC), il metodo di classificazione descritto da Reinert (1990)[66], "[t]he TS are clustered according to their vocabularies and distributed according to the reduced forms frequencies" (Camargo & Justo 2021, p. 11). Su IRaMuTeQ, la dimensione dei segmenti è intesa come il numero *n* di unità testuali contenute in ciascun segmento ed è impostata di default a 40 occorrenze: in un'analisi standard, condotta su testi tradizionali (e.g.., testi narrativi, articoli di giornale, discorsi politici), ogni TS è costituito da una sequenza di 40 parole. Si tratta tuttavia di un valore di riferimento, un parametro indicativo per il ricercatore, il quale può definire la dimensione dei segmenti a sua discrezione (Retinaud 2015), vale a dire sulla base della struttura del proprio corpus e della tipologia e lunghezza dei testi analizzati. In presenza di un corpus costituito da testi molto brevi è ad esempio consigliabile ridurre il valore di segmentazione predefinito dal software o trattare ciascun testo come un unico TS, come nel caso di tweet o di risposte brevi a domande aperte dei questionari (Moreno & Retinaud 2022; Camargo & Justo 2021).

Al fine di individuare la dimensione di segmentazione adeguata all'analisi del corpus complessivo, quest'ultimo, previa lemmatizzazione, è stato dapprima testato su due

[66] "[i] segmenti di testo vengono raggruppati in base al loro vocabolario e distribuiti secondo le frequenze delle forme ridotte".

valori: 23 e 40 occorrenze, corrispondenti – rispettivamente – alla lunghezza media di una frase lungo i quattro subcorpora (cfr. *Tabella* 3) e al numero predefinito di occorrenze, suggerito dal software e regolarmente impiegato in letteratura. Nel primo caso (si veda *Figura 14*), l'applicazione del metodo Reinert produce un numero limitato di classi, perlopiù sbilanciate: la classe 1, con una classificazione del 47,1% dei segmenti, domina sulla classe 2 (27,6%) e sulla classe 3 (25,2%). L'adozione di 23 occorrenze come valore di segmentazione permette dunque di individuare le tre tematiche principali del corpus in esame, ma richiede di sacrificare le sfumature presenti al loro interno, non individuate a causa della brevità dei segmenti. Per contro, il valore predefinito di 40 occorrenze (si veda *Figura 16*) consente di cogliere la complessità interna del discorso sull'AI individuando un numero maggiore di classi semantiche, tra le quali vige inoltre una distribuzione più equilibrata di TS: la percentuale dei segmenti di testo assegnati a ciascuno dei cinque cluster oscilla tra l'11,33% e il 31,1%. Confrontando i risultati della DHC condotta su 23 (*Figura 14*) e 40 (*Figura 16*) occorrenze per segmento, è possibile concludere che, pur raggruppando commenti online, tipicamente brevi e frammentati, il corpus oggetto della presente ricerca presenta caratteristiche che non lo rendono adatto a bassi valori di segmentazione. Tali caratteristiche sono riconducibili all'ampia variabilità della lunghezza dei testi che lo compongono e alla sua organizzazione strutturale: in primo luogo, all'interno delle sequenze dialogiche della piattaforma, commenti brevi si alternano a commenti più estesi e discorsivi, propri di un forum di discussione; in secondo luogo, in fase di acquisizione i commenti sono stati aggregati per submission, con variabili definite a monte per ogni blocco. L'unità testuale di analisi non è dunque costituita dal singolo commento, ma dall'insieme dei commenti associati a ciascuna submission. Per questo motivo, un terzo test di segmentazione è stato condotto su un valore pari a 100 occorrenze per segmento (si veda *Figura 15*), ipotizzato idoneo a ottenere classi semantiche maggiormente stabili e interpretabili a fronte della lunghezza complessiva dei blocchi di commenti.

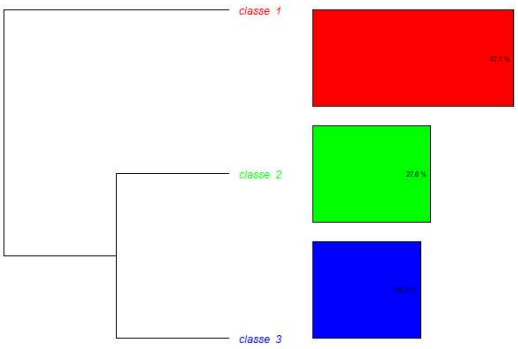

Figura 14 Dendogramma risultante dalla suddivisione del corpus in segmenti costituiti da 23 occorrenze.

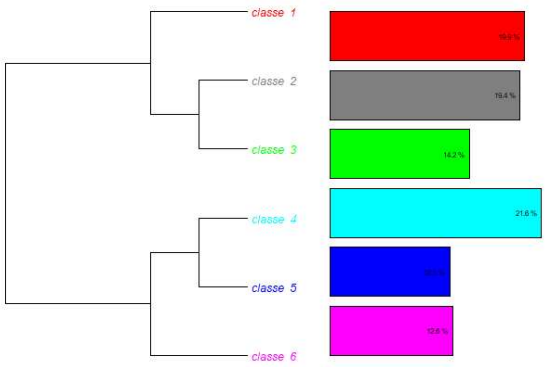

Figura 15 Dendogramma risultante dalla suddivisione del corpus in segmenti costituiti da 100 occorrenze.

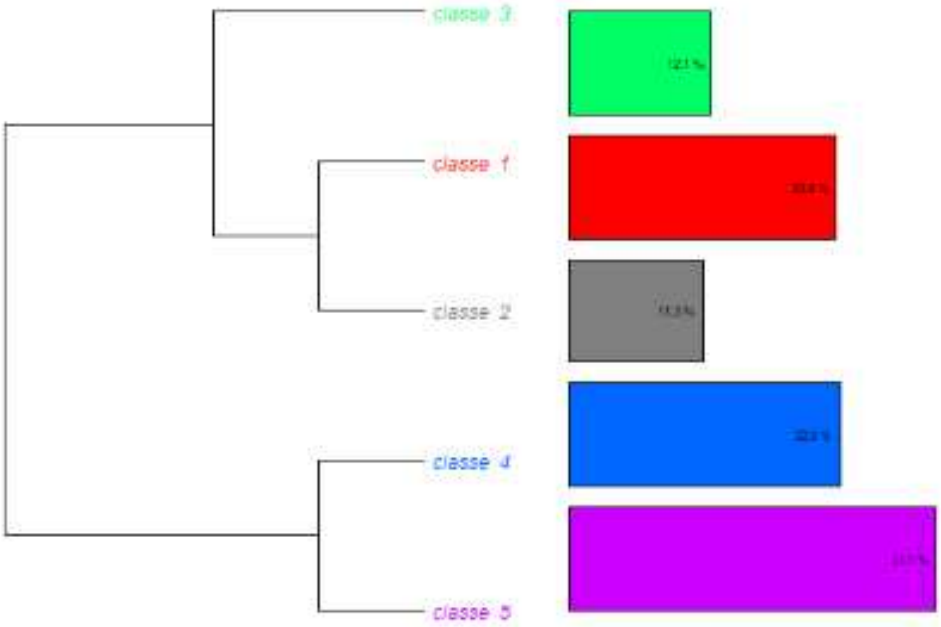

Figura 16 Dendogramma risultante dalla suddivisione del corpus in segmenti costituiti da 40 occorrenze.

Comparato alla *Figura 16*, il dendogramma in *Figura 15* non presenta differenze sostanziali in termini di frammentazione: un valore di segmentazione pari a 100 occorrenze produce sei cluster, uno in più rispetto alla classificazione condotta con 40 occorrenze per segmento. Tuttavia, un'osservazione qualitativa delle classi sancisce quest'ultimo valore – il numero di occorrenze predefinito da IRaMuTeQ – come il più

adatto per il corpus in esame, concludendo questa preliminare fase sperimentale[67], cruciale per una corretta clusterizzazione del corpus. La *Discending Hierarchical Classification* su 40 occorrenze produce difatti clusters più coerenti, compatti e facilmente interpretabili: nell'analisi dei commenti Reddit, segmenti da 40 occorrenze massimizzano le differenze lessicali e rivelano nuclei tematici in netto contrasto, evitando quell'aggregazione di temi eterogenei che è invece causata dall'adozione di segmenti più lunghi (*i.e.* 100 occorrenze).

Nello specifico, applicando il metodo Reinert, il corpus costituito da 28 testi (submission) viene suddiviso in 7.312 segmenti, organizzati in base al loro contenuto in cinque classi semantiche o cluster (si veda *Figura 16*). Con un valore di segmentazione pari a 40 occorrenze per segmento, la percentuale dei segmenti analizzati e classificati è pari al 98,85% (7.228 su 7.312). La DHC ha dapprima suddiviso a metà il corpus in esame, producendo due insiemi di cluster (prima divisione o iterazione). Una seconda scissione ha successivamente dato origine alle classi 4 e 5, unite dallo stesso ramo e contenenti rispettivamente il 22,9% e il 31,1% dei segmenti di testo, e alla classe 3, con il 12,06% dei segmenti di testo. In un terzo passaggio, l'applicazione del metodo Reinert ha generato un sottoinsieme del cluster 3, dal cui ramo hanno pertanto avuto origine la classe 1 (22,61%) e la classe 2 (11,33%), che hanno concluso la classificazione gerarchica.

Il dendogramma in *Figura 16* permette di osservare con chiarezza le relazioni di appartenenza tra i cinque cluster: l'appartenenza della classe 3 al medesimo ramo del sottoinsieme costituito dai sub-cluster 1 e 2 ne segnala la somiglianza tematica, la quale si accentua particolarmente tra queste due ultime classi poiché unite da un ulteriore legame interno. Analogamente, anche le classi 4 e 5 risultano semanticamente molto simili tra loro. Il grafico sottostante (si veda *Figura 17*) offre una rappresentazione alternativa della DHC, presentando per le cinque classi semantiche un elenco di parole caratterizzanti e dunque maggiormente associate a ciascuna di esse. All'interno di ogni elenco, le parole sono difatti ordinate per valore del chi-quadrato, che misura la forza di

[67] In linea con le indicazioni di Camargo & Justo (2021), che per gli studi mirati al contenuto (*text contents*) propongono di intervenire sul riconoscimento degli elementi linguistici (parole attive, supplementari o eliminate), in ciascuna delle tre analisi sono considerate forme attive gli aggettivi, i nomi comuni, i verbi e le forme non riconosciute. Nomi propri e verbi modali sono invece classificati come forme supplementari, mentre altri elementi grammaticali – quali aggettivi dimostrativi o possessivi, verbi ausiliari, preposizioni, etc. – sono eliminati dall'analisi.

associazione tra una determinata forma lessicale e il cluster di appartenenza rispetto al resto del corpus (Arditi et al. 2020). La *Figura 17* riporta dunque le parole maggiormente distintive di ciascuna classe. Ad esempio, il lemma "lavorio|lavoro", primo per valore di chi-quadrato ($\chi^2 = 215{,}79$) nel cluster 5, costituisce la parola più distintiva all'interno di tale classe, poiché appare in essa con più frequenza rispetto alle altre classi. È, in breve, sovra-rappresentata nel cluster 5 rispetto alla distribuzione attesa nel corpus complessivo. Si potrebbe dunque concludere che la forma "lavorio|lavoro", accompagnata da "azienda", "lavorare", "tecnologico", "futuro" e "soldo" (con χ^2 compreso tra 203,31 e 116,21), contribuisce significativamente alla definizione del profilo lessicale del cluster, il quale viene con facilità definito, da un punto di vista interpretativo, il cluster dell'ambito lavorativo.

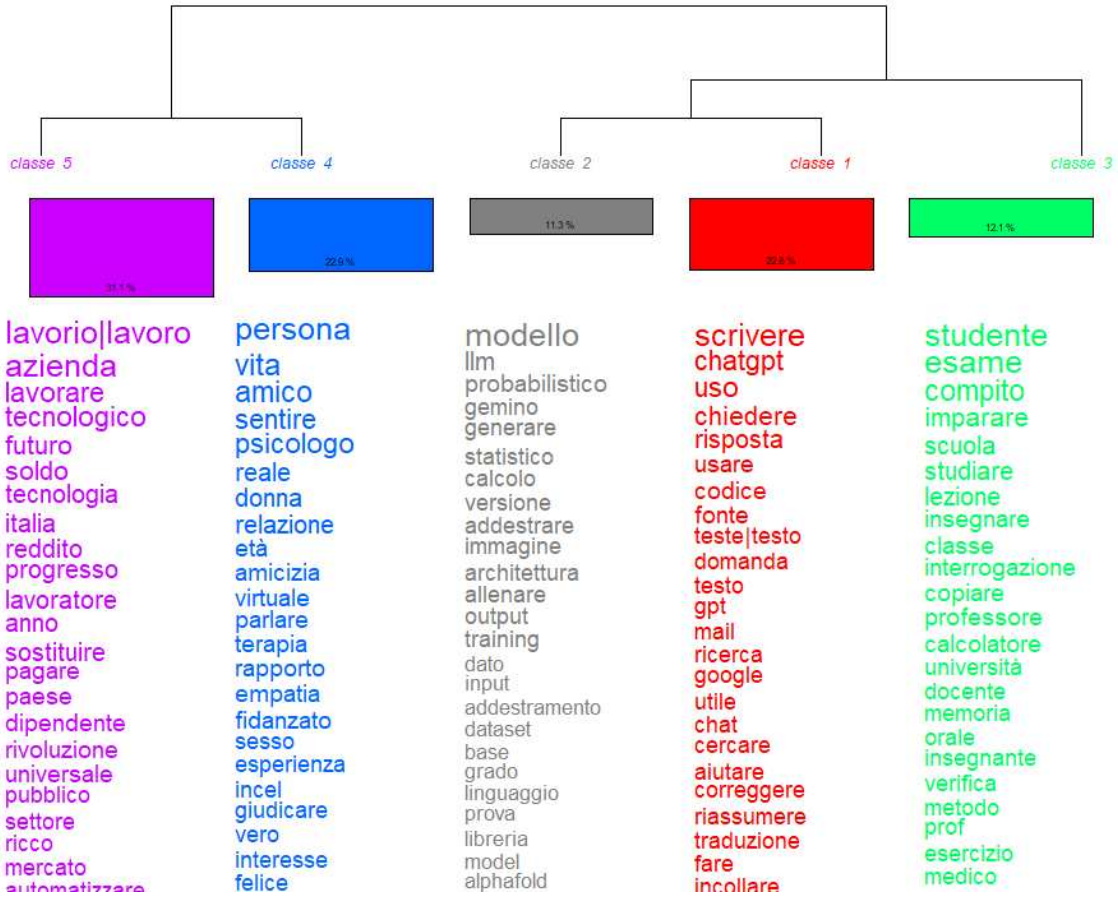

Figura 17 Anteprima dei profili lessicali delle cinque classi semantiche emerse dalla DHC del corpus complessivo.

A tal proposito, l'osservazione del vocabolario di ciascuna classe semantica rende subito evidente la forte corrispondenza tra i cluster e i subcorpora tematici, decretando il valore confermativo dell'analisi (Bernardi & Campostrini 2005): pur discutendo un'analisi di tipo esplorativo, la *Descending Hierarchical Classification* contribuisce nella presente ricerca a verificare la coerenza della classificazione tematica operata *ex ante*. In altri termini, la clusterizzazione del corpus prodotta da IRaMuTeQ corrobora in gran parte la suddivisione manuale delle submission nei quattro subcorpora. Come descritto nella fase di costruzione del corpus (cfr. *§3.1*), la lettura esplorativa dei post ha condotto all'identificazione di quattro macro-tematiche all'interno del discorso sull'intelligenza artificiale. Queste, nella DHC prodotta dal software, trovano corrispondenza con i cluster 5, 4, 3 e 1.

La classe 5, che detiene il maggior peso all'interno del corpus, è sovrapponibile a *sub1_lav*, il subcorpus costituito dai commenti di submission relative all'AI in ambito lavorativo, come indicato dai lemmi sopracitati. A seguire, la classe 4 mostra evidenti rimandi ai commenti inerenti all'impiego dell'AI come terapeuta o strumento di ascolto e sfogo, appurando la corrispondenza semantica con *sub4_rel*. A destra del grafico, il cluster 3, anch'esso originatosi dalla prima divisione del metodo Reinert, riguarda l'utilizzo dell'AI in ambito educativo, scolastico e universitario: la classe è pertanto riconducibile a *sub2_edu*. Le classi 1 e 2, infine, condividono lo stesso ramo, e presentano infatti un certo grado di vicinanza semantica. Mentre gli altri cluster si presentano come una cornice tematica, delineando i principali ambiti contrassegnati dall'impatto dell'AI, i cluster più interni realizzati nell'ultimo passaggio della *Descending Hierarchical Classification* si fanno rappresentativi del soggetto impattante, il tema protagonista del corpus: l'intelligenza artificiale. Entrambi i sub-cluster trattano l'AI in via diretta: da un lato, la classe 1 ne approfondisce le applicazioni nel quotidiano, collocandosi in prossimità di *sub3_app*; dall'altro, la classe 2 ne riassume gli aspetti tecnici e di funzionamento.

Proprio quest'ultima rappresenta il dato più significativo della classificazione. Come anticipatamente precisato, l'efficacia del *text clustering* risiede anche nella capacità di rivelare, in un'analisi complessiva del corpus, aree tematiche aggiuntive rispetto a quelle già identificate, avvalorando il suo ruolo nel far emergere pattern latenti. I risultati della DHC ne offrono una dimostrazione concreta: se, al pari delle classi precedenti, la classe

1 trova piena corrispondenza con uno dei subcorpus definiti a priori (*sub3_app*), la classe 2 rappresenta una nuova classe semantica, un'area tematica aggiuntiva passata inosservata al *close reading* delle submission pubblicate sulla piattaforma. Il *distant reading* ha così permesso di identificare il nucleo tematico di presentazione dell'AI generativa. Appuratane la pervasività, la letteratura recente propone di ascrivere i modelli di intelligenza artificiale generativa alle già menzionate *General-Purpose Technologies* (GPT): questi, poiché addestrati ("addestrare", "addestramento", "allenare", "training") su enormi volumi di dati ("dato", "dataset"), sono facilmente adattabili a molte applicazioni a basso costo e favoriscono l'innovazione in molteplici settori. Rientrano tra le tipologie di GenAI i *Large-Language Models* (LLM), modelli probabilistici in grado di "generare" testi stimando, sulla base di calcoli statistici ("calcolo", "statistico"), la probabilità che una determinata unità testuale (*token*) segua la precedente sequenza di *token*. I segmenti di testo tipici del cluster 2 permettono di osservare come all'interno dei subreddit si discuta, ad esempio, la tendenza a sovrastimare le capacità dei LLM, i quali, proprio in quanto modelli di linguaggio addestrati su rappresentazioni linguistiche, diventano inaffidabili con calcoli complessi o con le richieste di generazione di codice. Assieme all'acronimo "llm", i termini "modello" e "probabilistico" rappresentano i lemmi più caratterizzanti del cluster 2, con un valore del chi-quadrato compreso tra 542,58 e 248,34. All'interno del profilo lessicale della classe, Gemini (erroneamente lemmatizzato dal software in "gemino"), sviluppato da Google, costituisce il principale esempio di LLM. ChatGPT, l'LLM sviluppato da OpenAI, compare invece come secondo lemma più rappresentativo del cluster 1.

Il filone discorsivo incentrato sugli aspetti tecnici e funzionali dell'AI generativa si ripresenta, con maggiore o minore incidenza, in ciascuno dei subcorpora analizzati: nelle discussioni relative all'ambito lavorativo (micro-cluster 1 in *Figura 20*), relazionale (micro-cluster 3 in *Figura 23*), applicativo (micro-cluster 2 in *Figura 27*) ed educativo (micro-cluster 1 e 2 *Figura 30*). È opportuno precisare che, pur costituendo il focus tematico della classe 2 e, più in generale, dell'intero corpus oggetto di analisi, la locuzione composta dai lemmi "intelligenza" e "artificiale" non risulta rappresentativa di alcuno dei macro-cluster, poiché pervasiva in ciascuno di essi.

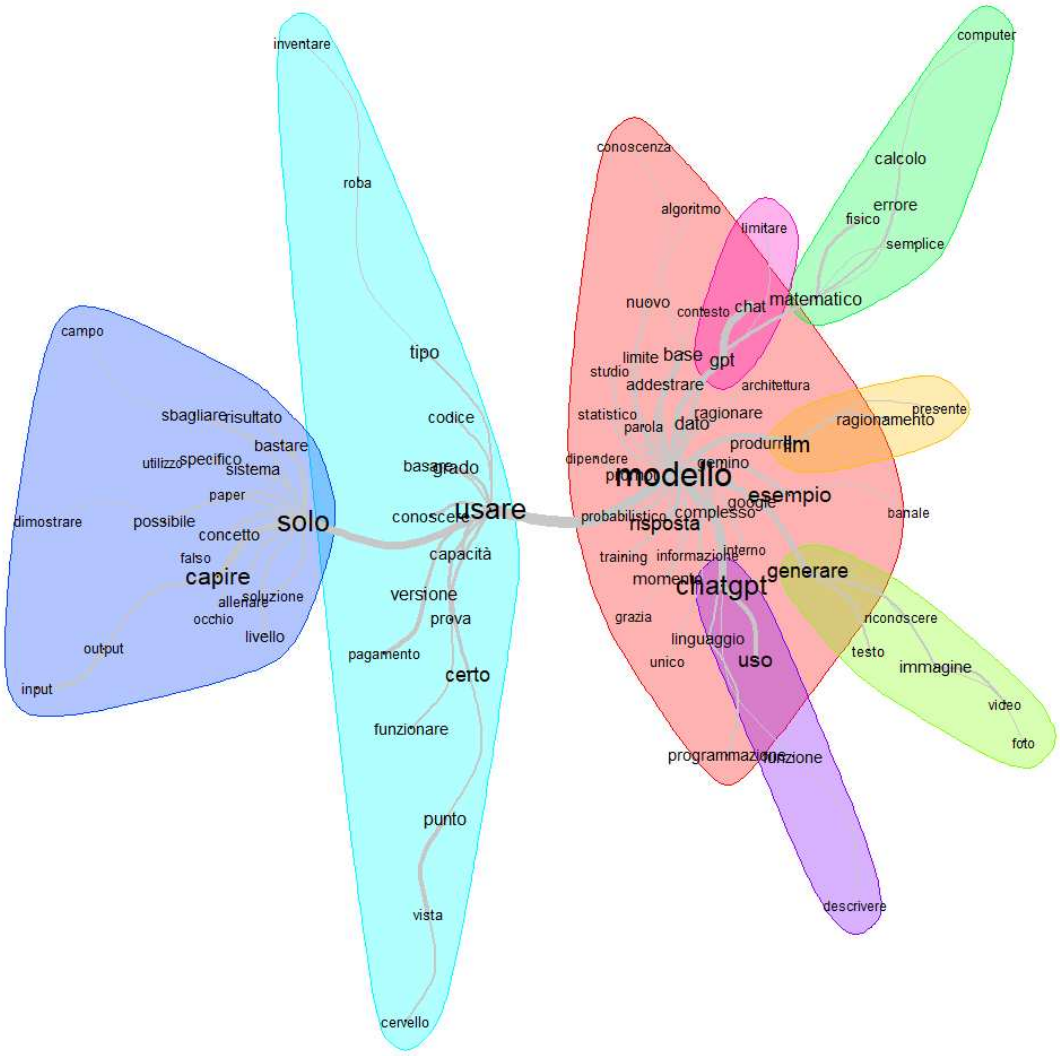

Figura 18 Grafico del cluster 2, identificato come classe semantica degli aspetti tecnici e computazionali dell'AI. Si includono i lemmi con una frequenza maggiore o uguale a 15. Le linee di collegamento tra i nodi illustrano la frequenza con cui i lemmi co-occorrono: maggiore è lo spessore della linea, maggiore è l'associazione tra i lemmi.

Le seguenti sezioni si occuperanno di esaminare i principali nuclei tematici del dibattito online. Seguendo un approccio *top-down*, si esploreranno brevemente gli ulteriori cluster emersi dall'analisi del corpus complessivo (anche definiti, per comodità espositiva, macro-cluster) e, mantenendo un'ottica confermativa, sarà successivamente condotta un'analisi secondaria su ciascuno dei subcorpora tematici di riferimento (da cui emergeranno i cosiddetti micro-cluster). L'applicazione del metodo Reinert su *sub1_lav*, *sub2_edu*, *sub3_app* e *sub4_rel* consente di ottenere un riquadro più dettagliato e

circoscritto dei frame di discussione all'interno di ogni nucleo tematico. L'interpretazione dei dati segue l'ordine di produzione dei macro-cluster ed è accompagnata da una loro rappresentazione grafica: nelle *Figure 18*, *19*, *22*, *26* e *29*, i lemmi[68] del cluster di riferimento sono raggruppati in nodi sulla base delle loro co-occorrenze nei segmenti di testo associati, delineando dei potenziali, ulteriori nuclei semantici all'interno del cluster analizzato.

4.2.1 Lavoro: le ansie occupazionali, la ridefinizione della società e il confronto tra intelligenza artificiale e umana

Tra i primi ad essere generati dalla DHC esplorativa, il cluster 5 detiene il maggior peso all'interno del corpus: la percentuale dei segmenti di testo ad esso assegnati, pari al 31,1%, suggerisce la preminenza del tema del lavoro all'interno del discorso pubblico sull'intelligenza artificiale. Il cluster 5 risulta difatti semanticamente allineato a *sub1_lav*, il subcorpus definito a priori per tematizzare l'impiego dell'AI in ambito lavorativo e, più in generale, ciò che la sua introduzione comporta nel contesto professionale: come già constatato, "lavorio|lavoro" ($\chi^2 = 215.79$) costituisce il lemma più distintivo del cluster, seguito, tra i lemmi più caratterizzanti, dalla sua forma verbale "lavorare" ($\chi^2 = 149.37$). Ciononostante, il suo profilo lessicale rivela una struttura ibrida le cui forme differiscono per tematizzazione diretta o indiretta del lavoro: ai lemmi "azienda", "soldo", "reddito", "lavoratore" e "dipendente", strettamente riconducibili all'ambito occupazionale, si alternano lemmi quali "tecnologico" e "tecnologia", "futuro" e "progresso", "italia" e "paese". Queste ultime forme, dalla forte connotazione trasformativa, evocano invece il cambiamento – già in atto o atteso – associato alla nuova "rivoluzione", e lo collocano all'interno di un confronto internazionale. L'attenzione è ridiretta dal "settore" economico alla dimensione sociale: tra i redditors, l'intelligenza artificiale è discussa in quanto fattore di ridefinizione del lavoro e, per estensione, della società nel suo complesso.

[68] Al fine di favorire la leggibilità dei grafici, per ogni cluster sono stati selezionati tutti i lemmi con una frequenza maggiore o uguale a 15. Si escludono, inoltre, il verbo modale "potere" e i verbi di supporto "fare" e "dare". Le linee di collegamento tra i nodi illustrano la frequenza con cui i lemmi co-occorrono: maggiore è lo spessore della linea, maggiore è l'associazione tra i lemmi.

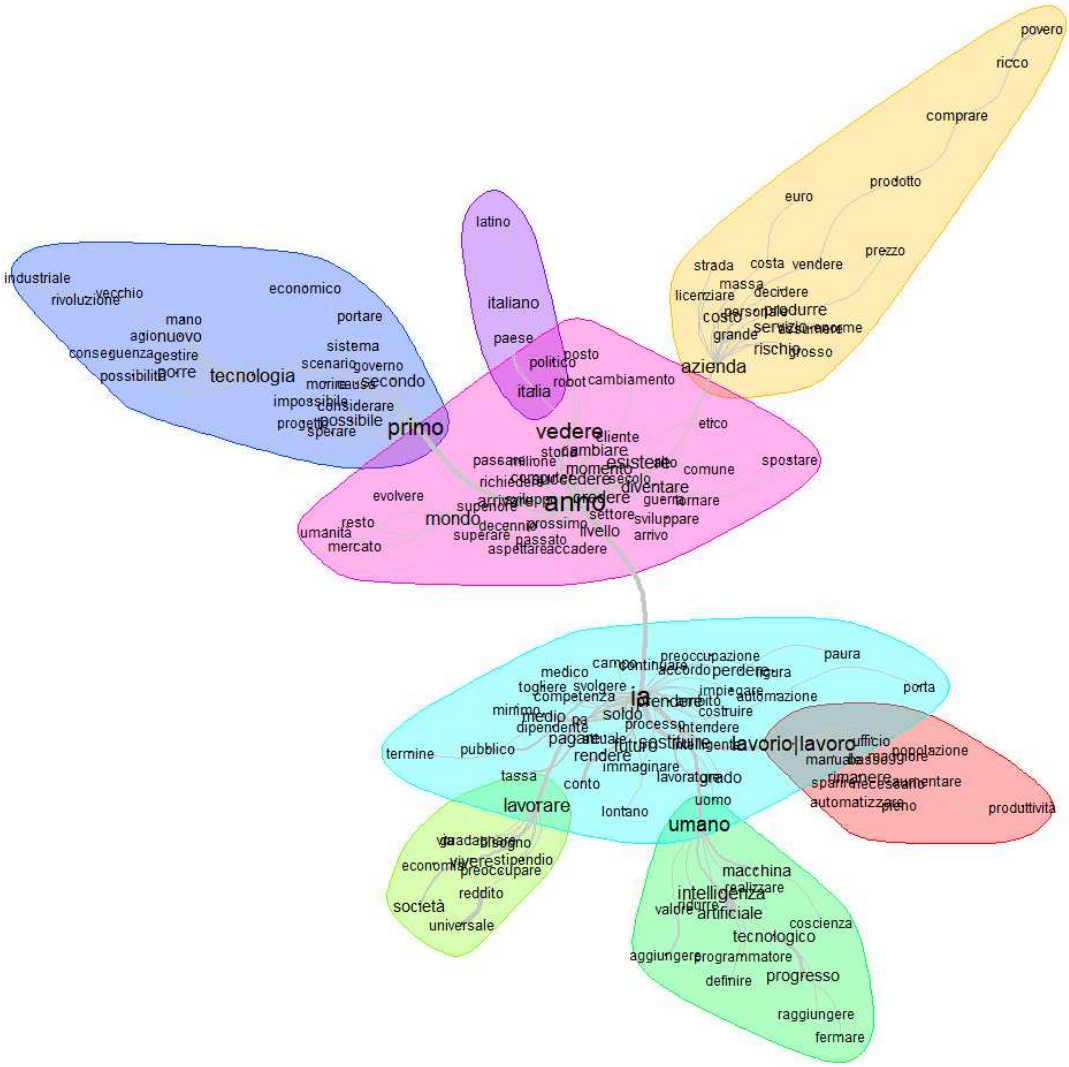

Figura 19 Grafico del macro-cluster 5, identificato come classe semantica del lavoro. Si includono i lemmi con una frequenza maggiore o uguale a 15. Le linee di collegamento tra i nodi illustrano la frequenza con cui i lemmi co-occorrono: maggiore è lo spessore della linea, maggiore è l'associazione tra i lemmi.

Questa duplice tematizzazione è avvalorata dall'analisi condotta sul subcorpus semanticamente corrispondente al macro-cluster 5. Il metodo Reinert applicato a *sub1_lav* produce quattro classi semantiche, organizzate in due coppie appartenenti allo stesso livello della classificazione gerarchica. Alla destra del grafico (si veda *Figura 20*), il cluster 3, che tratta il lavoro come ambito occupazionale, si lega al cluster 2, che lo rappresenta invece come oggetto di trasformazione sociale. Nel primo caso, "dipendente", il lemma maggiormente associato al cluster 3 ($\chi^2 = 114.18$), e i successivi

"licenziare", "stipendio" e "disoccupazione" mettono in risalto le ansie occupazionali dovute alla crescente adozione dell'intelligenza artificiale nel contesto lavorativo, sia esso "pubblico" o aziendale ("azienda"). Nonostante globalmente prevalga un atteggiamento positivo nei confronti dell'AI, le preoccupazioni più concrete associate alla sua adozione riguardano l'aumento della disoccupazione (Loewen et al. 2024). Secondo uno studio di Mayer et al. (2025) e pubblicato da McKinsey & Company, una delle maggiori società internazionali di consulenza manageriale, oltre il 70% dei lavoratori statunitensi ritiene che entro due anni l'AI generativa (di seguito anche indicata come GenAI) modificherà il 30% o più delle proprie attività lavorative.

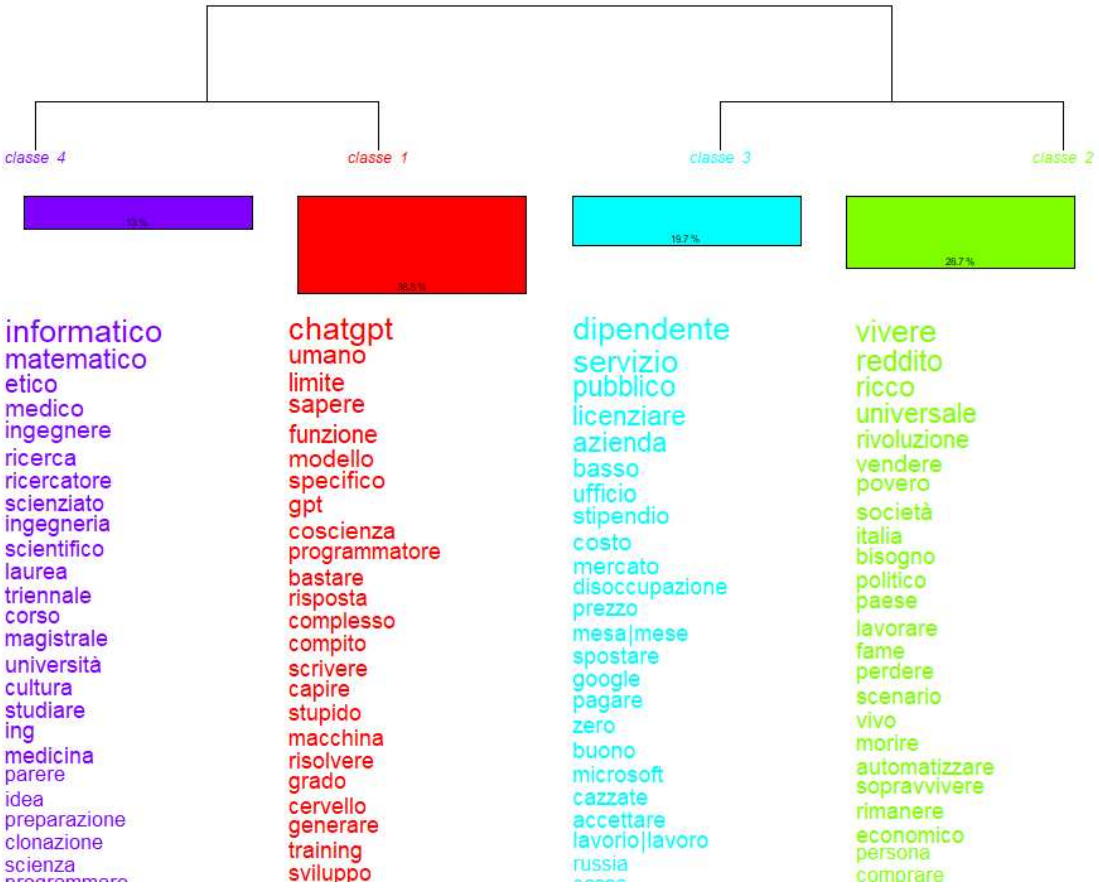

Figura 20 Anteprima dei profili lessicali dei micro-cluster emersi dalla DHC condotta sul subcorpus *sub1_lav*.

Sebbene motivati ad acquisire competenze legate all'utilizzo dell'intelligenza artificiale, "[…] AI optimists are only a slight majority in the workplace; a large minority (41

percent) are more apprehensive and will need additional support" (*ivi*, p. 4)[69]. In termini generali, i dipendenti sono tre volte più propensi dei leader a credere che l'AI sostituirà il 30% del loro lavoro già nel corso del prossimo anno. Analogamente, il "Future of Jobs Report 2025" del World Economic Forum (2025) classifica l'intelligenza artificiale come un fattore determinante nella rivoluzione del mercato del lavoro: a livello globale, entro il 2030 saranno creati 170 milioni di nuovi posti di lavoro, mentre 92 milion di ruoli potrebbero essere eliminati. Tuttavia, osservando i segmenti di testo tipici del cluster, le menzioni alle nuove opportunità di impiego sembrano restare marginali.

Nel dendogramma in *Figura 20*, le forme "licenziare" e "disoccupazione", "automazione"*[70] e "sostituire"* (che nel cluster 5 in *Figura 17* compare con un maggiore valore del chi2) suggeriscono una convergenza con i risultati delle suddette indagini, evidenziando – pur trattandosi di metodologie e/o contesti differenti – come anche il panorama italiano sia segnato dai timori per la futura situazione occupazionale. A supporto di tale interpretazione, il sondaggio "Global Public Opinion on Artificial Intelligence" (Loewen et al. 2024) pubblicato dallo Schartz Reisman Institute (SRI) rivela che in Italia, nonostante nel complesso prevalga una visione positiva dell'AI[71], il 48%[72] della popolazione ritiene che il proprio lavoro sarà "probabilmente" (11%) o "certamente" (37%) sostituito da una macchina nei prossimi dieci anni. Il 73%[73] degli italiani dichiara inoltre che, nell'ipotesi che accada, sarebbe "abbastanza difficile" (49%) o "estremamente difficile" (33%) trovare un'altra fonte di reddito o un'occupazione equivalente. Il grafico sottostante (si veda *Figura 21*), che illustra le co-occorrenze tra i lemmi appartenenti al cluster 3 a partire dalla sua forma più rappresentativa, evidenzia difatti la forte co-occorrenza tra "dipendente" e le forme "ai", "pubblico", "sostituire" e "licenziare". In particolare, l'aggettivo "pubblico" – assieme al già citato acronimo "pa",

[69] "[…] coloro che hanno una visione ottimista sull'impiego dell'AI costituiscono solo una leggera maggioranza, mentre una minoranza considerevole (41 percento) manifesta maggiore apprensione e necessiterà di supporto aggiuntivo".
[70] L'asterisco (*) segnala le forme non visibili nella figura di riferimento.
[71] Le percentuali di risposta alla domanda "Generally speaking, do you have a very positive, fairly positive, neither positive nor negative, fairly negative or very negative view of AI?" si suddividono come segue: 10% *positive*, 35% *fairly positive*, 34% *neither positive nor negative*, 17% *fairly negative*, 11% *very negative*. Per la domanda "Which statement comes closer to your view: 'I think AI will make the future better' or 'I think AI will make the future worse.'", le risposte si distribuiscono come segue: 48% *better*, 30% *unsure*, 22% *worse*.
[72] Nel restante 52%, il 29% ritiene che "probabilmente" non avverrà (*probably no*) e il 23% ritiene che invece non avverrà per certo (*definitely no*).
[73] Nel restante 27%, circa il 22% ritiene che sarebbe "abbastanza semplice" (*somewhat easy*), mentre solo il 4% dichiara che sarebbe "estremamente semplice" (*extremely easy*).

il *word-type* più discriminante per *sub1_lav* (cfr. *§3.2.2*) – è riconducibile a una submission specificamente incentrata sul possibile impatto dell'AI sul settore pubblico. L'elevato numero di commenti, criterio metodologicamente determinante per l'inclusione della submission nel corpus, mostra come il rischio della sostituzione tecnologica stimoli un'intensa partecipazione al dibattito, ma soprattutto come anche il pubblico impiego possa essere percepito vulnerabile ai processi di automazione.

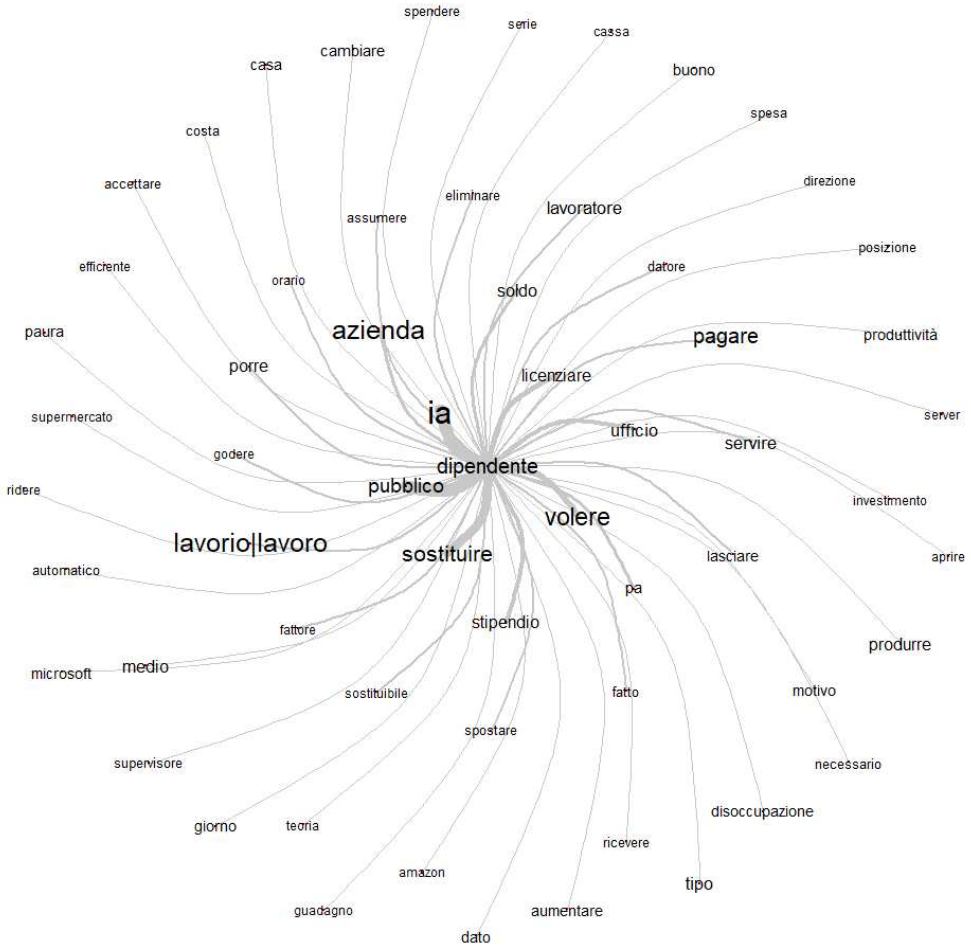

Figura 21 Rappresentazione grafica delle co-occorrenze del lemma "dipendente". Si includono i lemmi con frequenza maggiore o uguale a 5. Lo spessore dei rami è direttamente proporzionale al punteggio di co-occorrenza delle parole, mentre la dimensione delle parole riflette la loro frequenza all'interno del cluster.

D'altra parte, il lemma più distintivo della classe 2 evidenzia il carattere sociale del cluster: con un valore del chi-quadrato pari a 106.74, il verbo "vivere" si colloca nettamente al di sopra delle altre forme. Se il profilo lessicale della classe 3 tematizza la

ridefinizione del lavoro indotto dall'AI, quello della classe 2 tematizza la redistribuzione economica e l'indiretta ridefinizione del tessuto sociale: l'antonimia tra "ricco" e "povero" e lo stesso lemma "rivoluzione" richiamano la trasformazione della "società" come esito più ampio dell'introduzione dell'AI, mettendone in luce le ripercussioni sul lungo periodo. L'intelligenza artificiale è discussa su Reddit in quanto risorsa che, all'interno di una società capitalistica ("capitalismo"*, "capitalista"*), risiede nelle mani di attori economicamente e tecnologicamente avanzati. Diversi studi evidenziano difatti come l'AI sia maggiormente utilizzata da grandi aziende già produttive, da lavoratori ad alta istruzione e nei settori ad alta qualifica (Appel et al. 2026; Chatterji et al. 2025; Kergroach & Héritier 2025), ai quali si oppongono classi sociali svantaggiate dall'assenza di competenze digitali. In Italia, ad esempio, solo il 33% degli italiani dichiara di possedere le competenze necessarie per l'utilizzo dei servizi di AI generativa (Anelli et al. 2025): l'automazione ("automatizzare"), sostituendo le mansioni dei lavoratori meno qualificati, riduce le opportunità occupazionali e incrementa le disparità economiche. Anche su scala globale, "The Next Great Divergence" (Muthukrishn et al. 2025), un recente rapporto del Programma delle Nazioni Unite per lo Sviluppo (UNPD), segnala che l'intelligenza artificiale potrebbe ampliare le disuguaglianze già esistenti tra Paesi ricchi e Paesi poveri. Coerentemente con questo assunto, nel cluster 2 il lemma "italia" è frequentemente utilizzato come termine di paragone con le altre nazioni: sul fronte lavorativo, l'Italia è difatti considerata ancora indietro rispetto ad altri Paesi UE. Il profilo lessicale del cluster 2 suggerisce pertanto una visione del futuro prossimo caratterizzata da precarietà e vulnerabilità, tipicamente propria dei Paesi europei e anglofoni rispetto ai Paesi asiatici (Loewen et al. 2024). Tale interpretazione, oltre ad essere corroborata dai dati sopracitati, è accentuata dall'occorrenza dei lemmi "bisogno", "fame", "perdere", "vivo", "morire" e "sopravvivere", che confermano come per molti il potenziamento delle capacità dell'AI sia associato persino a scenari estremi di estinzione dell'umanità (*ibid.*). La classe, infine, mette in rilievo un piano socio-"politico" meno evidente nella DHC del corpus complessivo: il lemma "reddito", rappresentativo del subcorpus (cfr. *§3.2.2*) e fortemente caratteristico del micro-cluster 2 ($\chi^2 = 78.01$), è nella maggior parte dei casi associato all'aggettivo "universale" ($\chi^2 = 68.71$). Il "reddito universale", locuzione che occorre 26 volte, rappresenta per gli utenti l'unica soluzione in grado di mitigare le potenziali criticità sociali legate ai processi di automazione.

Alla sinistra del grafico (si veda *Figura 20*), la coppia di classi 1 e 4 introduce una dimensione che si discosta notevolmente da quella socio-economica racchiusa dai cluster 2 e 3. La classe 1, a cui il metodo Reinert assegna circa il 39% dei segmenti di testo totali, rivela una nuova sfaccettatura del discorso, rimasta latente tanto nell'analisi del corpus complessivo quanto nell'analisi degli altri subcorpora: quando oggetto del dibattito è l'impiego dell'AI in ambito lavorativo, la discussione fa spazio all'antropomorfizzazione dell'AI e al confronto tra intelligenza artificiale e umana. Quest'ultimo risulta immediatamente evidente osservando in due lemmi più caratteristici del cluster: "chatgpt" ($\chi^2 = 82.63$) e "umano" ($\chi^2 = 48.83$) delineano due poli tematici in cui lemmi tendenzialmente associati alla sfera computazionale ("funzione", "modello", "macchina") e lemmi associati alla sfera cognitiva ("sapere", "coscienza", "cervello") si intrecciano all'interno di un'unica classe semantica. In linea con questi risultati, il già citato "Global Public Opinion on Artificial Intelligence" (Loewen et al. 2024) discute la difficoltà dei rispondenti nel delineare con chiarezza il rapporto tra ciò che è intelligenza artificiale e ciò che invece è umano: "[m]any define these as opposites: what is AI is not human, although it might do – or try to do – what humans do. More commonly, AI is described as a technology (or a robot, or a computer and so on) that can do or try to do what humans do or have the intelligence of a human" (*ivi*, p. 7)[74]. In termini generali, i rispondenti si dividono tra coloro che lodano l'intelligenza artificiale, superiore a quella umana, e coloro che invece la giudicano in quanto tecnologia che tenta di simulare, in modo incompleto o insoddisfacente, quei processi tipicamente associati alla cognizione umana. Il profilo lessicale del cluster suggerisce difatti uno scambio di riflessioni tanto sulle "capacità" quanto sui limiti ("limite") dell'AI: in particolare, i redditors descrivono ChatGPT come un "modello" che è in grado di "risolvere" problemi di vario tipo e "generare" codici, video e immagini, ma che frequentemente si rivela inefficace, fornendo risposte ("risposta") accondiscendenti piuttosto che fondate. Il GPO-AI (Loewen et al. 2024) torna a supporto di questi dati, registrando un calo nella soddisfazione degli utenti rispetto all'accuratezza delle risposte dell'AI generativa, nonostante il suo utilizzo in costante aumento: nel 2025 il livello di soddisfazione passa dal 60% del 2024 al 56%,

[74] "[m]olti li definiscono come poli opposti: ciò che è AI non è umano, sebbene possa svolgere – o tentare di svolgere – attività umane. Più comunemente, l'AI viene descritta come una tecnologia (o un robot, o un computer e simili) in grado di fare, o provare a fare, ciò che fanno gli esseri umani, o che può possedere un'intelligenza paragonabile a quella umana".

mentre il 13% dei rispondenti dichiara che gli strumenti di GenAI non forniscano risposte o soluzioni precise. Il cluster 1 esemplifica il disallineamento tra le capacità cognitive che gli utenti attribuiscono all'AI generativa e le sue reali capacità operative: come altri *Large-Language Models* ("llm"*), ChatGPT è in grado di produrre output linguistici altamente efficaci, motivo per cui vige tra i suoi fruitori la tendenza a sovrastimare le capacità cognitive di un "modello" che opera, in realtà, esclusivamente sulla base di stime probabilistiche di co-occorrenza delle parole. Esattamente in quanto modello probabilistico fondato su enormi quantità di dati testuali di addestramento (*"training" set*) e privo di qualunque forma di "coscienza" o capacità di "ragionare", la correttezza e l'accuratezza delle sue risposte – denominate, in tal caso, "allucinazioni" (28 occorrenze all'interno del corpus) – non è in realtà garantita. L'esaminazione dei segmenti di testo tipici del cluster suggerisce infatti una tendenza – in particolare da chi dichiara la propria professione di informatico o matematico – a sminuire le capacità operative dell'intelligenza artificiale e, in particolare, a mettere in discussione la possibilità che essa possa sostituire integralmente la figura del "programmatore", lo stesso responsabile del suo sviluppo e della sua implementazione. Ciononostante, gli utenti di Reddit ne riconoscono il potenziale evolutivo. Il lemma "limite", ad esempio, viene utilizzato perlopiù in riferimento alle carenze dei modelli attuali, considerate transitorie e risolvibili in virtù del costante "sviluppo" tecnologico.

All'interno del dibattito, le riflessioni sulle relazioni tra intelligenza artificiale e umana danno origine ad un nucleo semantico distinto, seppur affine a quello del cluster 1. A differenza di quest'ultimo, che introduce unicamente la figura specifica del programmatore, il cluster 4 raggruppa una serie di ruoli professionali riconducibili all'ambito STEM. Innanzitutto, "informatico", il lemma più distintivo del cluster (χ^2 = 320.65) e tra le forme con il più elevato TF-IDF per *sub1_lav*, estende la conversazione a tutti i soggetti coinvolti nel settore IT. Ad esso si affiancano "matematico", "ingegnere" e "ricercatore", che contribuiscono a delineare il cluster 4 come la classe semantica del sapere specialistico umano. Il cluster 4 tematizza il paradosso secondo cui coloro che progettano e sviluppano l'AI rischierebbero di essere sostituiti da una loro stessa creazione. Tuttavia, contrariamente ai micro-cluster 1, 2 e 3, nei quali il rapporto che intercorre tra intelligenza artificiale e intelligenza umana è definita in termini di competizione o sostituzione, il cluster 4 mette in luce una dipendenza costitutiva.

L'intelligenza artificiale è riconosciuta come il prodotto dell'intelletto umano: senza i processi di "ricerca" e un'elevata "preparazione" tecnica dell'uomo, algoritmi e modelli non potrebbero esistere. Tale interpretazione è rafforza dalla presenza, nel profilo lessicale della classe 4, di termini appartenenti al campo semantico accademico, quali "laurea", "triennale", "corso", "magistrale", "università", "cultura" e "studiare", che segnalano la formazione specialistica come prerequisito fondamentale per lo sviluppo dell'AI. Il cluster è infine caratterizzato dai lemmi "medico" e "etico", che evidenziano le implicazioni dell'adozione dell'intelligenza artificiale in contesti ad alto impatto sociale: i redditors ricorrono ai lemmi "medico" e "medicina" per confermare il settore sanitario come uno dei principali ambiti di applicazione dell'AI; il lemma "etico", invece, fa riferimento ai rischi e ai quesiti etici sollevati dall'impiego dell'AI nella ricerca informatica, come nel caso della progettazione di veicoli a guida autonoma.

4.2.2 Salute mentale e relazioni: l'antropomorfizzazione del chatbot come amico, psicologo e partner virtuale

Nella *Descending Hierarchical Classification* del corpus complessivo (si veda *Figura 17*), il cluster 4 è gerarchicamente equivalente al cluster 5: le due classi semantiche sono prodotte nella medesima fase dell'applicazione del metodo Reinert e condividono, pertanto, profili lessicali affini. Ciononostante, la classe 4, a cui viene assegnato il 22,9% dei segmenti testuali analizzati, si distingue immediatamente per la natura relazionale e/o esperienziale dei suoi lemmi caratterizzanti. Con un valore del chi-quadrato pari a 298.53, "persona" costituisce il lemma più distintivo, a cui seguono termini appartenenti a campi semantici più circoscritti: i lemmi "vita", "sentire", "parlare", "età", "esperienza" rimandano al vissuto soggettivo degli utenti, mentre "amico", "donna", "relazione", "amicizia", "rapporto", "empatia", "fidanzato" e "sesso" collocano il discorso su un piano relazionale, sia esso sentimentale o amicale; a questi, si alternano "psicologo" e "terapia", che richiamano la presenza di difficoltà personali, la ricerca di un supporto professionale e un bisogno di ascolto.

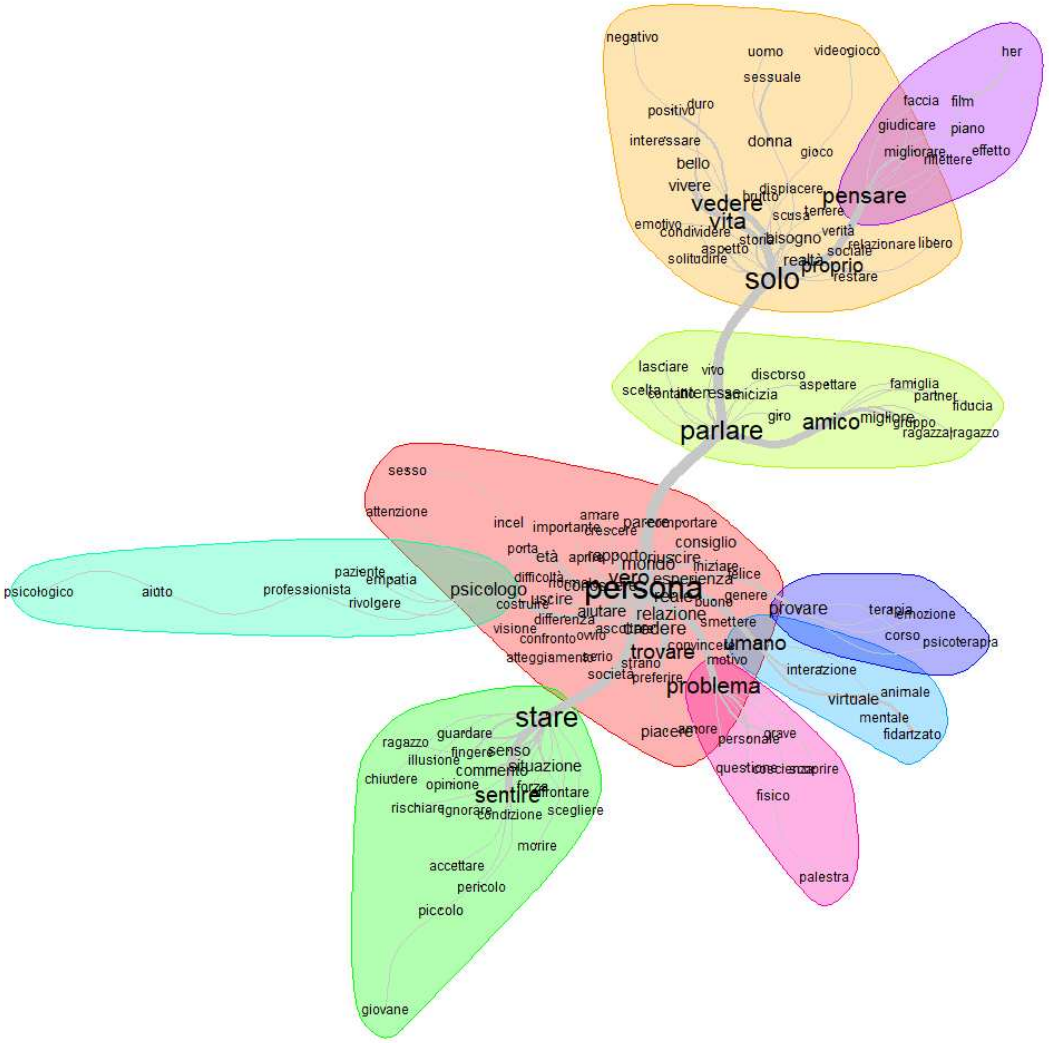

Figura 22 Grafico del macro-cluster 4, identificato come classe semantica della salute mentale e delle relazioni sociali. Si includono i lemmi con una frequenza maggiore o uguale a 15. Le linee di collegamento tra i nodi illustrano la frequenza con cui i lemmi co-occorrono: maggiore è lo spessore della linea, maggiore è l'associazione tra i lemmi.

Alla luce di tali osservazioni, a livello semantico il cluster 4 risulta corrispondente a *sub4_rel*, il subcorpus definito a priori per tematizzare l'impiego dell'AI come strumento di ascolto o sfogo e il suo impatto sulla salute mentale e sulle relazioni sociali. Esaminando il profilo lessicale della classe, il primo dato degno di nota è l'antropomorfizzazione dell'intelligenza artificiale, con cui l'utente instaura un rapporto interpersonale: essa non emerge più solo come strumento di supporto cognitivo od operativo, bensì come un interlocutore ritenuto in grado di fornire all'utente sostegno

emotivo o psicologico. Il cluster 4 getta pertanto luce su una nuova, più recente, modalità di interazione con l'AI, che è tanto strumento funzionale quanto strumento di regolazione emotiva. Essa assume diverse declinazioni, svolgendo il ruolo di un amico con cui confidarsi ("amico", "amicizia"), di una figura professionale di supporto ("psicologo", "terapeuta", "terapia") o persino di un partner virtuale ("donna", "relazione", "rapporto", "fidanzato", "sesso", "incel", "amore"). In ciascuno dei tre casi, l'intelligenza artificiale è protagonista di un uso compensativo, in cui i confini tra interazione umana e interazione mediata dai chatbot risultano sempre meno definiti: l'assenza di un interlocutore "reale" accogliente ed empatico è compensata dal ricorso ad un interlocutore "virtuale" in grado di simulare le emozioni umane ("emozione"), di fronte al quale ci si sente ("sentire") liberi di "parlare" perché, programmato per essere accondiscendente, ascolta ("ascoltare"*) senza "giudicare".

Negli ultimi anni, la letteratura sull'affidamento del supporto emotivo a un algoritmo è aumentata in modo consistente, pur mostrando risultati contrastanti. Ad un anno dal boom dell'AI generativa, una ricerca del National Bureau of Economic Research (NBER) ha analizzato milioni di conversazioni avvenute su ChatGPT a partire da novembre 2022, anno del suo rilascio, fino a luglio 2025, quando il suo utilizzo ha raggiunto il 10% della popolazione globale adulta. I risultati dello studio, condotto da Chatterji et al. (2025), riconducono al tema della *Self-Expression* solo il 5.3% dei messaggi complessivi con il chatbot, distribuito tra le sottocategorie *Greetings and Chitchat*, *Relationships and Personal Reflection* e *Games and Role Play*. L'espressione identitaria e, in senso lato, la ricerca di una connessione emotiva sembrano pertanto ricoprire solo una quota marginale delle conversazioni con ChatGPT. Ciononostante, questa categoria registra il maggior livello di soddisfazione da parte dell'utente rispetto alle risposte del chatbot, con un rapporto *good-to-bad ratio* pari a 7.86 a indicare che, sebbene attestato come poco frequente, il tentativo di instaurare un contatto emotivo risulta particolarmente gratificante. Come anticipatamente osservato, l'utente medio si rivela sempre meno soddisfatto delle risposte fornite dalla GenAI: coerentemente con i risultati del GPO-AI (Loewen et al. 2024), l'NBER (Chatterji et al. 2025) rileva un livello di soddisfazione insufficiente per le altre categorie d'uso di ChatGPT, tra le quali *Practical Guidance* (4.37), *Seeking Information* (4.75) e *Writing* (3.11), le tre categorie principali, saranno discusse successivamente (si veda *§4.2.3*). Ciò suggerisce che, laddove lo scopo

dell'interazione sia la ricerca di un supporto emotivo o relazionale, l'accuratezza delle risposte del chatbot incida marginalmente sul grado di soddisfazione: l'utente che si rivolge all'intelligenza artificiale per un supporto emotivo non è alla ricerca di una soluzione a un problema concreto, ma tenta esclusivamente di rispondere al proprio bisogno di natura emotiva e soddisfare il desiderio di essere ascoltato. Opponendosi alla marginalità della *Self-Expression* dimostrata dall'NBER, lo studio di Zao-Sanders (2025) riconosce la crescente frequenza con cui l'AI generativa è impiegata per finalità socio-emotive. Il report, dal titolo *How People are Really Using Generative AI Now*, presenta le prime 100 tipologie di utilizzo della GenAI nel 2025 e classifica *Therapy/Companionship* come il caso d'uso più diffuso. A seguire, *Organise my life* e *Find purpose* occupano la seconda e terza posizione: sul podio si collocano dunque tre modalità di utilizzo che l'autore ascrive alla categoria *Personal & Professional Support*. Il confronto con i risultati del 2024[75] evidenzia una netta transizione da un uso prevalente tecnico e orientato alla produttività verso applicazioni incentrate sul benessere personale, sull'organizzazione della vita quotidiana e sull'esplorazione esistenziale.

I risultati della DHC condotta su *sub4_rel* (si veda *Figura 23*), consentono di esplorare il modo in cui la ricerca di un supporto mediato dall'AI si articola all'interno del dibattito tra redditors. Il grafico sottostante illustra i tre micro-cluster emersi dall'analisi condotta sul subcorpus semanticamente corrispondente al macro-cluster 4 (cfr. *Figura 17*). I dati rispecchiano le tre suddette modalità di interazione con l'intelligenza artificiale, sebbene non si pongano sullo stesso piano gerarchico. Nel dendogramma, essi sono difatti strutturati in funzione del ruolo attribuito all'AI: nella classe 2, che costituisce un nucleo tematico a sé stante, il ricorso all'AI assume un ruolo secondario, mentre nelle classi 1 e 3, prodotte dalla medesima biforcazione del primo ramo, la sua concreta implementazione emerge in maniera più evidente.

[75] *Generate ideas* (categoria: *Content Creation & Editing*), *Therapy / companionship* (categoria*: Personal & Professional Support*), *Specific Search* (categoria: *Research, Analysis & Decision Making*), che nel 2024 rappresentavano le tre principali tipologie di utilizzo dell'AI, nel 2025 si classificano – rispettivamente – al sesto, primo e undicesimo posto.

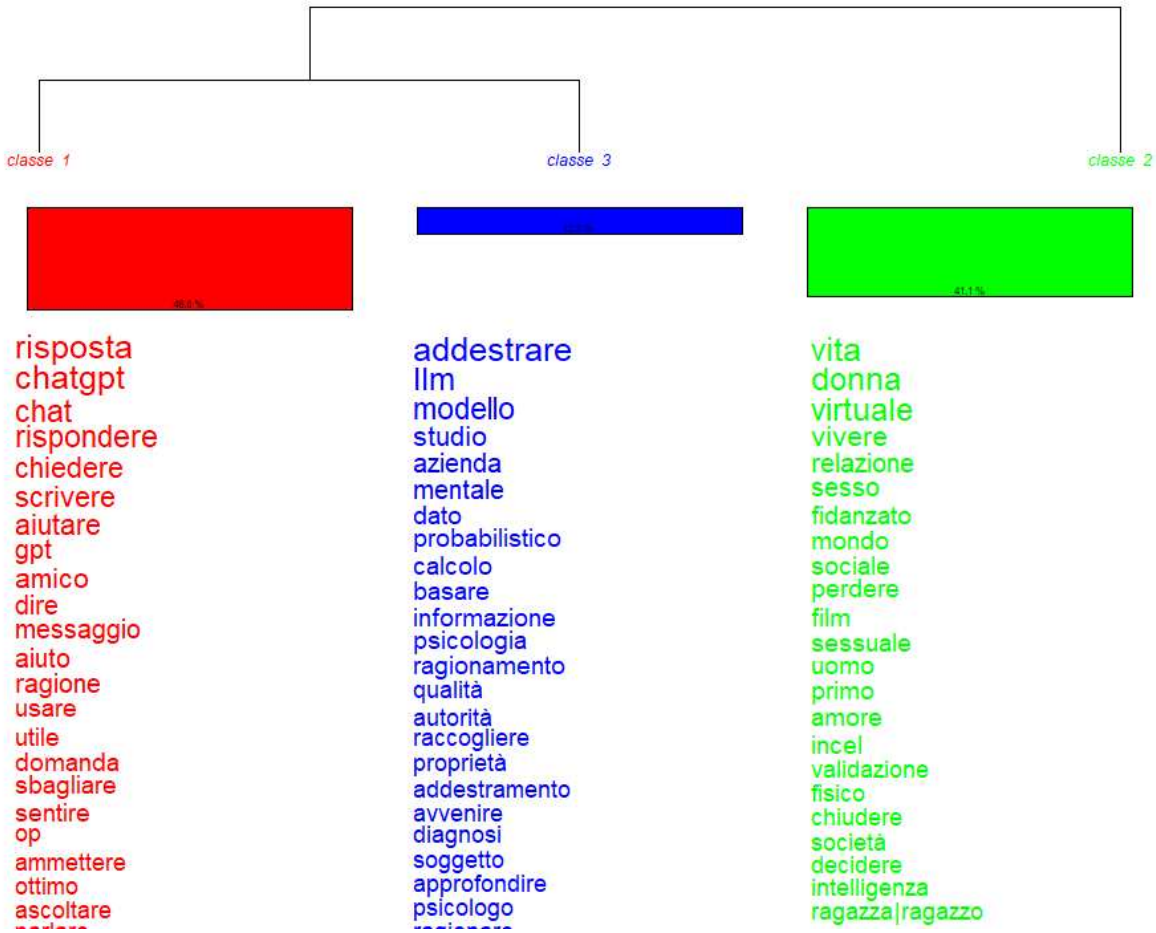

Figura 23 Anteprima dei profili lessicali dei micro-cluster emersi dalla DHC condotta sul subcorpus *sub4_rel*.

Come si evince dal lemma più distintivo – oltre che più frequente – all'interno del cluster, la classe 2 è prevalentemente incentrata sulla "vita" (χ^2 = 62.03) degli utenti, che indugiano sul racconto delle proprie esperienze e riflettono su quelle condivise dagli altri. In particolare, la classe 2 sembra tematizzare le motivazioni di natura sentimentale e sessuale che sottendono il ricorso all'AI. Il cluster rispecchia un dato rilevante già osservabile nello scenario nazionale: secondo un sondaggio condotto da Censuwide per conto di LELO (2025), azienda svedese leader nel settore dei sex toys, il 71% degli italiani[76] (il 76% degli uomini e il 66% delle donne) dichiara di utilizzare o di aver utilizzato l'AI per aspetti legati alla propria vita intima, con lo scopo di migliorare la

[76] A livello globale, il dato si attesta al 60%, mentre ulteriori picchi di attività si registrano per gli spagnoli (81%) e i francesi (70%).

qualità dei loro rapporti sessuali e delle loro relazioni (Moro 2026; PR Newswire 2026). Viene ad esempio impiegata per ricevere consigli in merito a flirt o appuntamenti (34%), per aumentare la fiducia in sé stessi prima di un incontro (22%) o per generare scenari o idee di giochi di ruolo che possano rendere più soddisfacente l'esperienza sessuale (27%). Il profilo lessicale del cluster è difatti caratterizzato da lemmi come "donna", "relazione", "sesso", "fidanzato", "sessuale", "uomo" e "amore", che, osservando i segmenti testuali tipici del cluster, occorrono entro un dialogo fortemente polarizzato tra legittimazione e stigma: il ricorso all'AI è sia una legittima risposta alla solitudine sia sintomo di un crescente degrado sociale in cui la dipendenza affettiva e/o sessuale si fa anche virtuale.

Per meglio illustrare il dibattito, si propone un nuovo grafico (si veda *Figura 24*) raffigurante le co-occorrenze tra i lemmi appartenenti al micro-cluster 2. In questo caso, si sceglie di porre al centro il lemma aggettivale "virtuale", il terzo termine più distintivo ($\chi^2 = 57.1$) ma al contempo il più singolare all'interno della classe semantica. Osservando la *Figura 24* si rende immediatamente evidente la contrapposizione con "reale" e l'elevata associazione con i lemmi "realtà", "mondo" e, in particolare, il lemma "fidanzato". Quest'ultimo, accostato ai termini "attrazione", "intimità", "contatto", "sesso" e "sessuale", testimonia la sempre più diffusa e discussa tendenza a "instaurare" un legame non solo emotivo ma anche sessuale con i chatbot di intelligenza artificiale, progressivamente integrati tanto nella dimensione sentimentale quanto in quella intima della propria vita privata. Domina dunque l'uso dell'AI come alternativa praticabile, frequentemente descritta come una vera e propria "scelta" ("scegliere", "decidere") di "vita" poiché fonte di "conforto" e "validazione" in una società da cui ci si sente abbandonati ("abbandonare"). Oltre che una "cura" alla "depressione" e al "dolore", l'AI si configura come uno spazio sicuro di sperimentazione ("sperimentare"), in cui l'utente si sente "lontano" e "libero" da ogni tipo di giudizio ("giudicare"). D'altra parte, la discussione è contornata da una prospettiva stigmatizzante nei confronti di coloro che si affidano in modo eccessivo all'AI come forma di supporto emotivo, seppur espressa con registri diversi da parte dei redditors: da un lato, il degrado delle relazioni sociali è da attribuire all'uomo stesso, unico colpevole della propria alienazione perché incapace di affrontare il mondo reale; dall'altro, emerge una retorica paternalistica in cui lo stigma è accompagnato da consigli pratici ("cambiare", "smettere", "riuscire", "migliorare",

"sperare", "riflettere"), volti a incoraggiare il miglioramento personale o il ricorso a supporti psicologici tradizionali.

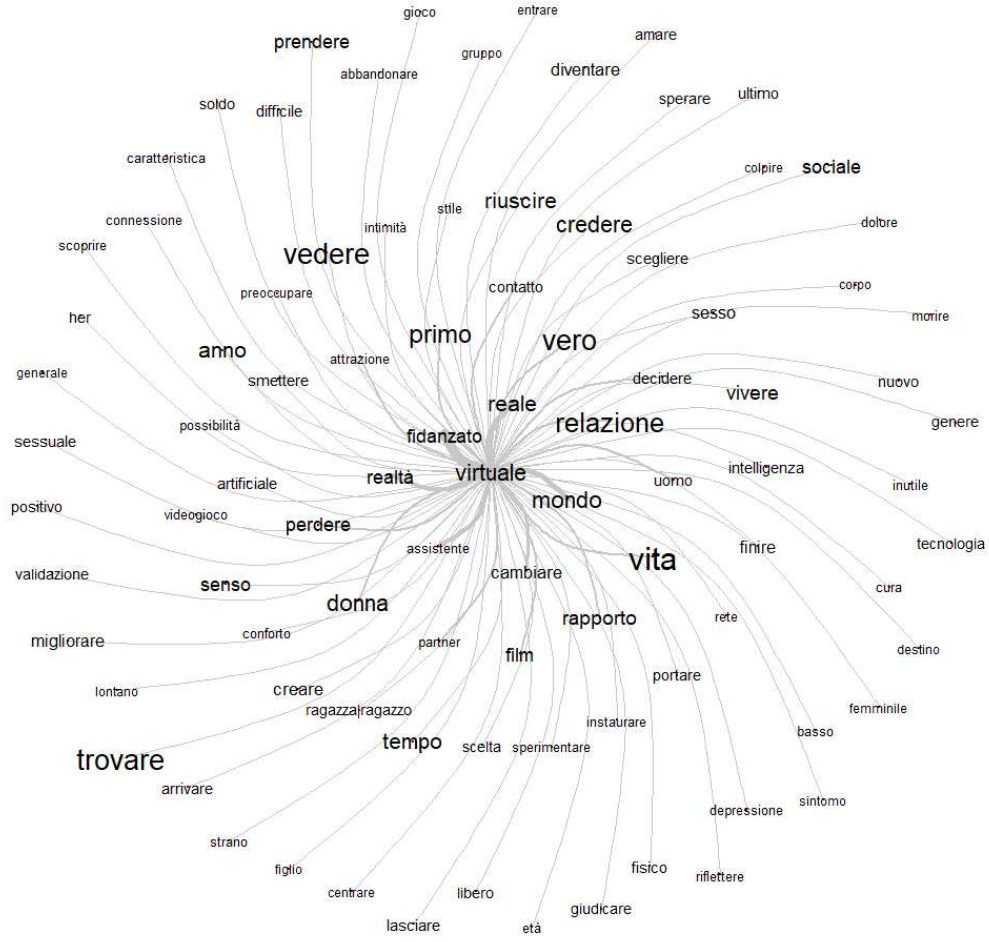

Figura 24 Rappresentazione grafica delle co-occorrenze del lemma "virtuale". Si includono i lemmi con frequenza maggiore o uguale a 5. Lo spessore dei rami è direttamente proporzionale al punteggio di co-occorrenza delle parole, mentre la dimensione delle parole riflette la loro frequenza all'interno del cluster.

A tal proposito, è significativa l'appartenenza al micro-cluster 2 del forestierismo "incel", la forma con il maggior peso TF-IDF all'interno del corpus complessivo e la più discriminante per *sub4_rel* (cfr. *§3.2.2*). Se ne illustrano le co-occorrenze in *Figura 25*. Come anticipato, con l'appellativo di "incel" si fa riferimento ai cosiddetti "celibi involontari". Il termine veniva in origine utilizzato per descrivere la condizione di individui che non riuscivano a instaurare una relazione sentimentale o sessuale nonostante il desiderio di farlo. Nel tempo, l'appellativo ha tuttavia assunto una connotazione

negativa. Al giorno d'oggi è difatti associato ai membri di una subcultura online costituita in larga parte da uomini dall'ideologia misogina, che si definiscono incapaci di instaurare relazioni intime a causa di fattori percepiti come insormontabili, quali lo status socio-economico e/o l'aspetto fisico ("fisico", "brutto"* e "basso"* emergono tra i lemmi associati al cluster 2).

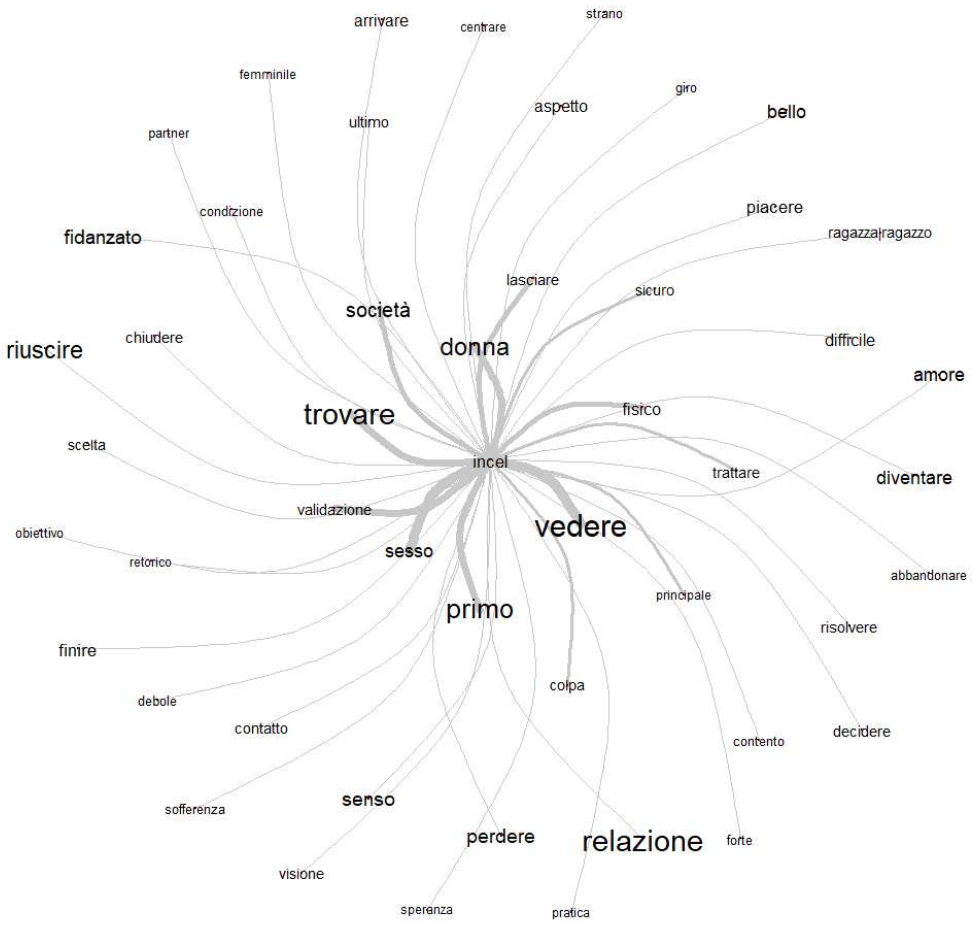

Figura 25 Rappresentazione grafica delle co-occorrenze del lemma "incel". Si includono i lemmi con frequenza maggiore o uguale a 5. Lo spessore dei rami è direttamente proporzionale al punteggio di co-occorrenza delle parole, mentre la dimensione delle parole riflette la loro frequenza all'interno del cluster.

Tratti caratteristici dell'incel sono il risentimento verso la società, l'autocommiserazione e la marcata tendenza ad attribuire la responsabilità del loro insuccesso relazionale alle donne, per questo motivo spesso denigrate e vittime di oggettificazione sessuale (Rodríguez et al. 2025; DeCook 2021). Dall'esaminazione del grafico soprastante emerge

difatti l'elevata co-occorrenza con il lemma "donna", oltre che con "sesso", "validazione", "società", "colpa" e "fisico". In questo contesto, la marcata correlazione tra i termini "incel" e "donna" testimonia la presenza di un conflitto di genere che vede da un lato gli uomini, che percepiscono un marcato senso di ingiustizia, e dall'altro le donne, ritenute responsabili delle loro difficoltà relazionali. Inoltre, il lemma "donna", secondo per valore del chi-quadrato ($\chi^2 = 57.1$), compare con maggiore frequenza rispetto al suo opposto "uomo" (44 occorrenze contro 19), consolidando la centralità della figura femminile nel discorso sulle difficoltà relazionali percepite dagli uomini.

La frequenza del termine "incel" (pari a 20 all'interno del cluster e a 27 all'interno del subcorpus), assieme al valore rappresentativo che ricopre per la classe semantica, lascia spazio a diverse interpretazioni. In primo luogo, suggerisce come il dibattito affronti il tema della solitudine maschile, dalla quale il ricorso ai chatbot costituisce – legittimamente o meno – una via di fuga. In questo contesto, può essere riconosciuta all'intelligenza artificiale una funzione di regolazione emotiva: i chatbot consentono agli utenti che ne fanno uso di sfogare i ricorrenti sentimenti di odio e frustrazione, riducendo il rischio di arrecare danno diretto ad altre persone. D'altra parte, la loro antropomorfizzazione potrebbe rafforzare le dinamiche di isolamento sociale a cui gli incel sono già soggetti: i chatbot rispondono alla ricerca di validazione costante per cui sono programmati e che gli utenti percepiscono come negata dalla società stessa. Al micro-cluster 2 appartiene difatti anche il lemma "film", verosimilmente riconducibile a *Her*, il film discusso in precedenza per il suo valore discriminante all'interno di *sub4_rel* (cfr. *§3.2.2*). In secondo luogo, la presenza del termine, spesso impiegato in modo dispregiativo e dunque difficilmente autodichiarato, si fa indice di come la discussione tra i redditors si dispieghi su un ulteriore livello, diverso da quello del conflitto di genere: il denominativo "incel" funge da linea di frattura tra gruppi sociali, vale a dire tra coloro che vengono associati alla categoria degli *involuntary celibates* e coloro che, stigmatizzando tale impiego dell'AI – sia con atteggiamento critico, moralistico o per timore dei rischi futuri – vi si oppongono con decisione.

Tornando ad osservare la *Figura 23*, i profili lessicali delle classi 1 e 3 mancano della dimensione emotiva fortemente caratterizzante del cluster 2. Come anticipato, si tratta di due sub-cluster semanticamente affini nei quali a predominare è la concreta implementazione dell'intelligenza artificiale. L'esaminazione dei profili lessicali

evidenzia tuttavia due differenze rilevanti. In primo luogo, l'implementazione dell'AI è verbalizzata secondo prospettive diverse: se l'uno valorizza l'utente e le sue richieste, l'altro si focalizza sulla macchina e sul suo funzionamento. La classe 1 ne rappresenta difatti l'uso operativo: tra i lemmi caratterizzanti, i verbi "rispondere", "chiedere", "scrivere", "aiutare" descrivono pragmaticamente il tipo di scambio dialogico iniziato dall'utente ed eseguito dal chatbot ("chatgpt"). D'altra parte, la classe 3 ne tematizza l'aspetto tecnico, come testimoniato dai lemmi "addestrare", "llm", "modello", "studio", "dato", "probabilistico", che esplicitano i meccanismi sottostanti al funzionamento dell'AI.

In secondo luogo, sebbene i lemmi associati ai due cluster ribadiscano la ricerca di un interlocutore, essi differiscono nella tipologia di figura che l'intelligenza artificiale e, più nello specifico, i chatbot ("chatgpt") sono chiamati a sostituire o a integrare: l'AI assume il ruolo di amico e/o di psicologo. I segmenti di testo appartenenti al cluster 1 descrivono la ricerca di un confidente informale, un "amico" paziente e disponibile ("ascoltare" "parlare"*, "confronto"*, "conversazione"*, "capire"*, "sfogare"*, "consiglio"*) che l'utente tende talvolta a preferire alle amicizie tradizionali. Alla classe 1 sono difatti associati lemmi come "ragione" e "accondiscendere"*, che sottolineano il carattere non conflittuale di ChatGPT. Difatti, se l'interazione mediata dall'AI costituisce uno spazio di confronto sicuro, nelle relazioni tradizionali l'utente si percepisce più esposto al giudizio sociale e teme di rappresentare un carico emotivo. A tal proposito, una delle submission mostra un'ulteriore modalità di utilizzo dell'intelligenza artificiale: un tema di discussione prevalente è l'eticità di ricorrere a ChatGPT per rispondere a un amico in difficoltà, delegando al modello la scrittura di messaggi di conforto. In entrambi i casi – che si tratti dell'utente in cerca di supporto emotivo o dell'utente chiamato a fornirlo – l'interazione con l'intelligenza artificiale sostituisce il reale dialogo con amici e familiari, incidendo sulla vita personale dell'utente e intaccando l'autenticità dei legami interpersonali.

Il cluster 3 delinea invece il chatbot come alternativa alla figura professionale dello "psicologo". Il profilo lessicale della classe semantica è caratterizzato da termini quali "psicologia", "mentale" (spesso preceduto da "salute"*), "diagnosi" e "psicoterapeuta"*, che tuttavia risultano meno discriminanti rispetto al lessico tecnico. La predominanza di quest'ultimo evidenzia che, diversamente da quanto ci si potrebbe attendere, il cluster non

si concentra sui benefici o sui rischi derivanti dall'impiego dell'AI come sostituto di un supporto professionale. I potenziali danni causati dell'affidare la propria salute mentale a un "modello" che genera risposte basate su calcoli ("calcolo") probabilistici ("probabilistico") sono menzionati solo marginalmente. Contrariamente, il nucleo tematico del cluster è da ricondurre all'attuabilità tecnica di questa sostituzione. I lemmi più caratteristici del cluster sono infatti quelli relativi al funzionamento dei *Large-Language Models*: con valori del chi2 compresi tra 122.34 e 87.5, "addestrare", "llm" e "modello" occorrono nei segmenti di testo che mettono in risalto, da un lato, la consapevolezza che simulare un reale consulto psicologico richieda sistemi addestrati su grandi quantità di dati ("addestramento", "dato", "informazione") e, dall'altro, la convinzione che questo progresso possa realmente concretizzarsi in futuro. Il chatbot, al momento, resta un modello probabilistico progettato per essere accondiscendente e privo di capacità di "ragionamento". A conclusione dell'analisi dei sub-cluster 1 e 3, è opportuno precisare che le figure dell'amico e dello psicoterapeuta, pur rappresentative delle rispettive classi semantiche, compaiono solo dopo i lemmi più operativi o tecnici, suggerendo come, anche nella ricerca di un'interazione di tipo umano, la "macchina"* venga in larga parte percepita come tale.

4.2.3 Utilizzo quotidiano: un supporto operativo tra scrittura, ricerca e potenziali dilemmi etici

Il macro-cluster 1 è tematicamente corrispondente al subcorpus incentrato sull'utilizzo quotidiano dell'AI generativa (*sub3_app*), a cui oggi ricorre il 48% degli italiani (il 20% in più rispetto al 2024), spesso od ogni giorno nel 14% dei casi (Anelli et al. 2025): la presenza delle forme "uso" e "usare" esplicita il focus tematico della classe. All'interno del 22,6% dei segmenti testuali assegnati a tale cluster, "scrivere", "chatgpt", "uso", "chiedere", "risposta" e "usare" rappresentano difatti i sei lemmi più caratterizzanti, con valori del chi-quadrato compresi tra 518 e 175.42. In particolare, per quanto riguarda il termine "chatgpt", l'elevato contributo del chi2 ad esso associato ($\chi^2 = 354.52$) – accanto alle 747 occorrenze all'interno del corpus complessivo – ribadisce il primato del chatbot di OpenAI tra i modelli di intelligenza artificiale generativa preferiti dal pubblico.

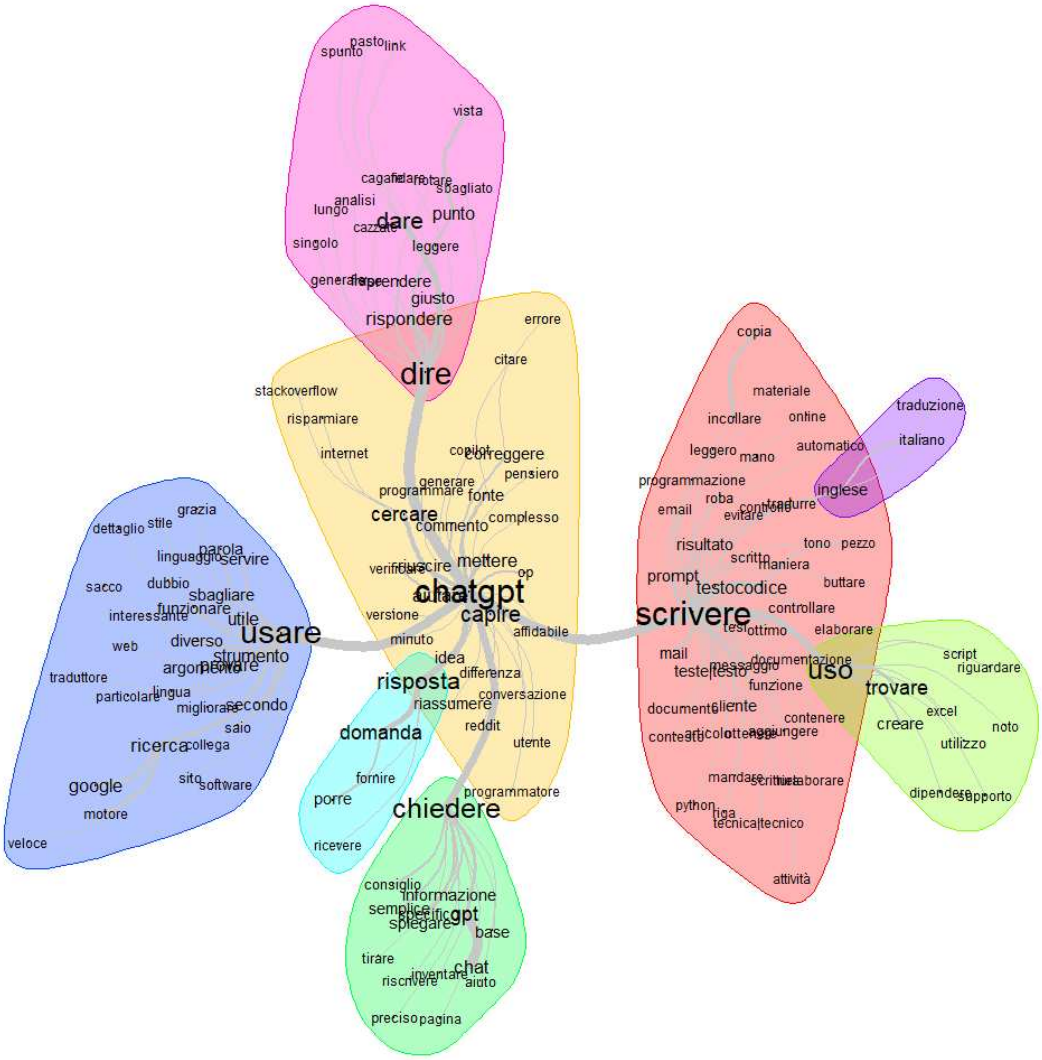

Figura 26 Grafico del macro-cluster 1, identificato come classe semantica delle applicazioni nella vita quotidiana. Si includono i lemmi con una frequenza maggiore o uguale a 15. Le linee di collegamento tra i nodi illustrano la frequenza con cui i lemmi co-occorrono: maggiore è lo spessore della linea, maggiore è l'associazione tra i lemmi.

Lo studio dell'NBER (Chatterji et al. 2025), già citato nella sezione precedente, offre importanti approfondimenti sull'utilizzo odierno del chatbot, costituendo una fonte indispensabile per l'interpretazione dei lemmi associati alla classe. Secondo il report, dal titolo *How People Use ChatGPT*, quasi l'80%[77] dell'utilizzo complessivo di ChatGPT rientra in tre ampie categorie, denominate *Practical Guidance* (ricevere consigli pratici),

[77] Le percentuali totali si articolano come segue: *Practical Guidance* (28.8%), *Seeking Information* (24.4%), *Writing* (23.9%), *Multimedia* (7,3%), *Self-Expression* (5,3%), *Technical Help* (5,1%), *Other/Unknown* (5,2%).

Seeking Information (cercare informazioni) e *Writing* (scrivere). All'interno del cluster 1, i lemmi "chiedere", "risposta", "domanda" "ricerca", "google", "cercare" o "aiutare" rimandano alle prime due categorie di utilizzo: come si legge sul report, "*Practical Guidance* […] includes activities like tutoring and teaching, how-to advice about a variety of topics, and creative ideation. *Seeking Information* includes searching for information about people, current events, products, and recipes, and appears to be a very close substitute for web search" (*ivi*, p. 2)[78]. Nella DHC condotta limitando l'analisi a *sub3_app* (si veda *Figura 27*), emerge difatti una classe semantica orientata all'utilizzo dell'AI come motore di ricerca (classe 1), come testimoniato dalle forme "google", "fonte", "ricerca", "cercare", "motore", "chatgpt", "link" e "search". A supporto dei suddetti risultati, il nuovo "Digital 2026" di We Are Social (We Are Social & Meltwater 2025c) mostra la diminuzione della percentuale di persone che, in Italia, visitano mensilmente i motori di ricerca e l'aumento del traffico web generato dalle piattaforme di intelligenza artificiale, tra le quali ChatGPT, producendo l'84.5% dei referral, si attesa in prima posizione, seguita da Microsoft Copilot (8.55%) e Perplexity (4.23%).

La terza categoria di utilizzo, *Writing*, trova esplicita realizzazione nel lemma più caratterizzante del cluster 1: "scrivere" ($\chi^2 = 518.2$). Essa include "[…] the automated production of emails, documents and other communications, but also editing, critiquing, summarizing, and translating text provided by the user" (*ibid.*)[79]. Tra le prime posizioni del profilo lessicale, troviamo difatti "mail", "documento", "tesi", ma anche "correggere", "riassumere", "traduzione", "riscrivere", "riformulare" e "rielaborare", tutte forme che confermano come due terzi dei messaggi appartenenti alla categoria *Writing* siano in realtà perlopiù richieste di modifica, rielaborazione e traduzione, piuttosto che di scrittura di un testo *ex novo*. Nell'applicazione del metodo Reinert su *sub3_app*, il profilo lessicale del micro-cluster 5 riassume analogamente le principali modalità di impiego dell'AI e riporta parte di tali lemmi caratterizzanti ("scrivere", "mail", "uso", "teste|testo"). Assumono particolare rilievo "inglese e "lingua", due forme che segnalano un utilizzo diffuso dell'AI per le attività di traduzione, specialmente verso l'inglese e dall'inglese. Osservando il grafico sottostante, si nota come il ramo della classe

[78] "*Practical Guidance* comprende attività come il tutoraggio e l'insegnamento, consigli pratici su una varietà di argomenti e sviluppo di idee. *Seeking Information* riguarda la ricerca di informazioni su persone, eventi attuali, prodotti e ricette, e sembra costituire un sostituto molto vicino alla ricerca sul web".
[79] "[…] la produzione automatizzata di email, documenti e altre forme di comunicazione, ma anche la revisione, l'analisi critica, la sintesi e la traduzione di testi forniti dall'utente".

5 dia origine, oltre che al già discusso cluster 1, al cluster 2. In modo comparabile alla classe 2 prodotta nell'analisi del corpus complessivo (si veda *Figura 17*), il cluster 2 di *sub3_app* è incentrato sul funzionamento "statistico"-computazionale dei "modelli" di intelligenza artificiale. Oltre a ChatGPT, rientrano nel profilo lessicale del micro-cluster i già menzionati "copilot" e "perplexity"*, "claude" e "deepseek"*. Emergono, inoltre, le "difficoltà" legate alle piattaforme AI e la tendenza, da parte dei redditors, a confrontarne le prestazioni ("complesso", "errore", "migliore").

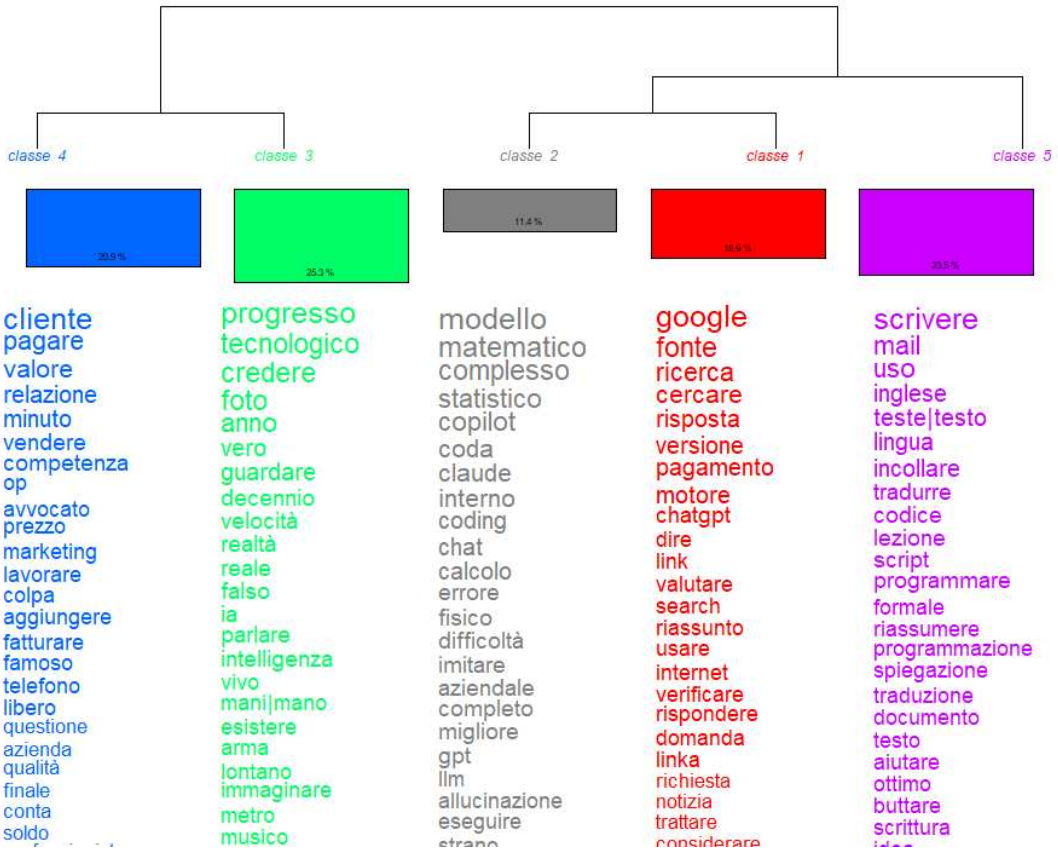

Figura 27 Anteprima dei profili lessicali dei micro-cluster emersi dalla DHC condotta sul subcorpus *sub3_app*.

All'interno del subcorpus dedicato all'ambito applicativo dell'AI, si identificano inoltre due nuclei tematici che richiamano il cluster associato alla dimensione lavorativa nell'analisi del corpus complessivo. Tali risultati suggeriscono che, nel discorso pubblico generale, la dimensione lavorativa e quella applicativa tendano ad essere trattate

distintamente (cluster 5 e cluster 1 in *Figura 17*); contrariamente, nel discorso pubblico specificamente incentrato sulle applicazioni, la dimensione lavorativa costituisce uno degli ambiti entro cui vengono discussi l'utilizzo e le funzionalità dell'intelligenza artificiale. In particolare, sebbene il nucleo semantico del micro-cluster 4 sia chiaramente riconducibile al lavoro, il suo profilo lessicale non rimanda direttamente all'ambito occupazionale, bensì ai potenziali dilemmi etici legati all'utilizzo dell'intelligenza artificiale nell'esercizio della propria attività professionale. Lemmi come "cliente", "pagare" e "vendere" si accostano difatti a lemmi come "valore", "relazione", "competenza" e "colpa". Il micro-cluster 3 si incentra invece sulla stessa intelligenza artificiale in quanto manifestazione dell'odierno "progresso tecnologico" (locuzione che occorre 11 volte all'interno del subcorpus), discusso tra i redditors come fonte di "preoccupazione"*.

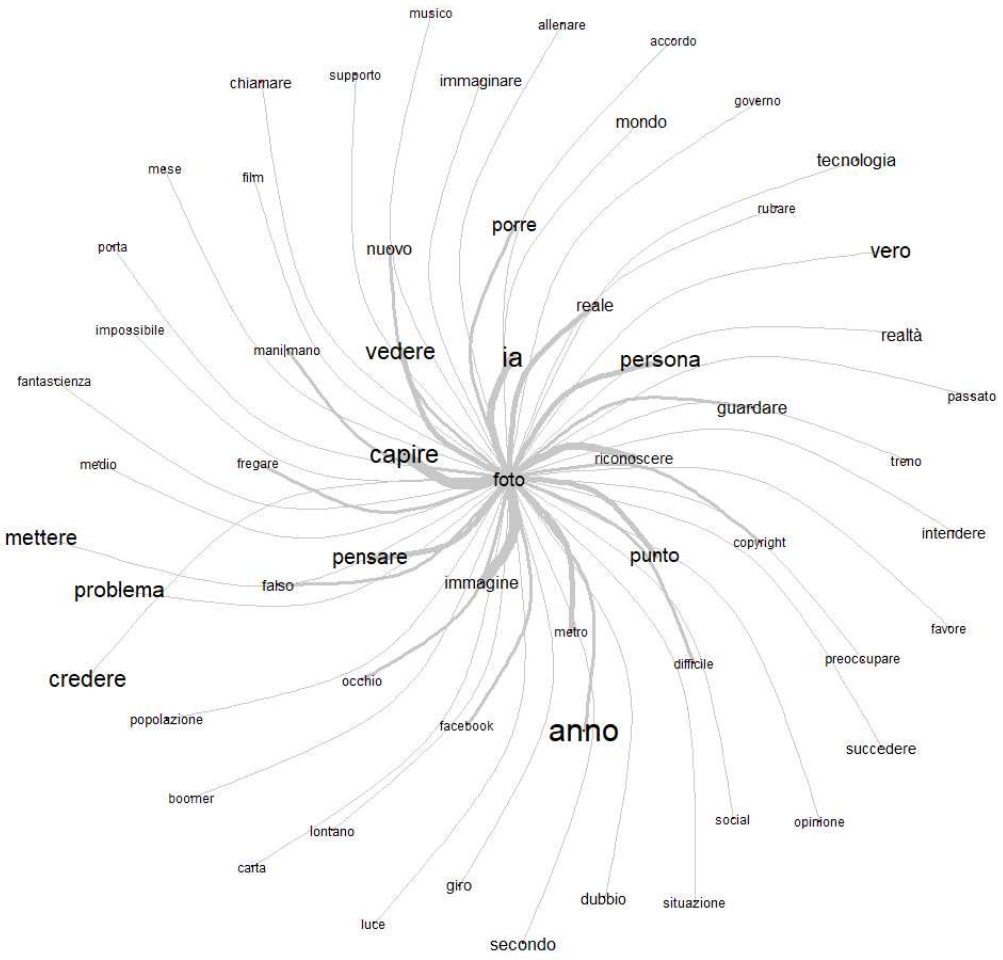

Figura 28 Rappresentazione grafica delle co-occorrenze del lemma "foto". Si includono i lemmi con frequenza maggiore o uguale a 5. Lo spessore dei rami è direttamente proporzionale al punteggio di co-occorrenza delle parole, mentre la dimensione delle parole riflette la loro frequenza all'interno del cluster.

Un esempio comunemente riportato sono le "foto" generate artificialmente, che, ormai indistinguibili da quelle "reali", mettono in crisi i confini tra il "vero" e il "falso". Il grafico in *Figura 28* illustra le co-occorrenze tra i lemmi appartenenti alla classe 3 a partire dal lemma "foto" ed evidenzia l'elevata associazione del termine, oltre che con l'acronimo "ia", con il verbo "riconoscere" e gli aggettivi "falso", "reale" e "difficile". Emerge, inoltre, la co-occorrenza con il lemma "copyright", ad evidenziare il dibattito sulla legittimità della raccolta e dell'utilizzo dei dati per l'addestramento dei *large-language models*.

Le classi semantiche 3 e 4, unite dallo stesso ramo, rappresentano dunque la componente socio-evolutiva associata all'utilizzo dell'AI, all'interno della quale emergono riflessioni sulle conseguenze etiche e sui processi di trasformazione che tale utilizzo comporta. Contrariamente, il grappolo opposto, costituito dal cluster 5 e dai sub-cluster 1 e 2, ne evidenzia invece l'aspetto tecnico-applicativo in una prospettiva più pragmatica.

4.2.4 Istruzione: l'*executive help-seeking* e l'indebolimento del pensiero critico

Il cluster 3, uno dei primi ad essere prodotti dalla DHC condotta sul corpus complessivo, costituisce una classe tematicamente ben definita, confermando l'istruzione come uno dei casi d'uso dell'AI maggiormente discussi. I già citati Chatterji et al. (2025) confermano tale dato, dimostrando come, su ChatGPT, il 10,2% dei messaggi degli utenti riguardi richieste di tutoraggio o insegnamento. In una chiara corrispondenza con il subcorpus relativo al contesto scolastico e universitario (*sub2_edu*), il cluster 3 include il 12,1% dei segmenti testuali analizzati e presenta come lemmi maggiormente caratterizzanti "studente", "esame", "compito", "imparare", "scuola", "studiare", con valori del chi-quadrato compresi tra 545.2 e 250.41.

Un'osservazione più approfondita del profilo lessicale del cluster 3 mette in evidenza la compresenza di due binomi concettuali. Da un lato, risaltano due diversi ambiti educativi, quello universitario ("esame", "università", "sbobine"*) e quello scolastico ("compito", "scuola", "classe", "interrogazione", "verifica", "liceo"*), con una maggiore predominanza di quest'ultimo. Si tratta di un dato che testimonia come la discussione pubblica sull'impiego dell'AI riguardi tanto i giovani adulti universitari quanto gli

adolescenti. Vale la pena precisare, inoltre, che la prevalenza dei lemmi afferenti al contesto scolastico potrebbe essere attribuibile all'età media degli utenti coinvolti nelle discussioni: in un confronto con i metodi di studio precedenti, i redditors, la cui età media è compresa tra i 18 e i 29 anni (Pew Research Center 2025), richiamano più frequentemente l'esperienza condivisa della scuola antecedente all'introduzione dell'AI generativa[80], piuttosto che la più recente o imminente esperienza universitaria.

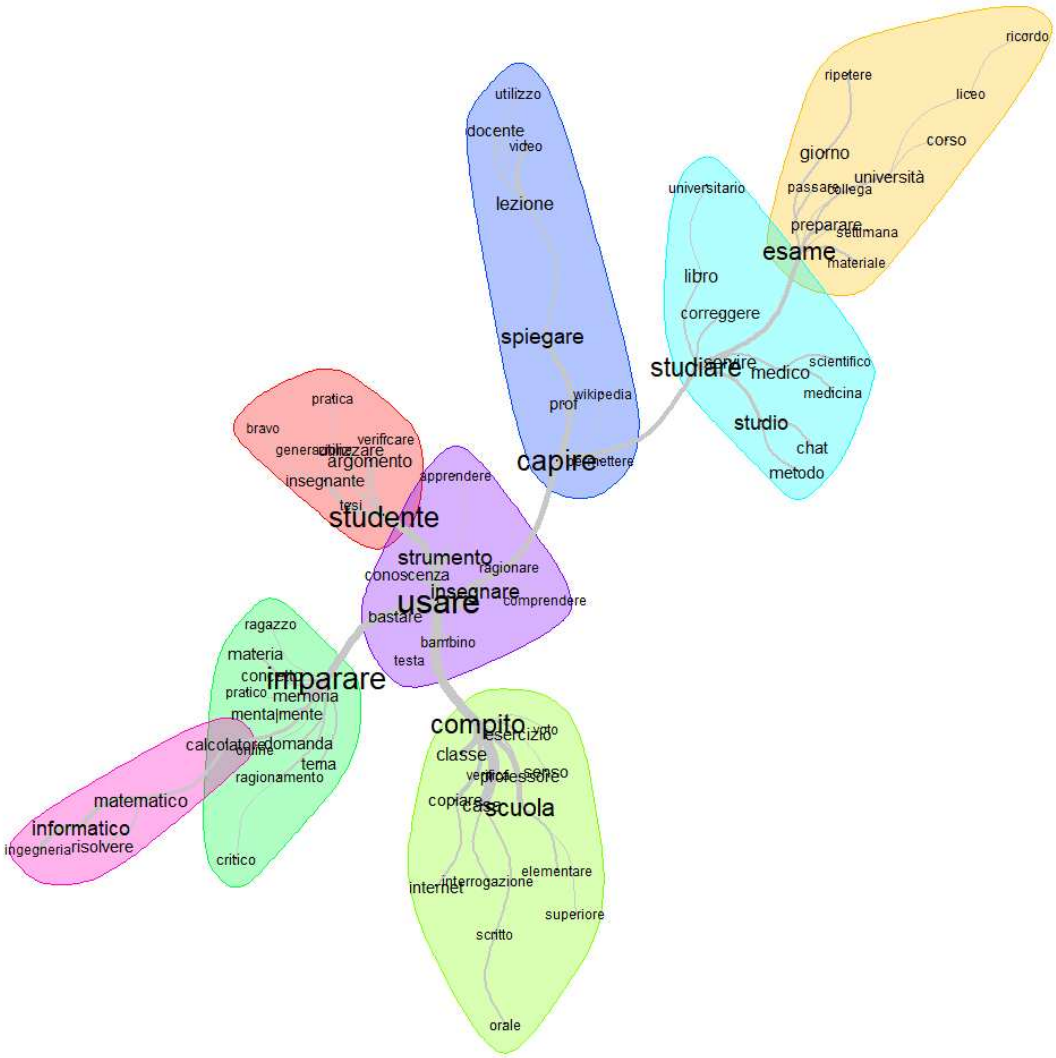

Figura 29 Grafico del macro-cluster 3, identificato come classe semantica dell'educazione e dell'istruzione. Si includono i lemmi con una frequenza maggiore o uguale a 15. Le linee di collegamento tra i nodi illustrano la frequenza con cui i lemmi co-occorrono: maggiore è lo spessore della linea, maggiore è l'associazione tra i lemmi.

[80] Tra le forme caratterizzanti del cluster 1 in *Figura 30*, che mostra la DHC condotta sul subcorpus tematico dedicato all'ambito scolastico (*sub2_edu*), compare infatti il lemma "enciclopedia", spesso utilizzato per fare riferimento ai metodi di apprendimento tradizionali.

Dall'altro lato, si distinguono i due principali ruoli all'interno del contesto educativo, quello dello "studente" e quello dell'"insegnante": l'AI generativa non si limita a trasformare solo il "metodo" di apprendimento degli studenti ("imparare", "studiare", "copiare"), ma condiziona profondamente anche l'approccio di "docenti" o "professori" all'insegnamento ("insegnare"). Un recente report dell'USC Center for Generative AI and Society (Aguilar et al. 2025) ha fornito dati rilevanti sul modo in cui studenti e professori di tutto il mondo si stanno adattando all'intelligenza artificiale. Lo studio, che guarda separatamente ai due ruoli, ha dimostrato come gli studenti universitari facciano degli strumenti di GenAI (ChatGPT, Claude, Copilot) un uso più esecutivo (*executive help-seeking*) che strumentale (*instrumental help-seeking*): l'intelligenza artificiale è, in altre parole, maggiormente impiegata "[…] to get direct answers with minimal effort" (*ivi*, p. 5)[81] piuttosto che "[…] to better understand a concept or process" (*ibid.*)[82].

Allo scopo di delimitare l'analisi al campo semantico dell'istruzione e dell'educazione e coglierne pertanto le sfumature interne, anche nel caso del macro-cluster 3 è stata condotta una seconda DHC sul subcorpus tematico di riferimento (si veda *Figura 30*). In linea con i risultati riportati dall'USC Center, nella classificazione dei segmenti testuali contenuti in *sub2_edu* la classe 1 risulta fortemente caratterizzata dai lemmi "copia" ($\chi^2 = 63.94$) e "incollare" ($\chi^2 = 52.1$). Le due forme si susseguono spesso in espressioni come "copia e incolla" o "copia-incollando", evidenziando l'uso passivo degli strumenti di intelligenza artificiale e il supporto alla scrittura come impiego principale ("testo"). Contrariamente, "ragionare" e "critico"* (quest'ultimo spesso preceduto da "pensiero"* o "spirito"*), compaiono solo successivamente nel profilo lessicale della classe, con un valore del chi2 rispettivamente pari a 22.6 e 16.13. I lemmi, che rimandano all'utilizzo dell'AI come strumento per comprendere e approfondire, sono spesso al centro di discussioni focalizzate sulla legittimità della sua adozione in ambito scolastico: la capacità di ragionare in modo autonomo e indipendente da parte dello studente è prerogativa essenziale per un corretto utilizzo della GenAI.

Secondo Aguilar et al. (2025), l'utilizzo dell'AI da parte degli studenti universitari va oltre le semplici richieste di miglioramento della scrittura, includendo la risoluzione di complessi problemi di programmazione mediante codice prodotto dall'AI. Tra le forme

[81] "[…] per ottenere risposte dirette con il minimo sforzo".
[82] "[…] per comprendere meglio un concetto o un processo".

caratterizzanti del micro-cluster 1, il lemma “risolvere” ne sottolinea l’impiego nella risoluzione dei compiti, ma non emerge in modo rilevante in relazione a problemi di programmazione: la forma “programmazione” compare come secondo lemma più rappresentativo (χ^2 = 83.43) del sub-cluster 2 e risulta prevalentemente associata alla stesura del codice, che viene tuttavia spesso descritta come inefficace. A seguire, con riferimento allo stesso sub-cluster, i lemmi “documento”, “paper”, “revisione” e “report” suggeriscono l’uso dell’AI come supporto alla redazione di elaborati accademici.

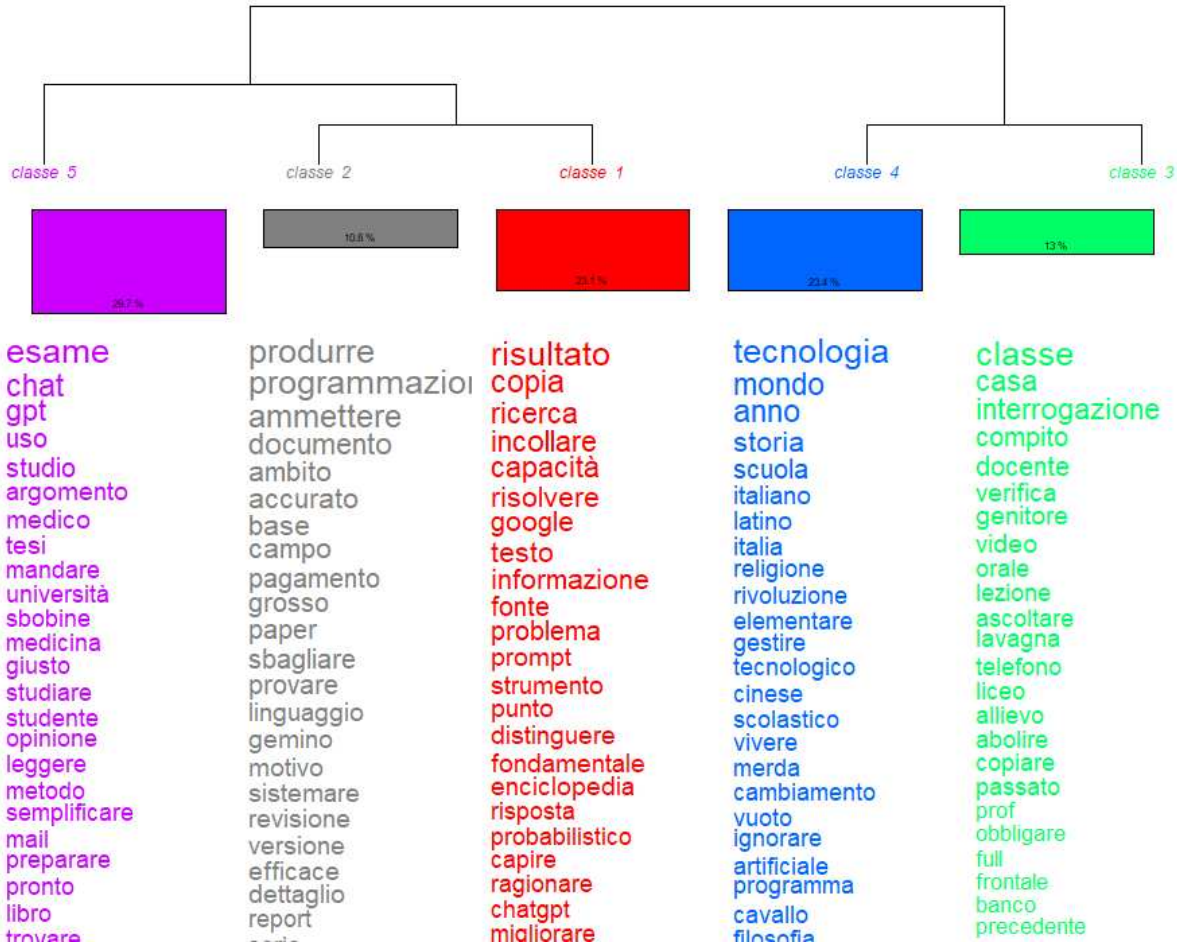

Figura 30 Anteprima dei profili lessicali dei micro-cluster emersi dalla DHC condotta sul subcorpus *sub2_edu*.

I sub-cluster 1 e 2 costituiscono l’ultima iterazione della DHC condotta su *sub2_edu*: essi sono il risultato della biforcazione del ramo condiviso con la classe 5, con la quale presentano dunque una somiglianza semantica. Nella prima fase della DHC, il subcorpus viene difatti suddiviso in due rami principali: dall’uno ha origine la classe 5, dall’altro

hanno origine le classi 2 e 4. Esaminando i profili lessicali delle tre classi, le prime ad essere prodotte, si riconosce il primo binomio osservato nell'analisi del corpus complessivo (si veda *Figura 17*). Da un lato, i lemmi "esame", "tesi", "università" e "sbobine" del cluster 5 delineano il nucleo tematico dell'università, interpretazione supportata dai sub-cluster 1 e 2, che, nidificati nel ramo comprendente anche il cluster 5, pongono l'accento sugli universitari raggruppandone sia gli usi sia le opinioni. È interessante notare come, tra le parole più rappresentative del cluster 5, occorrano anche le forme "medico" e "medicina", risultanti da uno specifico thread di discussione: il confronto tra uno studente di medicina, avvezzo all'uso di ChatGPT per il superamento degli esami, e gli altri utenti della community solleva riflessioni di natura etica. Dall'altro lato, i cluster 3 e 4 sono invece afferenti al nucleo tematico della scuola. Oltre ai più generici lemmi "classe", "interrogazione" e "compito", la classe 3 racchiude il secondo binomio oppositivo osservato in *Figura 17*, vale a dire l'antonimia tra studente ("interrogazione", "ascoltare", "allievo") e insegnante ("docente", "prof"): la tematica prevalente riguarda l'uso dell'AI per lo svolgimento dei "compiti" a "casa", mentre le "interrogazioni"-"orali" vengono percepite come il più affidabile metodo di valutazione per i docenti. Di primo acchito, la classe 4 si presenta invece come un raggruppamento di materie scolastiche ("tecnologia", "storia", "italiano", "latino", "religione"), ma lemmi come "mondo", "italia", "rivoluzione", "vivere", "cambiamento" e "vuoto" ne evidenziano la forte impronta sociale: il cluster raccoglie le riflessioni dei redditors sulla tendenza, nello scenario italiano, a temere il cambiamento ed eludere la rivoluzione, facendosi rappresentativo di quella porzione di comunità favorevole all'adozione delle nuove tecnologie in ambito educativo.

5. Il complesso sistema di relazioni di somiglianza e differenza

5.1 La distribuzione dei testi sul piano cartesiano

Il metodo Reinert così applicato ha consentito di identificare, in primo luogo, le classi semantiche che dominano il corpus nella sua interezza. Grazie alle analisi secondarie, sono stati successivamente estratti i contenuti ricorrenti nelle singole raccolte testuali: circoscrivendo l'analisi ai singoli subcorpora, la *Descending Hierarchical Classification* ha individuato gli argomenti e i temi che attraversano e caratterizzano ciascuna macro-tematica. Avendo ampiamente discusso i profili lessicali dei macro-clusters generati, è possibile ora concentrarsi sul modo in cui tali profili si rapportano tra di loro.

Il dendogramma in *Figura 16* illustra l'ordine gerarchico di produzione dei clusters e fornisce pertanto una prima rappresentazione delle relazioni di similarità e dissimilarità. Ciononostante, è l'analisi delle corrispondenze a chiarire il complesso sistema di relazioni sotteso tra i profili lessicali. Nella sua definizione classica, la *Correspondence Analysis* (di seguito anche indicata come CA) consente difatti di visualizzare su un piano cartesiano il complesso sistema di relazioni presenti tanto tra i testi che compongono un corpus quanto tra i termini ad essi associati. Come anticipato (*§2.4.2*), tale mappatura ha origine da un *Term-Document Matrix* che incrocia i *terms* contenuti nel corpus (il vocabolario *V*), distribuiti sulle righe, con i *documents* da cui è costituito (i testi *M*), distribuiti sulle colonne. In linea con l'approccio confermativo del presente lavoro, i seguenti paragrafi confronteranno due rappresentazioni grafiche che differiscono nella tipologia di testi[83] distribuiti sulle colonne della suddetta tabella di contingenza $V \times M$. La prima rappresentazione (si veda *Figura 31*), ottenuta tramite il software RStudio, è il risultato di una matrice che incrocia il vocabolario del corpus con i subcorpora che lo compongono; in una seconda rappresentazione (si veda *Figura 32*), invece, ai subcorpora si sostituiscono i macro-clusters generati dalla *Descending Hierarchical Classification* condotta su IRaMuTeQ e illustrata in *Figura 16*. Considerata la tradizionale matrice "parole × testi", si passa pertanto da una matrice "parole × subcorpora" a una matrice "parole × clusters", mantenendo

[83] È tuttavia necessario precisare che esiste una differenza anche in merito alla variabile *V* distribuita sulle righe: nel primo caso il vocabolario è costituito da forme flesse (così come estratte dai commenti Reddit), mentre nel secondo da forme ridotte, che hanno cioè subito il processo di lemmatizzazione previsto da IRaMuTeQ.

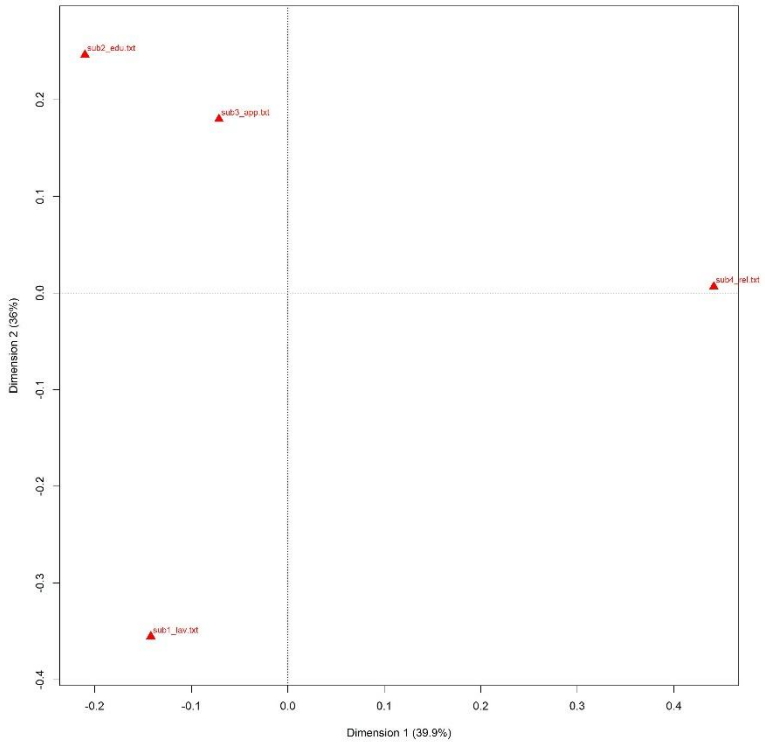

Figura 31 CA risultante dalla matrice che incrocia il vocabolario del corpus con i subcorpora definiti a priori.

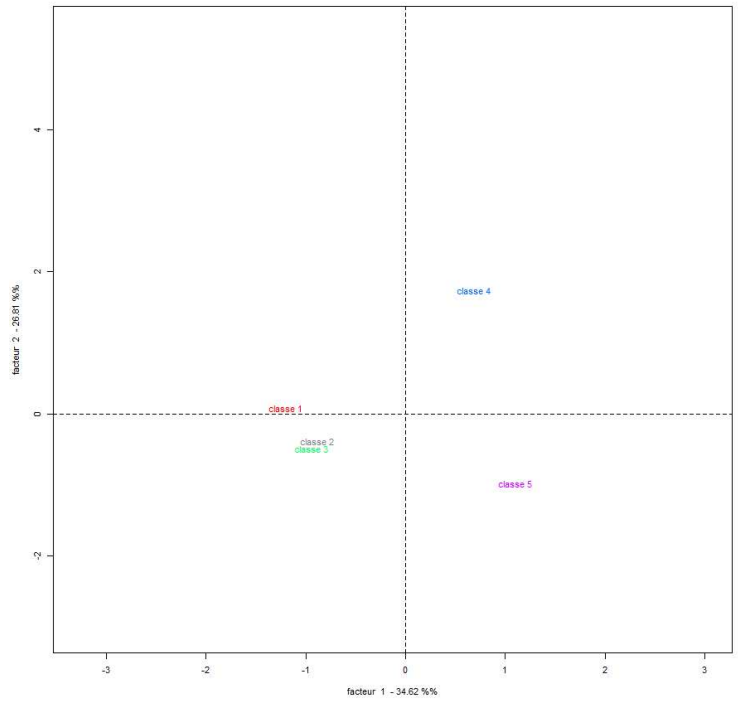

Figura 32 CA risultante dalla matrice che incrocia il vocabolario del corpus con i clusters prodotti empiricamente tramite DHC.

invariato il vocabolario *V* distribuito sulle righe, ma modificando l'unità testuale *M* distribuita sulle colonne. È stato precedentemente discusso come l'analisi delle corrispondenze riduca lo spazio multidimensionale prodotto dalle matrici di distanza chi-quadrato $V \times V$ e $M \times M$ a uno spazio costituito esclusivamente dalle prime due dimensioni, al fine di riassumere le relazioni più importanti e favorire l'interpretazione dei risultati. Il piano cartesiano costruito con la prima e con la seconda dimensione è difatti in grado di raccogliere la quota di inerzia più elevata e dunque di rappresentare al meglio la struttura di associazione complessiva (Tuzzi 2024). In *Figura 31* e *32*, i due assi da cui hanno origine i due distinti piani cartesiani raccolgono una quota di inerzia pari al 75,9% e al 61,4%, rispettivamente.

5.1.1 L'analisi delle corrispondenze sui subcorpora definiti a priori

Nel caso in esame, la distribuzione di testi – siano essi i subcorpora tematici o le classi semantiche – e parole sui quattro quadranti del piano bidimensionale offre, in primo luogo, una migliore comprensione dei legami tra le macro-tematiche individuate, mettendone in rilievo le somiglianze e le differenze lessicali. Una prima esaminazione del piano cartesiano consente di valutare la somiglianza lessicale tra testi e parole sulla base della loro appartenenza allo stesso quadrante (Tuzzi 2024). Nella CA condotta sui subcorpora definiti a priori (si veda *Figura 31*), si evidenzia la somiglianza lessicale tra *sub3_app* e *sub2_edu*, entrambi appartenenti al secondo quadrante del piano. I due subcorpora risultano al contempo i più vicini all'origine degli assi. Ciò risulta particolarmente evidente nel caso di *sub3_app*: la raccolta di commenti tematicamente incentrati sulla varietà di impiego dell'intelligenza artificiale non presenta un lessico particolarmente distintivo rispetto al resto del corpus. È stato anticipatamente ipotizzato come l'AI sia ormai talmente pervasiva nella vita quotidiana degli utenti che le discussioni sulle sue applicazioni pratiche ricorrono con regolarità all'interno dei commenti online, a prescindere dalla macro-tematica delle submission. La CA, ponendo *sub3_app* in prossimità dell'origine, conferma la considerevole trasversalità del tema e, di conseguenza, il lessico medio del subcorpus, perlopiù privo di specificità lessicali di interesse.

Contrariamente, *sub1_lav* e *sub4_rel* rappresentano i subcorpora più distanti dall'origine, mostrando profili lessicali più caratterizzanti rispetto al profilo medio del

corpus. Collocandosi nel primo quadrante del piano, *sub4_rel* condivide con *sub2_edu* e *sub3_app* il semipiano alto, mentre *sub1_lav*, che si posiziona nel terzo quadrante, condivide con *sub2_edu* e *sub3_app* il semipiano sinistro. Da un lato, dunque, i subcorpora focalizzati sull'uso esecutivo ed operativo dell'AI (*sub2_edu*, *sub3_app*) e sulle ansie economico-occupazionali ad essa legate (*sub1_lav*) sono graficamente discostati dal subcorpus che ne discute i benefici e rischi a livello psicologico e relazionale (*sub4_rel*): il primo asse (*Dimension 1*) sembra pertanto definire una dimensione che va dal funzionale al relazionale, ribadendo l'opposizione tra l'oggetto tecnologico e chi ne fruisce. Dall'altro lato, il secondo asse (*Dimension 2*) suggerisce una dimensione che va dall'economico al sociale: *sub1_lav*, il subcorpus che tematizza l'incertezza economica, si contrappone sia a *sub4_rel*, che costituisce la massima espressione della sfera socio-relazionale, sia a *sub2_edu* e *sub3_app*, nei quali la discussione sull'uso pragmatico dell'AI pone l'accento sui contesti sociali di utilizzo (*i.e.*, la scuola e, in termini generali, la vita quotidiana) piuttosto che sulle potenziali conseguenze economiche di lungo periodo.

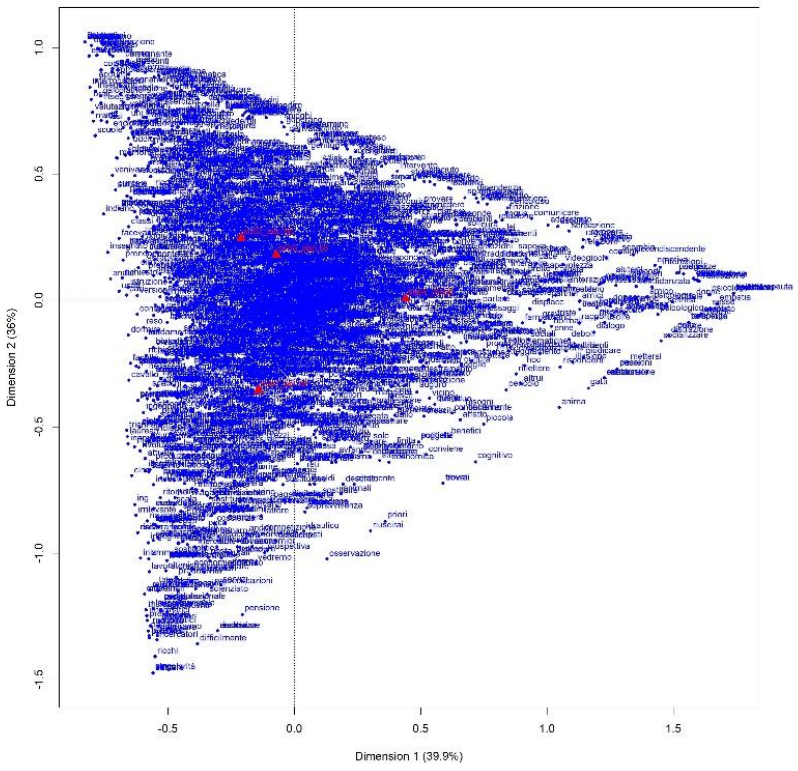

Figura 33 CA condotta sui subcorpora definiti a priori, considerando i *word-types* con frequenza maggiore o uguale a 8.

La *Figura 33* riporta i risultati della *Correspondence Analysis* condotta sui *word-types* con frequenza maggiore o uguale a 8, la soglia di frequenza stabilita al fine di garantire la rappresentatività della quota di forme selezionate per svolgere l'analisi rispetto alla totalità delle forme del corpus (cfr. *§3.1.1*). I profili lessicali dei 2.963 *word-types* più ricorrenti vengono così proiettati come punti sul piano cartesiano, fornendo una rappresentazione più dettagliata del modo in cui i quattro subcorpora si rapportono tra loro. Si osserva come i termini si concentrino in una posizione centrale dello spazio bidimensionale, dalla quale si estendono verso tre punti periferici: il secondo e il terzo quadrante, su cui si collocano *sub2_edu*, *sub3_app* e *sub1_lav*, sono caratterizzati da una distribuzione verticale e più compatta dei termini, mentre il semipiano destro si distingue per un maggior grado di dispersione, confermando l'autonomia di *sub4_rel*. Il subcorpus focalizzato sull'antropomorfizzazione dell'intelligenza artificiale presenta pertanto un profilo lessicale più specifico rispetto alla media del corpus.

5.1.2 L'analisi delle corrispondenze sulle classi semantiche definite empiricamente

Il capitolo precedente ha ampiamento trattato l'ordine gerarchico di produzione dei clusters tramite applicazione del metodo Reinert: le iterazioni della *Descending Hierarchical Classification* hanno distinto le classi 4 e 5 dalla classe 3, dal cui ramo ha infine avuto origine un sottoinsieme costituito dai sub-clusters 1 e 2. In *Figura 32*, la CA condotta sulle cinque classi semantiche fornisce un'interpretazione complementare alla DHC: avendo individuato le diverse classi semantiche, l'analisi delle corrispondenze consente di visualizzarne i profili lessicali nello spazio bidimensionale e di renderne chiaramente visibili le relazioni di similarità o dissimilarità.

La disposizione dei cinque clusters in cui il corpus è stato suddiviso empiricamente riflette in larga parte quella dei quattro subcorpora definiti a priori. Le classi 4 e 5, tematicamente corrispondenti a *sub 4_rel* e *sub1_lav*, si confermano clusters a sé stanti, caratterizzati da profili lessicali significativamente diversi tra loro e rispetto a quelli delle altre classi: la classe 4, che occupa il primo quadrante del piano cartesiano, e la classe 5, che si colloca nel quarto quadrante, costituiscono le classi semantiche più distanti dall'origine degli assi e, pertanto, portatrici di un lessico più distintivo rispetto al profilo lessicale medio del corpus. L'analisi delle corrispondenze spiega dunque il motivo per cui nella DHC le classi 4 e 5 condividono lo stesso ramo del dendogramma,

distinguendosi dal secondo insieme di clusters: la loro correlazione non è tanto riconducibile a una forte somiglianza lessicale – dal momento che la CA evidenzia profili lessicali distinti – quanto piuttosto alla loro comune distanza rispetto al cluster 3 e ai sub-clusters 1 e 2.

Per contro, la proiezione sullo spazio bidimensionale delle classi 1, 2 e 3 riflette pienamente la struttura individuata dalla DHC: il fatto che i sub-clusters 1 e 2 derivino dalla medesima biforcazione da cui ha origine il cluster 3 è indice della reale similarità dei profili lessicali delle classi, la quale risulta naturalmente marcata tra i due sub-cluster. In modo analogo a *sub3_app* e *sub2_edu*, le classi 1 e 3 si collocano difatti in prossimità dell'origine: in *Figura 32*, il cluster delle applicazioni quotidiane e il cluster dell'istruzione ribadiscono la loro vicinanza lessicale, pur non condividendo esattamente il medesimo quadrante. Si osserva, difatti, come la classe 1 coincida con il primo asse. La neutralità e la similarità dei profili delle classi semantiche 1 e 3 sono in realtà accentuate – rispetto a quelle degli equivalenti *sub3_app* e *sub2_edu* in *Figura 31* – dall'identificazione della già citata classe 2, che il metodo Reinert identifica come un'area tematica aggiuntiva di presentazione della GenAI. Come anticipatamente discusso, la sua introduzione rivela la presenza, all'interno del dialogo tra i redditors, di sezioni discorsive autonome incentrate sugli aspetti tecnici e di funzionamento dell'intelligenza artificiale generativa, ossia sulla GenAI in quanto modello probabilistico addestrato su enormi quantità di dati testuali. L'applicazione del metodo Reinert isola il suddetto lessico tecnico, considerevolmente presente laddove il tema centrale delle submission è l'istruzione (cluster 3) e, in particolare, l'uso applicativo dell'AI nel quotidiano (sub-cluster 1), mentre meno pervasivo nei commenti di submission incentrate sull'ambito lavorativo (cluster 5). Ciò comporta, da un lato, un maggiore avvicinamento tra le classi 1 e 3 (rispetto a *sub3_app* e *sub2_edu*) – con le quali la classe 2 mostra una sovrapposizione rilevante – e, dall'altro, un maggiore avvicinamento tra le classi 4 e 5 (rispetto a *sub4_rel* e *sub1_lav*). Se in *Figura 31* il tema del lavoro e il tema del benessere psico-relazionale si oppongono lungo ciascuno dei due assi, in *Figura 32* essi condividono un semipiano, ponendosi entrambi sul polo positivo del primo asse. Ne consegue una ridefinizione delle dimensioni rappresentate dagli assi: il secondo asse (*facteur 2*) ribadisce una dimensione economico-relazionale, contrapponendo in particolare il tema del lavoro (*classe 5*), che verte sulle suddette preoccupazioni

economico-occupazionali, a quello delle relazioni interpersonali e del benessere psicologico (*classe 4*); il primo asse definisce invece una dimensione che contrappone nuovamente il polo semantico dell'oggetto tecnologico al polo semantico dei soggetti che ne fruiscono, confermandola come dimensione funzionale-relazionale. Si può concludere che la principale differenza tra le due mappature risiede nel fatto che, lungo quest'ultima dimensione, la CA prodotta a partire dai dati empirici non colloca il cluster del lavoro sul polo oggettivo, bensì su quello soggettivo: contrariamente a *sub1_lav*, una volta distaccato dal lessico tecnico anche il cluster 5 emerge come classe semantica attinente all'impatto sociale dell'AI. Il semipiano destro tematizza, pertanto, sia i rischi e i benefici dell'integrazione tra AI e relazioni interpersonali (*classe 4*) sia le preoccupazioni dei redditors in merito all'automazione e alla redistribuzione economica (*classe 5*). Nel semipiano sinistro si proiettano invece le classi 1, 2 e 3, i cluster per l'appunto caratterizzati da profili in cui predomina il lessico relativo al funzionamento tecnico dell'intelligenza artificiale.

5.2 Un confronto ravvicinato: le forme discriminanti

È stato anticipatamente discusso come risultato finale di un'analisi delle corrispondenze sia la traduzione grafica della distanza chi-quadrato calcolata per tutte le coppie di parole e per tutte le coppie di testi in esame (cfr. *§2.4.2*). Una volta restituite le relazioni sottese tra i testi (tra i quattro subcorpora in *Figura 31* e tra le cinque classi semantiche in *Figura 32*), è doveroso esaminare i profili lessicali delle parole, che assumono anch'essi la forma di coordinate proiettate come punti sul piano bidimensionale. I grafici sottostanti arricchiscono, dunque, l'interpretazione dell'analisi visualizzando le relazioni di somiglianza e differenza presenti non solo tra i testi, ma anche tra le parole e tra testi e parole. La *Figura 34* mostra il sistema di relazioni tra le forme caratteristiche di ciascun subcorpus, mentre la *Figura 35* illustra il sistema di relazioni tra i lemmi caratteristici di ciascun cluster[84]. Al fine di favorire la leggibilità e l'interpretabilità dei dati, nella CA condotta sui subcorpora definiti a priori si è scelto di visualizzare esclusivamente le forme

[84] Si ricorda che le parole grammaticali, poiché naturalmente distribuite in modo uniforme lungo il corpus, non possiedono alcuna forza discriminante. Di conseguenza, non influenzando la formazione delle classi, non vengono visualizzate nell'analisi delle corrispondenze prodotta a partire dalla DHC.

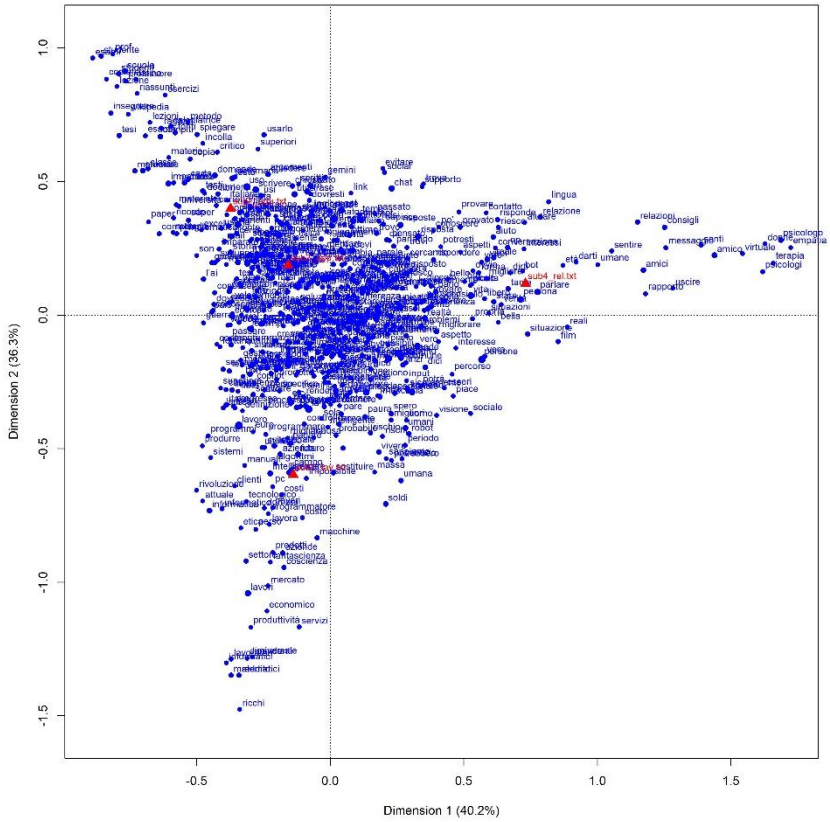

Figura 34 CA raffigurante il sistema di relazioni tra le forme caratteristiche (frequenza maggiore o uguale a 30) di ciascun subcorpus.

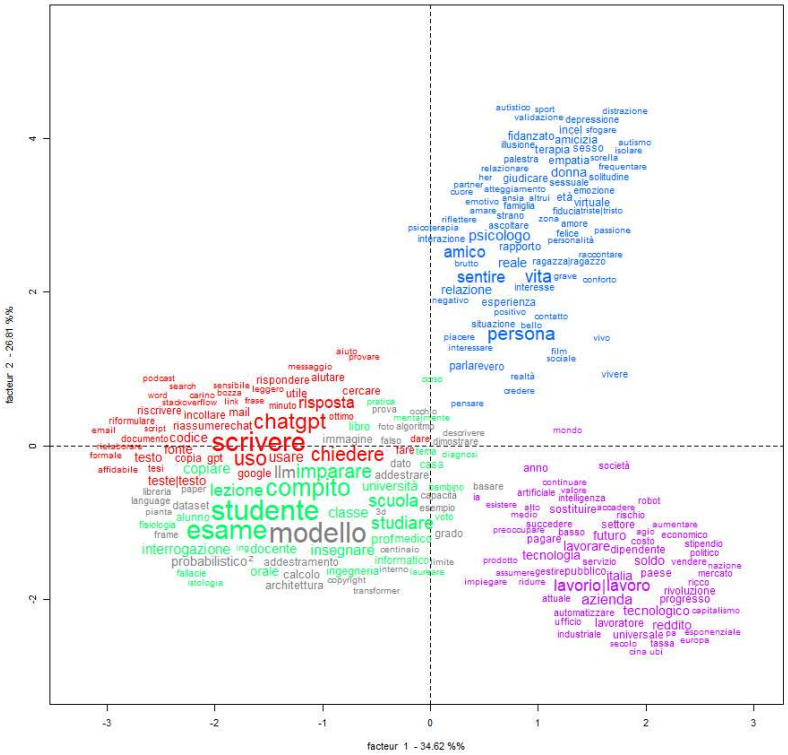

Figura 35 CA raffigurante il sistema di relazioni tra i lemmi caratteristici di ciascun cluster.

con frequenza maggiore o uguale a 30. Inoltre, si garantisce la coerenza del confronto considerando unicamente le parole di contenuto.

Per interpretare correttamente la distribuzione grafica delle parole, sono necessarie alcune precisazioni. È innanzitutto opportuno ribadire che nella *Correspondence Analysis* la distanza (o la prossimità) che intercorre tra le forme riflette il grado di dissimilarità (o di somiglianza) dei loro profili lessicali: maggiore è la divergenza tra i profili lessicali delle parole, maggiore è la distanza grafica presente tra queste sul piano bidimensionale. In secondo luogo, si ricorda come la posizione assunta da un termine acquisisca significato non solo in relazione a quella assunta da tutti gli altri termini, ma anche in relazione all'origine degli assi: un termine che si colloca lontano dall'origine possiede un profilo lessicale fortemente specifico, che cioè contribuisce in modo considerevole a distinguere il testo a cui è associato dagli altri testi. Infine, si precisa come nella CA eseguita su IRaMuTeQ la dimensione dei lemmi dei cinque cluster, separati cromaticamente, è direttamente proporzionale al valore del chi2 ad essi associato: maggiore è la dimensione della parola, maggiore il è grado di sovra-rappresentazione del lemma all'interno del cluster. Un alto valore del chi2 ed una maggiore distanza della parola dall'origine degli assi indicano, rispettivamente, la forte rappresentatività della parola all'interno del cluster a cui è associata e la sua capacità di distinguere tale cluster dagli altri nello spazio bidimensionale. Per chiarire questa distinzione concettuale, si parlerà pertanto di "rappresentatività" nel primo caso e di "forza discriminante" nel secondo.

Alla luce di queste considerazioni, si osserva la distribuzione delle forme sui rispettivi piani cartesiani. In linea con quanto discusso nelle sezioni precedenti, *sub1_lav* e il cluster 5 si distinguono nel raggruppare le preoccupazioni associate alla crescente adozione dell'AI nel contesto lavorativo, raramente espresse nei restanti subcorpora. Osservando le forme più discriminanti poste nel terzo quadrante in *Figura 34* e nel quarto quadrante in *Figura 35*, si osserva come la struttura del discorso definita dalla CA condotta sui subcorpora combaci in parte con quella della CA condotta sui cluster empirici. In entrambe le figure, emergono difatti termini che ne chiariscono l'appartenenza al polo economico. Ciononostante, le due analisi delle corrispondenze mettono in risalto una lieve differenza tra *sub1_lav* e il cluster 5. I *word-types* più discriminanti per *sub1_lav* sembrano riassumere ansie occupazionali e, per l'appunto, di

natura economica: tra i termini più distanti dall'origine degli assi si osservano "ricchi", "reddito", "dipendente", "produttività", "servizi", "economico", "lavori", "mercato", "macchine", "rivoluzione", "costo", "soldi", "sostituire". Tra questi, "matematici" e "informatici" identificano le categorie professionali frequentemente evocate dai redditors. L'elevata specificità di tali termini segnala la capacità discriminante delle professioni tecnico-scientifiche nei commenti online che discutono il potenziale impatto dell'AI sul lavoro. Anche nel cluster 5 si osserva la presenza di numerose fome appartenenti alla sfera economica. Tra i lemmi più periferici risaltano tuttavia lemmi che rimandano alla sfera socio-politica già discussa nelle sezioni precedenti: a "reddito", "lavoratore", "automazione" e "ricco" si affiancano "ubi", "tassa, "europa", "paese", "italia", "progresso" e "politico", che alludono alla ridefinizione della società e alle politiche di redistribuzione economica. Da un lato, il confronto tra i due grafici consente di evidenziare ancora una volta come la crescente automazione e il rischio di disoccupazione siano fortemente distintivi dei commenti online a submission incentrate sul rapporto tra lavoro e intelligenza artificiale. Dall'altro, mette in luce sezioni discorsive che pongono l'impatto dell'automazione su un piano collettivo: la CA condotta sulle classi definite empiricamente non attribuisce la specificità del cluster 5 alle sole ripercussioni sul lavoro individuale, bensì su ciò che il rischio di sostituzione tecnologica implica per le politiche economiche. È interessante infine notare come, contrariamente al cluster 5, tra i termini associati a *sub1_lav* compaia il *word-type* "coscienza", semanticamente distinto dal resto delle forme appartenenti al terzo quadrante. Il termine ribadisce come il subcorpus includa riflessioni sul rapporto tra sfera cognitiva e computazionale, il quale emerge qui con maggior chiarezza rispetto alla CA condotta sulle classi empiriche. Il confronto è richiamato anche dai lemmi "umani" e "robot", due forme caratterizzate da profili lessicali notevolmente simili e collocate in posizione intermedia lungo la dimensione funzionale-relazionale.

Analogamente a *sub1_lav* e al cluster 5, anche nel caso di *sub4_rel* e del cluster 4 si riconosce una parziale corrispondenza tra i profili lessicali dei termini discriminanti, collocati nel primo quadrante dei piani bidimensionali. In particolare, in *Figura 34* i termini che maggiormente contribuiscono a separare *sub4_rel* dai restanti subcorpora identificano i ruoli sociali in cui i chatbot di GenAI vengono antropomorfizzati: il polo relazionale definito dal primo asse è dominato dai *word-types* "amico", "donne",

"psicologi" e "psicologo" (ai quali si aggiunge "terapia"), la cui posizione periferica permette di concludere che l'utilizzo della GenAI come surrogato delle relazioni interpersonali sia una componente discorsiva specifica di *sub4_rel*, poco distintiva per altri subcorpora. I termini, ad esempio, non compaiono con pari evidenza in *sub3_app*, che pure raccoglie discussioni inerenti alle molteplici applicazioni quotidiane dell'AI. La specificità di "virtuale", caratterizzato da un profilo lessicale simile a quello dei suddetti *word-types*, esalta inoltre la componente fittizia dell'interazione. Si distinguono nel primo quadrante in *Figura 34* anche "aiuto", "aiutare", "parlare", "relazioni", "consigli", "messaggi" e "rapporto", ulteriori forme significativamente afferenti alla macro-tematica del subcorpus e, in particolare, alle specifiche richieste di supporto. Le forme fino ad ora analizzate si limitano tuttavia ad esprimere la pragmaticità del ricorso ai chatbot di GenAI, seppur giustificato da bisogni di natura affettiva. La sfera emozionale appare difatti alquanto latente nell'analisi delle corrispondenze condotta sui subcorpora definiti a priori, nella quale un'unica forma richiama esplicitamente la dimensione emotiva: "empatia", l'emozione umana maggiormente ricercata e che i chatbot si mostrano in grado di simulare, costituisce il *word-type* più distintivo per *sub4_rel*. Le emozioni assumono invece particolare rilievo in *Figura 35*, in cui si distinguono lemmi quali "cuore", "emozione", "fiducia", "amare", "ansia", "amore", "felice", "emotivo" e "triste|tristo". Operando un confronto trai due grafici, si nota difatti come la CA condotta sulle classi semantiche definite empiricamente non associ il ricorso ai chatbot esclusivamente alla gestione delle relazioni interpersonali: l'analisi delle corrispondenze in *Figura 35* non attribuisce ai lemmi "amico" e "psicologo" la medesima forza discriminante illustrata in *Figura 34*. Questi, pur costituendo – assieme a "persona", "vita" e "sentire" – i termini maggiormente rappresentativi del cluster 4, riducono la propria distanza dall'origine degli assi. Nel polo relazionale definito dal secondo asse in *Figura 35*, i termini che maggiormente contribuiscono a distinguere il cluster 4 dalle altre classi semantiche sono difatti lemmi che esprimono il disagio emotivo degli utenti. Essi identificano i motivi che sembrano giustificare la ricerca di un supporto artificiale per la regolazione di stati affettivi negativi: "isolare", "sfogare", "depressione", "solitudine" esplicitano l'esperienza personale dell'utente e la presenza di difficoltà personali. La prossimità dei lemmi "autismo", "autistico", "sport", "palestra", "validazione", "incel", "sesso" suggerisce che il disagio emotivo si manifesta in particolare tra coloro che presentano

Disturbi dello Spettro Autistico (*Autism Spectrum Disordsers*) (condizione che determina un maggior isolamento sociale), che manifestano insicurezze legate al proprio aspetto fisico e/o che lamentano un senso di ingiustizia nei confronti della società per il loro insuccesso sentimentale o sessuale. Profili lessicali simili si osservano anche per "terapia", "empatia", "donna" e "virtuale", che distanziandosi nuovamente dall'origine degli assi confermano la loro specificità per il cluster semanticamente incentrato sul supporto emotivo fornito dall'AI.

In *Figura 34*, il secondo quadrante costituisce il terzo punto periferico di estensione dei *word-types*: il polo che definisce la componente funzionale (*Dimension 1*) e sociale (*Dimension 2*) del discorso pubblico sull'intelligenza artificiale è contraddistinto dai termini associati a *sub2_edu* e, in misura minore, da quelli associati a *sub3_app*. Nel caso di *sub2_edu*, le parole che si collocano a maggiore distanza dall'origine identificano chiaramente il campo semantico dell'istruzione: si osservano "prof", "studente", "esami", "professore", "lezione", "riassunti", "esercizi", "tesi", "metodo", "compiti" e "spiegare". I due *word-types* che più contribuiscono alla variazione del grafico, "prof" e "studente", rimandano ai due principali ruoli del contesto educativo, mentre l'antonimia tra l'ambito scolastico e quello universitario emerge con minore evidenza. Relativamente a *sub3_app*, è già stato ampiamente discusso come la prossimità del subcorpus all'origine degli assi sia indice del suo lessico medio. Nei subreddit, parlare degli usi quotidiani dell'intelligenza artificiale significa fare riferimento a una molteplicità di utilizzi pratici, che attraversano diversi ambiti della vita quotidiana. Nella CA condotta sui subcorpora definita a priori, la distribuzione dei punti sul piano bidimensionale corrobora tale ipotesi: *sub3_app* è caratterizzato da *word-types* che non possiedono un profilo lessicale abbastanza discriminante da rendere l'ambito applicativo dell'AI un nucleo tematico e lessicale a sé stante. In *Figura 34*, la bassa specificità delle forme ad esso associate incide difatti sulla leggibilità dei risultati: sebbene si osservino forme quali "gemini", "link", "chat" e "social", chiaramente riconducibili al campo semantico della navigazione in rete, l'elevata trasversalità del tema impedisce una chiara identificazione di ulteriori forme discriminanti, che si amalgamano a quelle dei subcorpora circostanti e, in particolare, a quelle di *sub2_edu*.

D'altro canto, la CA eseguita sulle classi semantiche definite empiricamente offre una rappresentazione grafica che favorisce l'interpretabilità dei risultati. Tuttavia,

contrariamente alle due macro-tematiche espresse da *sub1_lav* e il cluster 5 da un lato e da *sub4_rel* e il cluster 4 dall'altro, le due analisi delle corrispondenze differiscono laddove il discorso verta sull'ambito educativo e sulla varietà di applicazioni quotidiane dell'AI. Se in *Figura 34* il terzo punto periferico di estensione delle forme coincide con *sub2_edu*, in *Figura 35* coincide con la commistione di tre diversi cluster. Tale differenza è da attribuire, per l'appunto, all'introduzione di un quinto nucleo tematico, rappresentato dal cluster 2: nonostante i confini tra i cluster 1 e 3 siano ancora visibili, le forme associate all'aspetto tecnico e di funzionamento dell'AI (cluster 2, in grigio) mostrano profili lessicali marcatamente simili a quelli delle forme associate al contesto applicativo (cluster 1, in rosso) e, in particolare, all'istruzione (cluster 3, in verde). I profili lessicali di lemmi appartenenti alle tre classi semantiche del semipiano sinistro mostrano difatti una somiglianza tale da amalgamarsi in un unico blocco lessicale, chiaramente contrapposto alla specificità espressa, singolarmente, dai profili lessicali dei lemmi associati ai cluster 4 e 5 del semipiano destro. Ciò segnala la pervasività, nei cluster 1 e 3, di un lessico tecnico-funzionale che viene invece a mancare nei cluster 4 e 5. Contrariamente a quanto ci si potrebbe attendere, la classe del funzionamento tecnico (cluster 2) mostra maggiore somiglianza alla classe semantica dell'istruzione (cluster 3) piuttosto che a quella degli usi quotidiani (cluster 2). Tale distribuzione consente di giugnere ad una seconda conclusione che al contempo chiarisce la vicinanza tra *sub2_edu* e *sub3_app* illustrata in *Figura 31*: i commenti online che discutono l'integrazione dell'AI nei processi di apprendimento e di insegnamento sono i commenti in cui emergono con maggiore evidenza sezioni discorsive inerenti al funzionamento computazionale dei modelli di intelligenza artificiale. I lemmi "modello", "probabilistico", "calcolo", "dato", "llm", "generare", associati al cluster 2, si accostano a "studente", "esame", "compito", "imparare", "studiare", "scuola" del cluster 3. L'analisi condotta sulle classi semantiche funge pertanto da strumento interpretativo per quella eseguita sui subcorpora, poiché permette di concludere che il lessico condiviso da *sub2_edu* e *sub3_app* – e che ne giustifica la prossimità sul piano bidimensionale in *Figura 31* – è un lessico di tipo tecnico-funzionale: le discussioni che si sviluppano a partire da submission incentrate sull'istruzione non si limitano a trattare il contesto scolastico e universitario in quanto tale, ma includono riflessioni sul funzionamento dell'intelligenza artificiale (come suggerito dalla commistione del cluster 2 nel cluster 3) e, secondariamente, sulla

molteplicità di usi e applicazioni a cui essa si presta. A conferma di quest'ultima interpretazione, si evidenzia come la disposizione dei lemmi appartenenti ai cluster 1 e 3 in *Figura 35* sembri segnalare le principali modalità di impiego dell'intelligenza artificiale da parte degli studenti: "scrivere", "uso", "fonte", "mail", "codice", "scrittura", associati al cluster 1, si collocano in prossimità di "spiegare", "imparare", "copiare", "tema", "università", "esame", "verifica", "studente", associati al cluster 2. Per contro, i lemmi che rimandano alla figura del docente e il suo ruolo pedagogico si posizionano ad una distanza maggiore tanto dal cluster 2 quanto dall'origine degli assi: "interrogazione", "verifica", "assegnare", "prof", "docente" costituiscono i lemmi che più contribuiscono a separare il cluster 3 dalle restanti classi semantiche. I lemmi con maggiore forza discriminante per il cluster 2 appartengono a un lessico più specialistico, proprio della pratica informatica. Si distinguono "set", "debuggare", "4o" – in riferimento a GPT-4o, una versione di modello multimodale introdotta da OpenAI a Maggio 2024 e non più disponibile – e, in particolare, "rag", acronimo di *Retrieval-Augmented Generation*, un'architettura che ottimizza le prestazioni dei modelli di intelligenza artificiale, aiutando gli LLM a fornire risposte più pertinenti e con una qualità superiore, evitando allucinazioni (Gao et al. 2024). Infine, "search", "brainstorming", "word", "tutorial" e "python" emergono come i termini maggiormente distintivi il cluster 1, rivelando l'eterogeneità delle pratiche d'uso dell'AI, specialmente di quelle digitali.

È doveroso, in ultima analisi, dedicare un breve inciso alla distribuzione delle *stop words* del corpus sul piano bidimensionale. Queste, pur essendo elementi grammaticali privi di significato semantico, contribuiscono ad arricchire l'interpretazione dei risultati. La *Figura 36* illustra nuovamente la CA condotta sui subcorpora definiti a priori. Si sceglie questa volta di considerare i *word-types* con un frequenza maggiore o uguale a 195[85] e di includere nella visualizzazione, per l'appunto, le parole grammaticali: si collocano nel primo quadrante, verbi, aggettivi e pronomi in prima e seconda persona singolare ("io", "me", "ho", "tu", "ti", "mia", "tua"), ad enfatizzare il carattere fortemente personale ed esperienziale di *sub4_rel*. Contrariamente, appartengono a *sub1_lav*, posto nel terzo quadrante, forme grammaticali opposte, impersonali e/o plurali ("hanno",

[85] Si considera la totalità dei *word-types* appartenenti alla fascia di alta frequenza e metà dei *word-types* appartenenti alla fascia di media frequenza.

"siamo", "sarà", "chi"), che richiamano la collettività e sottolineano la componente sociale del subcorpus.

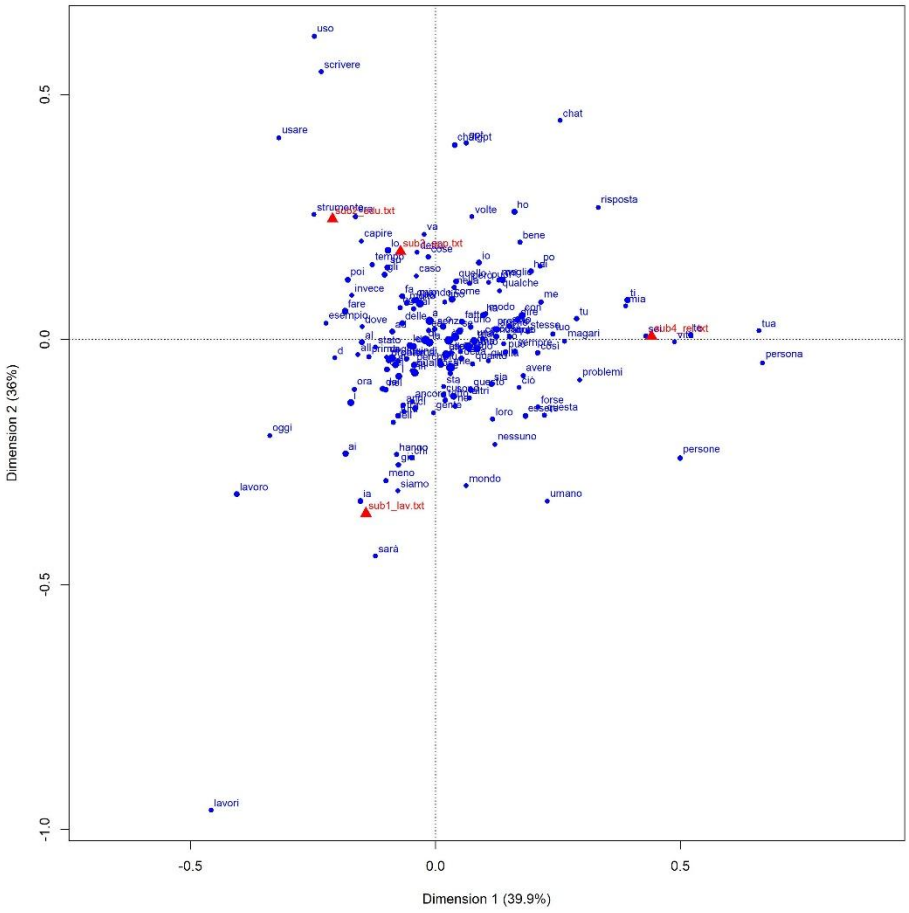

Figura 36 CA raffigurante il sistema di relazioni tra le forme caratteristiche (frequenza maggiore o uguale a 195) di ciascun subcorpus, includendo la visualizzazione delle *stop words*.

6. Conclusioni

6.1 Un bilancio generale dei risultati emersi

La ridefinizione delle interazioni sociali ha determinato, nella società contemporanea, l'emergere di una preziosa fonte empirica per l'analisi quantitativa dell'opinione pubblica. Le *threaded online conversations* dei forum di discussione si traducono oggi in corpora testuali di grandi dimensioni in grado di offrire uno spaccato autentico delle opinioni degli utenti su tematiche socialmente controverse. Tra i temi maggiormente discussi online, l'intelligenza artificiale e, in particolare l'intelligenza artificiale generativa (GenAI), emerge come argomento di dibattito pubblico caratterizzato da opinioni divergenti. Si comprende come l'integrazione di componenti quantitative e qualitative evidenzi il ruolo fondamentale che l'ASDT può svolgere nell'esplorare temi di crescente centralità: l'applicazione di tecniche statistiche, unita a una successiva attività di restituzione e interpretazione dei dati emersi, consente di identificare la molteplicità di declinazioni assunte da una tematica fortemente polarizzante, di chiarirne gli aspetti emergenti e di cogliere, tra questi, strutture di relazione latenti. Su una piattaforma come Reddit, che fa del dibattito collettivo il proprio fondamento, le comunità di utenti impegnati in scambi informativi incarnano la sfera pubblica dello spazio digitale.

L'analisi svolta nel presente elaborato ha previsto l'organizzazione dei commenti pubblicati dai redditors in quattro subcorpora tematici. Tale suddivisione è stata operata previa lettura delle submission selezionate, che ha condotto all'identificazione di quattro categorie concettuali ritenute dominanti all'interno del dibattito sull'intelligenza artificiale. Esse sono così sintetizzate: AI e lavoro (*sub1_lav*), AI e istruzione (*sub2_edu*), AI e applicazioni nel quotidiano (*sub3_app*), AI, salute mentale e relazioni (*sub4_rel*). Una prima rielaborazione dei dati testuali ha rivelato la presenza, sulla piattaforma, di tendenze diverse a seconda del contesto d'uso dell'AI: l'ambito lavorativo produce discussioni più estese, le pratiche di utilizzo quotidiano favoriscono un maggior livello di interazione e il contesto educativo determina un minor numero di interventi, seppur argomentativi e sintatticamente più complessi. Successivamente, la rappresentazione del vocabolario mediante liste di frequenze e valori di *term frequency-inverse document frequency* (TF-IDF) ha consentito di esaminare le parole di contenuto più ricorrenti all'interno del corpus e di individuare, all'interno di ciascuna raccolta di commenti, cornici discorsive altrimenti passate inosservate alla classificazione per frequenze.

Considerati i *word-types* più frequenti, un dato degno di nota è la predominanza del chatbot di OpenAI, introdotto a novembre del 2022: nel discorso pubblico attuale, parlare di AI implica inevitabilmente un riferimento immediato a ChatGPT, il primo assistente virtuale destinato al mercato di massa. L'identificazione delle forme caratterizzanti e la loro contestualizzazione tramite KWIC hanno invece fornito risultati integrativi: mentre le prime parole di contenuto dei singoli subcorpora esplicitano l'area semantica da essi rappresentata, il calcolo del TF-IDF evidenzia sequenze dialogiche latenti, dedicate alle eventuali ripercussioni dell'AI sul settore pubblico, agli interventi di politica sociale ed economica (*sub1_lav*), alla necessaria ridefinizione dei metodi didattici dei docenti (*sub2_edu*), alla versatilità d'uso dell'AI, spaziando dalla grafica digitale all'intrattenimento (*sub3_app*), e, infine, alle riflessioni critiche dei redditors sul modo in cui l'interazione tra umano e artificiale ridefinisce persino l'intimità (*sub4_rel*).

Il *text clustering* e l'analisi delle corrispondenze condotti nella seconda fase di analisi presuppongo l'esplorazione delle informazioni a disposizione senza ipotesi sulle strutture cercate. Pur presentandosi, pertanto, come uno studio esplorativo, l'applicazione dei metodi *unsupervised* ha decretato il valore confermativo dell'ASDT. Nella ricerca qui discussa, le analisi eseguite sulla classificazione tematica *ex ante* si dispiegano parallelamente alle analisi eseguite sul corpus complessivo, in un continuo confronto tra macro-tematiche definite a priori e classi semantiche emerse empiricamente.

Nel presente lavoro, il *clustering* stilometrico ha costituito il punto primario di applicazione dei metodi statistici. Raggruppando i segmenti dei subcorpora in gruppi di testi omogenei, la struttura gerarchica definita da *stylo* ha difatti introdotto tre dati che, ripresentandosi nel corso delle analisi successive, hanno preannunciato il valore confermativo della ricerca: l'algoritmo agglomerativo alla base della clusterizzazione stilometrica ha definito l'autonomia stilistica dei commenti inerenti al lavoro e al benessere psico-relazionale, la maggiore somiglianza lessicale tra i commenti sul contesto educativo e sulle pratiche d'uso nel quotidiano, e la trasversalità del lessico applicativo all'interno del corpus.

Per contro, l'algoritmo divisivo alla base del *clustering* tematico-contenutistico ha organizzato il corpus dei commenti online in cinque clusters coerenti, compatti e facilmente interpretabili, quattro dei quali hanno trovato piena corrispondenza con i subcorpora definiti a priori: osservando i lemmi più rappresentativi per ogni classe

semantica prodotta tramite metodo Reinert, si è resa evidente la concordanza tematica tra *sub1_lav* e il cluster 5, tra *sub2_edu* e il cluster 3, tra *sub3_app* e il cluster 1 e, infine, tra *sub4_rel* e il cluster 4. Si è dunque concluso che la clusterizzazione semantica corrobora in gran parte la coerenza della suddivisione manuale delle submission, confermando il lavoro, l'istruzione, le applicazioni pratiche e il benessere psico-relazionale come i principali nuclei di discussione tra gli ambiti contrassegnati dall'impatto dell'AI. Il dato più rilevante emerso dalla *Descending Hierarchical Classification* è tuttavia da ricondurre all'identificazione di un nuovo nucleo tematico, che avvalora l'efficacia del *text clustering* nel far emergere pattern latenti: il cluster 2 è la classe semantica di presentazione della GenAI, all'interno della quale un lessico incentrato sugli aspetti tecnici e funzionali dei modelli probabilistici suggerisce la tendenza a sovrastimare le capacità dei LLM.

Al fine di circoscrivere l'analisi ai suddetti nuclei di discussione e ottenerne un riquadro più dettagliato, una serie di DHC secondarie è stata eseguita sui singoli subcorpora. In *sub1_lav*, l'intelligenza artificiale è primariamente discussa in quanto fattore di ridefinizione del lavoro e, per estensione, della società nel suo complesso. I risultati più significativi si riconducono tuttavia a un ridimensionamento dell'AI, che si sostituisce al timore legato alla crescente automazione: da un lato, persiste il confronto tra le capacità cognitive attese dalla GenAI e le sue reali capacità operative, che ne incrementano l'insoddisfazione; dall'altro, si riconosce l'intelligenza artificiale come prodotto dell'intelletto umano, presupposto fondamentale per la sua stessa esistenza.

Nella DHC condotta su *sub2_edu*, l'aspetto di maggior rilievo riguarda la propensione degli studenti a prediligere un uso esecutivo degli strumenti di intelligenza artificiale (*executive help-seeking*), Questo utilizzo, prevalentemente passivo e spesso orientato alla produzione di testi, viene percepito come potenzialmente dannoso per lo sviluppo del pensiero critico e della capacità di ragionare in modo autonomo e indipendente. La clusterizzazione segnala, in tal senso, la coesistenza di diverse sezioni discorsive: alcune, proprio alla luce di questi rischi, delegittimano l'adozione dell'AI in ambito scolastico, mentre altre condannano la diffusa tendenza, nel panorama italiano, a temere il cambiamento ed eludere l'innovazione tecnologica.

Sub3_app riassume le principali pratiche di utilizzo dell'AI, regolarmente impiegata come motore di ricerca e strumento di supporto alla modifica, rielaborazione e traduzione

dei testi. Sebbene l'analisi riveli la pervasività di discussioni focalizzate sugli aspetti computazionali dei modelli probabilistici, il dato più significativo è la presenza, all'interno di un corpus dalla forte impronta pragmatica, di una componente sociale ed evolutiva. L'uso quotidiano degli strumenti di intelligenza artificiale mette difatti in discussione i confini tra ciò che è reale e ciò che artificiale e stimola, al contempo, riflessioni sui dilemmi etici connessi al suo impiego nelle attività professionali.

Data la natura e l'attualità del tema in analisi, è lecito definire *sub4_rel* come la raccolta più interessante da esaminare dal punto di vista semantico. I risultati della DHC condotta sul subcorpus rivelano che, all'interno del dibattito tra redditors, la ricerca di un supporto emotivo si traduce in un'antropomorfizzazione dei chatbot di intelligenza artificiale. Questi, percepiti in grado di simulare l'empatia umana, svolgono una funzione compensativa assumendo il ruolo di un amico con cui confidarsi, di una figura professionale di supporto o persino di un partner virtuale. Dalla clusterizzazione emergono difatti molteplici risultati degni di nota: la discussione online è attraversata dalla crescente integrazione dell'AI nella dimensione intima della vita privata, dallo stigma nei confronti di coloro ne fanno un eccessivo ricorso e da un conflitto di genere che pone le proprie radici in una diffusa solitudine maschile.

Se il *clustering* tramite DHC ha consentito di individuare i principali nuclei semantici che dominano il corpus nella sua interezza, l'analisi delle corrispondenze ha svelato le relazioni tra essi sottese. Le sezioni conclusive del presente lavoro hanno operato un confronto tra le distribuzioni, su piano cartesiano, dei subcorpora definiti a priori e delle classi semantiche emerse empiricamente. Per favorire l'esposizione dei risultati, la proiezione dei profili lessicali sul piano è stata discussa denominando le due dimensioni come segue: una dimensione funzionale-relazionale (*Dimension 1*) e una economico-sociale (*Dimension 2*) nel primo caso; una dimensione funzionale-relazionale (*facteur 1*) e una economico-relazionale (*facteur 2*) nel secondo. Dalla sola denominazione delle dimensioni, si osserva come le due analisi abbiano prodotto risultati alquanto sovrapponibili. In termini generali, essi hanno in parte evidenziato la coerenza della suddivisione tematica *ex ante* dei commenti online e delucidato le strutture gerarchiche prodotte dal clustering stilometrico e dalla DHC. In primo luogo, la rappresentazione su piano bidimensionale delle relazioni di similarità e dissimilarità tra profili lessicali ha dato prova dell'autonomia dei commenti orientati al lavoro e al benessere psico-

relazionale. Queste ultime si distinguono come le macro-tematiche che più contribuiscono alla strutturazione degli assi: nelle analisi precedenti, la loro correlazione non è tanto riconducibile a una forte somiglianza lessicale, quanto piuttosto alla loro comune dissimilarità rispetto al lessico educativo e applicativo. In secondo luogo, la CA condotta sui clusters empirici ha rivelato il lessico tecnico come fattore determinante nelle relazioni di somiglianza tra il contesto educativo e quello applicativo: le forme associate all'aspetto tecnico e di funzionamento dell'AI mostrano profili lessicali marcatamente simili a quelli delle forme associate al contesto applicativo e, in particolare, all'istruzione. Si evidenzia, infine, un ultimo risultato di considerevole importanza: laddove l'analisi venga eseguita sui clusters definiti empiricamente, il lessico inerente all'integrazione dell'AI nell'ambito lavorativo perde la propria componente tecnica ed emerge come classe semantica attinente all'impatto sociale.

6.2 Limiti dello studio e prospettive future

Nonostante la virtù del condurre un'analisi statistica dei dati testuali risieda proprio nella possibilità di dar voce ai dati, la loro acquisizione, rielaborazione e interpretazione richiede inevitabilmente un certo grado di discrezionalità da parte del ricercatore. Al di là dell'attività interpretativa, la selezione dei dati testuali e dei metodi statistici da applicare conserva un livello di arbitrarietà. Le scelte che il ricercatore opera, seppur mirate a una piena coerenza con gli obiettivi prefissati e con la natura dei testi oggetti di studio, vanno riconosciute come tali: una delle soluzioni adottate tra diverse alternative disponibili. È pertanto opportuno riconoscere i limiti metodologici che tali decisioni comportano e proporre possibili traiettorie di ricerca che possano, in futuro, produrre risultati integrabili o confrontabili con quelli discussi nel presente elaborato.

In primo luogo, la scelta di condurre l'analisi su un corpus di commenti in lingua italiana ha necessariamente posto dei limiti alla dimensione e alla rappresentatività del corpus rispetto a un fenomeno globale come il dibattito sull'intelligenza artificiale. Come avviene nella maggior parte delle piattaforme digitali, anche su Reddit l'ampiezza e la varietà delle discussioni in lingua inglese risultano nettamente predominanti rispetto a quelle in lingua italiana. Un'estensione dell'analisi a un corpus inglese potrebbe pertanto, da un lato, arricchire i risultati ottenuti e, dall'altro, offrire elementi di confronto tra due contesti distanti tanto da un punto di vista linguistico quanto culturale.

In secondo luogo, l'applicazione del *text clustering* e dell'analisi delle corrispondenze consente unicamente di formulare ipotesi sulla dimensione emotiva dei commenti online. Sebbene i due metodi si rivelino efficaci nel conferire maggiore chiarezza al complesso sistema di opinioni che scaturiscono da un tema socialmente controverso, lasciano irrisolto il quesito relativo al giudizio collettivo che lo attraversa, e dunque al *sentiment* con cui gli utenti si esprimono in merito all'intelligenza artificiale. Un possibile sviluppo futuro potrebbe dunque consistere nell'integrazione di un terzo metodo appartenente al ramo dei metodi *unsupervised*: la *Sentiment Analysis*, che tramite adozione del tradizionale *lexicon-based approach* (Liu 2015) aspira a comprendere e interpretare l'orientamento valutativo dei testi analizzati (Ceron et al. 2014) utilizzando, contrariamente ai due metodi statistici applicati, principalmente informazioni esterne al corpus. Applicare la *Sentiment Analysis* a un corpus di commenti Reddit risulterebbe coerente con il contesto in cui essa acquisisce popolarità: un ambiente digitale in cui l'analisi degli *opinion texts* – quali recensioni online, tweets, forum di discussione, blog, post sui social media o altre tipologie di UGC – è di costante interesse per aziende, attori politici ed enti pubblici. Conferendo un peso a sentimenti e opinioni, la *Sentiment Analysis* consentirebbe di rilevare la polarità complessiva (positiva o negativa) dei commenti in esame e di ricostruire gli "umori" dei redditors partecipanti al dibattito online, particolarmente rilevante nel caso di un'innovazione tecnologica il cui impatto si estende a molteplici ambiti, dall'economia alla politica, fino alle relazioni interpersonali.

La coniugazione dei risultati prodotti dal *text clustering* e dall'analisi delle corrispondenze (oltre che, come suggerito, da un'eventuale *Sentiment Analysis*) potrebbe infine costituire le fondamenta di una seconda fase sperimentale. L'esplorazione delle opinioni dei redditors tramite analisi statistica dei dati testuali può infatti essere interpretata come una fase preliminare di esplorazione del fenomeno: gli utenti della piattaforma possono essere intesi come soggetti intervistati – seppur non appartenenti a un campione rappresentativo – al fine di conoscere il fenomeno analizzato e, in altri termini, rilevare le informazioni che dominano il dibattito online sull'intelligenza artificiale. I risultati emersi potrebbero dunque costituire un fondamentale punto di partenza per la costruzione di un questionario altamente strutturato, mirato a verificare e integrare i dati raccolti nella prima fase. In particolare, dopo aver selezionato un numero di commenti sulla base della loro rilevanza analitica, questi potrebbero fungere da testi-

stimolo da includere nel questionario. In tal senso, lo strumento quantitativo potrebbe perseguire un ulteriore obiettivo: verificare se l'esposizione dei rispondenti alle opinioni altrui influisca sulla loro posizione originaria in merito a un tema fortemente divisivo.

La stesura di questo elaborato ha previsto l'uso di chatbot basati sull'intelligenza artificiale per supportare fasi specifiche del processo di ricerca. Tali strumenti sono stati impiegati per attività non decisionali: la generazione di codice Python per l'estrazione dei commenti pubblicati su Reddit, corpus testuale oggetto della presente ricerca; la generazione di codice R per l'esecuzione della *Correspondence Analysis*. Tutti i contenuti finali, le analisi e le interpretazioni critiche riflettono il mio lavoro e il mio contributo intellettuale. L'uso dell'intelligenza artificiale è stato esclusivamente di supporto e tutti gli output sono stati attentamente letti, revisionati e modificati.

Appendice A: il codice Python per l'estrazione dei commenti Reddit

```
import praw
import pandas as pd

# autenticazione Reddit
reddit = praw.Reddit(
    client_id="[…]",
    client_secret="[…]",
    user_agent="Tesi-CorpusReddit-AI"
)

# inserisci l'URL del post Reddit
post_url = "https://www.reddit.com/[...]/"

# ottieni l'oggetto Submission
submission = reddit.submission(url=post_url)
submission.comments.replace_more(limit=None)

# mappa ID commenti per ricostruire la gerarchia
comment_lookup = {}

# funzione ricorsiva per estrarre i commenti in ordine
def extract_comments(comments, parent_name=None):
    extracted = []
    for comment in comments:
        if isinstance(comment, praw.models.Comment):
            # escludi solo commenti effettivamente rimossi o vuoti
            if comment.body.strip().lower() in ["[removed]", "[deleted]"]:
                continue

            # etichetta se l'autore è deleted
            is_deleted_author = comment.author is None
            author_str = str(comment.author) if comment.author else "[deleted]"

            # trova l'autore a cui si sta rispondendo
            parent_name = comment_lookup.get(comment.parent_id, parent_name)

            # calcola il numero di risposte dirette al commento (replyCount)
            reply_count = len(comment.replies)

            data = {
                "author": author_str,
                "text": comment.body,
                "score": comment.score,
                "replyCount": len(comment.replies),
                "created_utc": comment.created_utc,
                "parent_id": comment.parent_id,
                "comment_id": comment.id,
                "post_id": submission.id,
                "post_title": submission.title,
                "subreddit": submission.subreddit.display_name,
                "isReplyToName": str(parent_name) if parent_name else None
            }

            extracted.append(data)

            # registra autore per commenti figli
            comment_lookup[f"t1_{comment.id}"] = author_str

            # estrai commenti figli mantenendo l'ordine
            extracted += extract_comments(comment.replies, parent_name=author_str)
    return extracted

# estrai i commenti ordinati
```

```
ordered_comments = extract_comments(submission.comments)

# salva in CSV
df = pd.DataFrame(ordered_comments)
df.to_csv("reddit_commenti_filtrati.csv", index=False)

print(f"Salvati {len(df)} commenti (escludendo quelli rimossi ma includendo quelli di
utenti [deleted]).")
```

Appendice B: struttura del corpus e collegamenti ipertestuali alle submission

	Post id	*Commenti*	*Commenti validi*	*Commenti-genitore*	*Trimestre*
sub1_lav	post1_lav	299	273	99	Q2 2025
	post2_lav	262	233	100	Q1 2025
	post3_lav	248	244	55	Q3 2022
	post4_lav	215	213	47	Q2 2024
	post5_lav	166	135	47	Q1 2024
	post6_lav	143	133	49	Q3 2024
	post7_lav	126	123	37	Q4 2023
	tot.	**1459**	**1354**	**434**	
sub2_edu	post1_edu	340	329	121	Q2 2025
	post2_edu	276	260	91	Q1 2025
	post3_edu	222	219	62	Q3 2022
	post4_edu	190	179	58	Q2 2024
	post5_edu	157	132	28	Q1 2024
	post6_edu	95	88	28	Q3 2024
	post7_edu	66	61	29	Q4 2023
	tot.	**1346**	**1268**	**417**	
sub3_app	post1_app	385	344	72	Q2 2025
	post2_app	262	249	121	Q1 2025
	post3_app	215	213	106	Q3 2022
	post4_app	208	191	96	Q2 2024
	post5_app	192	179	102	Q1 2024
	post6_app	135	128	49	Q3 2024
	post7_app	126	123	43	Q4 2023
	tot.	**1523**	**1427**	**589**	
sub4_rel	post1_rel	415	408	79	Q2 2025
	post2_rel	350	231	79	Q1 2025
	post3_rel	184	164	64	Q3 2022
	post4_rel	169	162	72	Q2 2024
	post5_rel	99	93	40	Q1 2024
	post6_rel	90	79	32	Q3 2024
	post7_rel	61	59	19	Q4 2023
	tot.	**1368**	**1196**	**385**	
corpus		**5696**	**5245**	**1825**	

Riferimenti bibliografici

Aguilar, S. J., Nye, B., Swartout, W. R., Macias, A., Xing, Y., & Xiu, R. L. (2025). *Fostering Critical Thinking in the Age of AI: How Students and Teachers Worldwide are adapting to AI.* Los Angeles, CA: USC Center for Generative AI and Society.

Aichner, T., & Jacob, F. H. (2015). Measuring the Degree of Corporate Social Media Use. *Internation Journal of Market Research, 57*(2), 257-276.

Aichner, T., Grünfelder, M., Maurer, O., & Jegeni, D. (2021). Twenty-Five Years of Social Media: A Review of Social Media Applications and Definitions from 1994 to 2019. *Cyberpsychology, Behavior, and Social Networking, 24*(4), 215-222.

Ali, S. M., & Hussein, K. S. (2014). The Comparative Power of Type/Token and Hapaxlegomena/Type Ratios: A Corpus-based Study of Authorial Differentiation. *Advances in Language and Literary Studies, 5*(3), 112-119.

Anelli, M., Napoli, L., & Cabrerizo, R. (2025). *Consumer Digital Empowerment Index*. Tratto da Consumer Empowerment Project (CEP): https://index.cep-project.org/

Anju, A. (2024). Exploring the Impact of Social Media on Public Opinion Formation: A Comparative Analysis. *International Journal for Research in Applied Science and Engineering Technology (IJRASET), 12*(4), 307-311.

Anthony, L. (2020). AntConc (versione 3.5.9) [Computer Software]. Tokyo, Japan: Waseda University. Tratto da https://laurenceanthony.net/software.html

Appel, R., Massenkof, M., McCrory, P., McCain, M., Heller, R., Neylon, T., & Tamkin, A. (2026). *The Anthropic Economic Index report: Economic Primitives.* Anthropic. Tratto da https://www.anthropic.com/research/anthropic-economic-index-january-2026-report

Aragón, P., Gómez, V., García, D., & Kaltenbrunner, A. (2017). Generative Models of Online Discussion Threads: State of the Art and Research Challenges. *Journal of Internet Services and Applications, 8*(15), 1-17.

Arditi, C., Walther, D., Gilles, I., Lesage, S., Griesser, A.-C., Bienvenu, C., . . . Peytremann-Bridevaux, I. (2020). Computer-assisted textual analysis of freetext comments in the Swiss Cancer Patient Experiences (SCAPE) survey. *BMC Health Services Research, 20*(1), 1-12.

Barbera, M., Corino, E., & Onesti, C. (2007). Cosa è un corpus? Per una definizione più rigorosa di corpus, token, markup. In M. Barbera, E. Corino, & C. Onesti (A cura di), *Corpora e linguistica in rete* (p. 25-88). Perugia: Guerra Edizioni.

Benzécri, J.-P. (1973). *L'analyse des données.* Paris: Dunod.

Bernardi, L., & Campostrini, S. (2005). Analisi, interpretazione e diffusione dei dati. In L. Bernardi (A cura di), *Percorsi di ricerca sociale. Conoscere, decidere, valutare* (p. 255-264). Roma: Carocci.

Berruto, G. (1987). *Sociolinguistica dell'italiano contemporaneo.* Roma: La Nuova Italia Scientifica (Carocci).

Bolter, J. D., & Grusin, R. (2000). *Remediation: Understanding New Media.* Cambridge: MIT Press.

Boulianne, S. (2015). Social media use and participation: A meta-analysis of current research. *Information, Communication & Society, 18*(5), 524-538.

Brezina, V. (2018). *Statistics in Corpus Linguistics: A Practical Guide.* Cambridge: Cambridge University Press.

Calvino, F., Haerle, D., & Liu, S. (2025). Is Generative AI a General-Purpose Technology? *OECD Artificial Intelligence Papers*(40).

Camargo, B. V., & Justo, A. (2013). IRAMUTEQ: A Free Software for Analysis of Textual Data. *Temas em Psicologia, 21*(2), 513-518.

Camargo, B. V., & Justo, A. (2021). *IRaMuTeQ Tutorial.* Tratto da IRaMuTeQ: Traduction du tutoriel portugais en anglais par Teresa Forte: https://pratinaud.gitpages.huma-num.fr/iramuteq-website/category/documentation.html

Castells, M. (1996). *The Rise of the Network Society.* Oxford: Blackwell Publishers; tr. it. (2002) La nascita della società in rete, Bocconi editore, Milano.

Ceron, A., Curini, L., & Iacus, S. M. (2014). *Social media e sentiment analysis. L'evoluzione dei fenomeni sociali attraverso la rete.* Milano: SPringer.

Chatterji, A., Cunningham, T., Deming, D. J., Hitzig, Z., Ong, C., Carl, S. Y., & Wadman, K. (2025). *How People Use ChatGPT.* Cambridge: NBER (National Bureau of Economic Research) Working Paper Series.

Childs, H. L. (1940). *An Introduction to Public Opinion.* New York: John Wiley and Sons.

Ciotti, F. (2021). Distant reading in literary studies: a methodology in quest of theory. *Testo e Senso*, 195-213.

Cohen, B. C. (1963). *The Press and Foreign Policy.* Princeton: Princeton University Press.

Cohen, B. K., Hunter, L. E., & Pressman, P. S. (2019). P-Hacking Lexical Richness Through Definitions of “Type” and “Token”. In L. Ohno-Machado, & B. Séroussi (A cura di), *MEDINFO 2019: Health and Wellbeing e-Networks for All* (p. 1433-1434). Lyon: IOS Press.

Corbetta, P. (2015). *La ricerca sociale: metodologie e tecniche* (Vol. IV. L'analisi dei dati). Bologna: Il Mulino.

Corchia, L. (2011). *La teoria dell'agenda-setting.* Dispensa universitaria, Università di Pisa, Pisa. Tratto da https://www.academia.edu/3793704/Luca_Corchia_La_teoria_dellagenda_setting_maggio_2011

Daft, R. L., & Lengel, R. H. (1986). Organizational information requirements, media richness, and structural design. *Management Science, 32*(5), 554-571.

DeCook, J. (2021). Castration, the Archive, and the Incel Wiki. *Psychoanalysis, Culture & Society, 26*, 234-243.

Delfanti, A., & Arvidsson, A. (2019). *Introduction to Digital Media.* Hoboken: John Wiley & Sons.

Dewey, J. (1927). *The Public and Its Problems.* New York: Henry Holt and Company.

Dobber, T., & Hameleers, M. (2024). The Social Media Comment Section as an Unruly Public Arena: How Comment Reading Erodes Trust in News Media. *Electronic News: Broadcast and Mobile Journalism, 19*(1), 3-18.

Douai, A., & Nofal, H. K. (2012). Commenting in the online Arab public sphere: Debating the Swiss minaret ban and the “Ground Zero Mosque” online. *Journal of Computer-Mediated Communication, 17*(3), 266-282.

Drugaș, M. (2022). Screenagers or "Screamagers"? Current Perspectives on Generation Alpha. *Psychological Thought, 15*(1), 1-11.

Eder, M. (2015). Does size matter? Authorship attribution, small samples, big problem. *Digital Scholarship in the Humanities, 30*(2), 167-182.

Eder, M., Rybicki, J., & Kestemont, M. (2016). Stylometry with R: A Package for Computational Text Analysis. *R Journal, 8*(1), 107-121.

Entman, R. M. (1993). Framing: Toward Clarification of A Fractured Paradigm. *Journal of Communication, 43*(4), 51-58.

Evert, S., Proisl, T., Jannidis, F., Reger, I., Pielström, S., Schöch, C., & Vitt, T. (2017). Understanding and explaining Delta measures for authoriship attribution. *Digital Scholarship in the Humanities, 32*(2), ii4-ii16.

Ferraresi, M., & Schmitt, B. H. (2018). *Marketing esperienziale: Come sviluppare l'esperienza di consumo.* Milano: FrancoAngeli.

Fobbe, S. (2026). *Key Word in Context (KWIC) Analysis and Lexical Dispersion Plots*. Tratto da Seán Fobbe: https://seanfobbe.com/tutorials/kwic-lexical-dispersion/

Gao, Y., Xiong, Y., Gao, X., Kangxiang, J., Jinliu, P., Yuxi, B., . . . Haofen, W. (2024). Retrieval-Augmented Generation for Large Language Models: A Survey. 1-21.

Gearhart, S., Coman, I. A., Moe, A., & Brammer, S. (2023). Facebook Comments Influence Perceptions of Journalistic Bias: Testing Hostile Media Bias in the COVID-19 Social Media Environment. *Electronic News, 17*(1), 3-18.

Gearhart, S., Moe, A., & Zhang, B. (2020). Hostile Media Bias on Social Media: Testing the Effect of User Comments on Perceptions of News Bias and Credibility. *Human Behavior and Emerging Technologies, 2*(2), 140-148.

Gherardi, L. (2022). Il fantasma dell'opinione pubblica. Cenni a una lunga storia di studi. In L. Gherardi (A cura di), *Lezioni brevi sull'opinione pubblica. Nuove tendenze nelle scienze sociali* (p. 19-32). Milano: Meltemi.

Giuliano, L., & La Rocca, G. (2008). *L'analisi automatica e semi-automatica dei dati testuali: Software e istruzioni per l'uso.* Milano: Edizioni Universitarie di Lettere Economia Diritto.

Goffman, E. (1959). *The Presentation of Self in Everyday Life.* New York: Doubleday Anchor Books.

Goffman, E. (1974). *Frame Analysis: An Essay on the Organization of Experience.* New York: Harper & Row.

Gundecha, P., & Liu, H. (2012). Mining Social Media: A Brief Introduction. *INFORMS TutORials in Operations Research*, 1-17.

Habermas, J. (1962). *Strukturwandel der Öffentlichkeit.* Neuwied: Hermann Luchterhand Verlag; tr. it. (1974) Storia e critica dell'opinione pubblica, Laterza, Roma-Bari.

Holmes, D. I. (1998). The Evolution of Stylometry in Humanities Scholarship. *Literary and Linguistic Computing, 13*(3), 111-117.

Horne, B. D., Adali, S., & Sikdar, S. (2017). Identifying the Social Signals that Drive Online Discussions: A Case Study of Reddit Communities. *International Conference of Computer Communications and Networks*, 1-10.

Imm, S., & Kang, T. U. (2020). Intimacy and Love with Artificial Intelligence in the Movie "Her". *Psychoanalysis, 31*(4), 91-97.

Jänicke, S., Franzini, G., Cheema, M. F., & Scheuermann, G. (2015). On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. *Eurographics Conference on Visualization (EuroVis).* Cagliari: The Eurographics Association.

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons, 53*(1), 59-68.

Kergroach, S., & Héritier, J. (2025). Emerging Divides in the Transition to Artificial Intelligence. *OECD Regional Development Papers*(147). doi:https://doi.org/10.1787/7376c776-en

Kietzmann, J. H., Hermkens, K., McCarthy, I. P., & Silvestre, B. S. (2011). Social media? Get serious! Understanding the functional building blocks of social media. *Business Horizons, 54*(3), 241-251.

Koukaras, P., Tjortjis, C., & Rousidis, D. (2020). Social Media Types: introducing a data driven taxonomy. *Computing, 102*(1), 295-340.

Lahjouji-Seppälä, M. Z., Rabus, A., & von Waldenfels, R. (2022). Ukrainian standard variants in the 20th century: stylometry to the rescue. *Russian Linguistics, 46*, 217-232.

Lasswell, H. D. (1927). *Propaganda Technique In The World War.* London: Kegan Paul, Trench, Trübner & Co., Ltd.

Lebart, L., & Salem, A. (1994). *Statistique textuelle.* Paris: Dunod.

Lebart, L., Salem, A., & Berry, L. (1998). *Exploring Textual Data.* Dordrecht: Kluwer Academic Publishers.

LELO. (2025). *Intergenerational Views on Relationships, Sex and Technology.* Tratto da LELO's Future of Sex and Love Report 2025: https://www.lelo.com/blog/wp-content/uploads/2025/06/LELO-Futurist-Report-2025.pdf

Lippmann, W. (1922). *The public opinion.* New York: Harcourt Brace; tr. it. (2004) L'opinione pubblica, Donzelli, Roma.

Liu, B. (2012). Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies, 5*(1), 1-167.

Liu, B. (2015). *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions.* New York: Cambridge University Press.

Loewen, P. J., Lee-Whiting, B., Arai, M., Bergeron, T., Galipeau, T., Gazendam, I., . . . Yusypovych, S. (2024). *Global Public Opinion on Artificial Intelligence (GPO-AI).* Toronto: Schwartz Reisman Institute for Technology and Society (SRI).

Mayer, H., Yee, L., Chui, M., & Roberts, R. (2025). *Superagency in the Workplace: Empowering People to Unlock AI's Full Potential.* New York: McKinsey & Company.

McCombs, M. E., & Shaw, D. L. (1972). The Agenda-Setting Function of Mass Media. *The Public Opinion Quarterly, 36*(2), 176-187.

McQuail, D. (2010). *McQuail's Mass Communication Theory* (6th edition ed.). London: Sage Publications.

Medvedev, A., Lambiotte, R., & Delvenne, J.-C. (2019). The Anatomy of Reddit: An Overview of Academic Research. In F. Ghanbarnejad, R. S. Roy, F. Karimi, J.-C. Delvenne, & B. Mitra, *Dynamics on and of Complex Networks III* (p. 183-2014). Cham, Switzerland: Springer Nature.

Menduni, E. (2007). *I media digitali: tecnologie, linguaggi, usi sociali.* Laterza.

Metzger, M. J., Flanagin, A. J., & Medders, R. B. (2010). Social and Heuristic Approaches toCcredibility Evaluation Online. *Journal of Communication, 60*(3), 413-439.

Microsoft. (2025, Agosto 6). Visual Studio Code (VS Code) [versione 1.103.1]. Tratto da https://code.visualstudio.com/

Montalescot, L., Lamore, K., Flahault, C., & Untas, A. (2024). What is the place of interpretation in text analysis? An example using ALCESTE® software. *Qualitative Research in Psychology*, 1-26.

Moreno, M., & Retinaud, P. (2022, Dicembre). *Manual para el usuario.* Tratto da IRaMuTeQ: Guia IRaMuTeQ: https://pratinaud.gitpages.huma-num.fr/iramuteq-website/category/documentation.html

Moretti, F. (2005). *Graphs, Maps, Trees: Abstract Models fora Literary History.* London: Verso.

Moro, E. (2026, Gennaio 19). *L'AI sta già influenzando il modo in cui facciamo sesso (ma forse non ce ne siamo ancora resi conto)*. Tratto da Cosmopolitan: https://www.cosmopolitan.com/it/sesso-amore/a70010634/intelligenza-artificiale-influenza-il-sesso/

Mosteller, F., & Wallace, D. L. (2007 [1964]). *Inference and disputed authorship: The Federalist.* Stanford: Center for the Study of Language and Information.

Muthukrishn, M., Peters, A. C., Di Canossa, V., & Zhu, L. J. (2025). *The Next Great Divergence: Why AI May Widen Inequality Between Countries.* New York: United Nations Development Programme (UNDP).

Naab, T. K., & Küchler, C. (2023). Content Analysis in the Research Field of Online User Comments. In F. Oehmer-Pedrazzi, S. H. Kessler, E. Humprecht, K. Sommer, & L. Castro, *Standardisierte Inhaltsanalyse in der Kommunikationswissenschaft - Standardized Content Analysis in Communitcation Research* (p. 441-450). Wiesbaden: Springer VS.

Naab, T., & Sehl, A. (2017). Studies of User-Generated Content: A Systematic Review. *Journalism, 18*(10), 1256-1273.

Newman, N., Fletcher, R., Kalogeropoulos, A., & Nielsen, R. K. (2019). *Reuters Institute Digital News Report 2019.* Oxford: Reuters Institute. Tratto da https://reutersinstitute.politics.ox.ac.uk/sites/default/files/inline-files/DNR_2019_FINAL.pdf

Nicodemo, F. (2017). *Disinformazia: La comunicazione al tempo dei social media.* Marsilio Editori.

Nobile, S. (2024). L'analisi dei dati testuali. In A. Fasanella, S. Mauceri, & S. Nobile, *Metodologia della ricerca sociale. Approcci, strategie e tecniche di indagine* (p. 577-593). Milano: FrancoAngeli.

OECD. (2007). *Participative Web and User-Created Content: Web 2.0, Wikis, and Social Networking.* Paris: Organisation for Economic Co-operation and Development.

Ondelli, S. (2018). Treat Texts as Data but Remember They Are Made of Words: Compiling and Pre-processing. In A. Tuzzi (A cura di), *Tracing the Life Cycle of Ideas in the Humanities and Social Sciences* (p. 133-150). Cham, Switzerland: Springer Nature.

Pew Research Center. (2025, Novembre 20). *Social Media Fact Sheet*. Tratto da Pew Research Center: https://www.pewresearch.org/internet/fact-sheet/social-media/

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. *motion: Theory, research, and experience, 1*(3), 3-33.

Posit team. (2023). RStudio: Integrated Development Environment for R. Boston, MA. Tratto da http://www.posit.co/

PR Newswire. (2026, Gennaio 22). *What's hot in 2026: LELO reveals trends that everybody needs to know*. Tratto da PR Newswire: https://www.prnewswire.com/news-releases/whats-hot-in-2026-lelo-reveals-trends-that-everybody-needs-to-know-302667891.html

Prensky, M. (2001). Digital Natives, Digital Immigrants. *On the Horizon, 9*(5), 1-6.

Quarta, A., & Smorto, G. (2020). *Diritto privato dei mercati digitali.* Firenze: Le Monnier Università.

Ratinaud, P. (2025 [2009]). IRaMuTeQ: Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires. Tratto da https://pratinaud.gitpages.huma-num.fr/iramuteq-website/

Reddit. (2023a). *r/reddit.com*. Tratto da Reddit: https://www.reddit.com/r/reddit.com/wiki/api/

Reddit. (2023b, Aprile 18). *Data API Terms*. Tratto da Reddit: https://redditinc.com/policies/data-api-terms

Reddit. (2025a, Giugno 30). *Reddit by the numbers*. Tratto da Reddit: https://redditinc.com/

Reddit. (2025b, Novembre 12). *Reddit Data API Wiki*. Tratto da Reddit: https://support.reddithelp.com/hc/en-us/articles/16160319875092-Reddit-Data-API-Wiki

Reddit. (2025c, Maggio 29). *Reddit User Agreement*. Tratto da Reddit: https://redditinc.com/policies/user-agreement

Reinert, M. (1990). Alceste une méthodologie d'analyse des données textuelles et une application: Aurelia De Gerard De Nerval. *Bulletin de Méthodologie Sociologique, 26*(1), 24-54.

Retinaud, P. (2015, febbraio 28). *Thread: [Iramuteq-users] Taille segments de texte [Messaggio nella mailing list iramuteq-users]*. Tratto da SourceForge: https://sourceforge.net/p/iramuteq/mailman/iramuteq-users/thread/54F1EB14.9030102%40univ-tlse2.fr/#msg33505982

Rieder, B. (2015, Maggio 04). *YouTube Data Tools (Version 1.42) [Software]*. Tratto da YouTube Data Tools: https://ytdt.digitalmethods.net/

Rodríguez, R., Gorostiza-Cerviño, A., Hidalgo-Tenorio, E., Moyano, M., & Maldonado, M. A. (2025). Exploring incel discourse through topic modeling: insights from Spanish-speaking contexts on X. *Humanities and Social Sciences Communications*.

Saif, M. M., & Turney, P. D. (2010). Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon. *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text* (p. 26-34). Los Angeles: Association for Computational Linguistics.

Schwab, K. (2016). *La quarta rivoluzione industriale.* Milano: FrancoAngeli.

Shaw, E. F. (1979). Agenda-Setting and Mass Communication Theory. *International Communication Gazette*, 96-105.

Short, J., Williams, E., & Christie, B. (1976). *The Social Psychology of Telecommunications.* Hoboken: John Wiley & Sons, Ltd.

Siegel, L. (2011). *Homo interneticus. Restare umani nell'era dell'ossessione digitale.* Prato: Piano B.

Sinclair, S., & Rockwell, G. (2026 [2016]). Voyant Tools (versione 2.0) [Web-based Software]. Tratto da http://voyant-tools.org/.

Stroud, N. J., Van Duyn, E., & Peacock, C. (2016). *Survey of Commenters and Comment Readers.* Austin: Center for Media Engagement. Tratto da https://mediaengagement.org/research/survey-of-commenters-and-comment-readers/

Technical Committee ISO/TC 215, Health informatics. (2025). *Reference architecture for syndromic surveillance systems for infectious diseases*. Tratto da ISO Online Browsing Platform (OBP): https://www.iso.org/obp/ui#iso:std:iso:ts:6226:ed-1:v1:en:term:3.1

Technical Committee ISO/TC 292, Security and resilience. (2024). *Authenticity, integrity and trust for products and documents: Guidelines for brand protection and enforcement procedures*. Tratto da ISO Online Browsing Platform (OBP): https://www.iso.org/obp/ui#iso:std:iso:ts:22386:ed-1:v1:en:term:3.4.1

Toffler, A. (1980). *The Third Wave.* New York: Bantam Books.

Trajtenberg, M., & Bresnahan, T. F. (1995). General purpose technologies: 'Engines of growth?'. *Journal of Econometrics*, 83-108.

Transnational Institute. (2024). *Digital Capitalism week 2: Big Tech and the Digital Overloads.* Tratto da https://www.tni.org/en/publication/digital-capitalism

Tuzzi, A. (2009). L'analisi statistica dei dati testuali come strumento di ricerca in contesti di sofferenza. In I. Testoni, D. Di Lucia Sposito, & F. Martini (A cura di), *Il Morire tra Ragione e Fede. Universi che orientano le pratiche di aiuto* (p. 1-15). Padova: Padova University Press.

Tuzzi, A. (2012). Reinhard Köhler's scientific production: words, numbers and pictures. In S. Naumann, P. Grzybek, R. Vulanović, & G. Altmann (A cura di), *Synergetic Linguistics. Text and Language as Dynamic Systems* (p. 223-242). Vienna, Austria: Praesens Verlag.

Tuzzi, A. (2024). *Fondamenti di analisi dei dati testuali.* Roma: Carocci.

von Sikorski, C. (2016). The Effects of Reader Comments on the Perception of Personalized Scandals: Exploring the Roles of Oomment Valence and Commenters' Social Status. *Journal of Communication*, 4480-4501.

We Are Social & Meltwater. (2025a, Febbraio 5). *Digital 2025 Global Overview Report.* Tratto da We Are Social: https://wearesocial.com/uk/blog/2025/02/digital-2025-the-essential-guide-to-the-global-state-of-digital/

We Are Social & Meltwater. (2025b, Febbraio 25). *Digital 2025: Italy.* Tratto da DataReportal: https://datareportal.com/reports/digital-2025-italy

We Are Social & Meltwater. (2025c, Dicembre). *Digital 2026: Italy*. Tratto da We Are Social: https://wearesocial.com/it/blog/2025/10/digital-2026/

Webster, F. (2014). *Theories of the Information Society.* New York: Routledge.

World Economic Forum. (2025). *Future of Jobs Report 2025.* Geneva: World Economic Forum.

Zampieri, S., Botturi, L., & Calvo, S. (2018). Giovani e tecnologie: tra nativi digitali e competenze effettive. *Schweizerische Zeitschrift für Bildungswissenschaften, 40*(2), 307-333.

Zao-Sanders, M. (2025). *How People are Really Using Generative AI Now.* Brighton, Massachusetts: Harvard Business Review.

Ziatdinov, R., & Cilliers, J. (2021). Generation Alpha: Understanding the Next Cohort of University Students. *European Journal of Contemporary Education, 10*(3), 783-789.

Ziegele, M., & Quiring, O. (2013). Conceptualizing Online Discussion Value: A Multidimensional Framework for Analyzing User Comments on Mass-Media Websites. *Communication Yearbook, 37*(1), 127-154.