

UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Scienze Statistiche

Corso di Laurea Magistrale in Scienze Statistiche

Topic Modeling Parametrico e Non-Parametrico per Dati Testuali
Complessi: Un'Applicazione ai Centri di Ascolto Caritas

Relatore:

Ch.mo Prof. Francesco Denti

Laureando:

Matteo Brognera

Matricola: 2157164

Anno Accademico 2025–2026

Ringraziamenti

Desidero ringraziare l'Università degli Studi di Padova e il Dipartimento di Scienze Statistiche per il percorso di formazione e per le opportunità di crescita accademica e personale che mi hanno offerto durante questi anni.

Un ringraziamento particolare va al mio relatore, il professor Francesco Denti, per avermi seguito nello sviluppo di questo lavoro, per la disponibilità, i preziosi suggerimenti e il supporto fornito durante tutte le fasi della ricerca.

Desidero esprimere la mia profonda gratitudine alla Caritas Diocesana di Treviso per aver reso possibile questo studio, mettendo a disposizione i dati necessari alla sua realizzazione e permettendomi di approfondire una realtà di grande valore sociale. Ringrazio in modo particolare don Bruno per la disponibilità, la fiducia e l'attenzione con cui ha sostenuto il progetto.

Un ringraziamento speciale va infine a Cecilia, operatrice della Caritas di Treviso, per il tempo e la competenza che mi ha dedicato. Le sue spiegazioni sono state fondamentali per comprendere la struttura dei dati, il funzionamento dei servizi e il contesto nel quale le informazioni vengono raccolte. La ringrazio inoltre per avermi accompagnato durante il lavoro e per avermi fornito i dati su cui si fonda questa tesi.

Abstract

Il presente lavoro di tesi sviluppa e applica una procedura metodologica basata su tecniche di *topic modeling* per valorizzare il patrimonio informativo contenuto nelle note narrative di colloquio dei Centri di Ascolto della Caritas Diocesana di Treviso. L'obiettivo è quello di trasformare i testi liberi redatti dagli operatori in conoscenza strutturata a supporto delle decisioni, nel pieno rispetto della riservatezza dei dati sensibili. Sul piano metodologico, la tesi confronta il modello parametrico *Latent Dirichlet Allocation* (LDA) e la sua estensione non parametrica *Hierarchical Dirichlet Process* (HDP). Nell'HDP, integrando le distribuzioni tematiche con le variabili socio-demografiche degli assistiti, emerge una forte polarizzazione della clientela: da una parte il topic dell'emergenza abitativa e della necessità di un rapido supporto burocratico per giovani uomini stranieri; dall'altra il sostegno di famiglie con radicata presenza sul territorio con fragilità di salute, economiche e relazionali. Lo studio mostra come l'analisi del testo permetta di andare oltre la rigidità dei tradizionali indicatori, fornendo alla direzione di un ente uno strumento basato su dati di fatto, scalabile e trasferibile, con cui pianificare i servizi.

Indice

1	Introduzione, ambito di applicazione e metodologie d'indagine	1
2	I Dati Caritas: Pre-processing e Profilazione	3
2.1	Il contesto operativo e l'origine dei dati	3
2.1.1	Il contesto istituzionale e i dati di registro	3
2.1.2	Analisi temporale della qualità informativa e criteri di selezione	4
2.2	Trattamento e pre-processing delle note	6
2.2.1	Costruzione del vocabolario iniziale	6
2.2.2	Approcci metodologici di riduzione del lessico	7
2.2.3	Valutazione comparativa tra le diverse tecniche utilizzate . .	12
2.3	Analisi descrittiva del campione e trattamento delle assenze informative	14
2.3.1	Distribuzione per genere	14
2.3.2	Distribuzione per stato civile	15
2.3.3	Distribuzione per età	15
2.3.4	Distribuzione per area geografica	16
2.3.5	Mappatura dei bisogni degli assistiti	17
2.3.6	Mappatura degli interventi negli assistiti	18
3	Modellazione Tematica tramite Latent Dirichlet Allocation	21
3.1	Fondamenti teorici della LDA	21
3.1.1	Inquadramento nel panorama del topic modeling	21
3.2	Struttura Probabilistica e Processo Generativo	22
3.2.1	Inferenza statistica tramite Collapsed Gibbs Sampling	24
3.3	Implementazione e Configurazione della LDA	26
3.3.1	Dalle note processate alla matrice documenti-termini	26
3.3.2	Algoritmo di Stima e Gibbs Sampling	27
3.3.3	Ottimizzazione del Numero di topic (K)	27
3.3.4	Taratura e Sensibilità degli Iperparametri	30
3.4	Caratterizzazione e Interpretazione dei topic	32
3.4.1	Relazioni Latenti e Strutture di Co-occorrenza	37
3.4.2	Validazione dei topic con i Dati di Registro	38

4	Modellazione Non-Parametrica: Hierarchical Dirichlet Process	43
4.1	Dal Modello Parametrico al Non-Parametrico: L'Hierarchical Dirichlet Process	43
4.1.1	Inferenza: Collapsed Gibbs Sampling	47
4.2	Implementazione e Configurazione dell'HDP	47
4.2.1	Iperparametri Iniziali e Setup del Modello	47
4.2.2	Ottimizzazione degli Iperparametri	48
4.2.3	Stima del Modello e Diagnostica di Convergenza	49
4.3	Caratterizzazione e interpretazione dei topic	49
4.3.1	Prevalenza dei topic nel corpus	50
4.3.2	Interpretazione semantica dei nove topic	51
4.3.3	Analisi di correlazione tra i topic	57
4.3.4	Validazione dei topic con i dati di registro	57
4.4	Valutazione Comparativa tra LDA e HDP	60
4.4.1	Analisi Quantitativa delle Metriche di Performance	61
4.4.2	Discussione Critica e Sintesi dei Modelli	62
4.5	Analisi delle covariate demografiche sui topic HDP	62
5	Considerazioni Finali	65
5.1	Sintesi del Lavoro e Obiettivi Raggiunti	65
5.2	Risultati Chiave e Confronto Metodologico: LDA vs HDP	66
5.3	Implicazioni Pratiche e Impatto Operativo	67
5.4	Limiti della Ricerca	67
5.5	Sviluppi Futuri	68
	Appendice	71
1	Prompt completo utilizzato per l'addestramento	71
2	Codice e materiali di supporto	72
	Bibliografia	73

Elenco delle figure

2.1	Confronto tra la qualità dei dati e la disponibilità di note narrative per anno di ultimo passaggio, dal 2009 al 2026.	5
2.2	Distribuzione degli assistiti Caritas Treviso (post-2020, $n = 2053$) per genere.	15
2.3	Distribuzione degli assistiti per stato civile.	16
2.4	Distribuzione degli assistiti per classe di età.	16
2.5	Distribuzione degli assistiti per area geografica di provenienza. . . .	17
2.6	Distribuzione dei macro bisogni per utente.	19
2.7	Distribuzione dei macro interventi per utente.	20
3.1	Rappresentazione grafica del modello LDA	24
3.2	Andamento della perplessità in-sample al variare del numero di topic K da 2 a 25. Modelli LDA stimati con metodo Gibbs (1.000 iterazioni).	28
3.3	Confronto della topic coherence media (Probabilistic Coherence) al variare del numero di topic K . Modelli LDA stimati con metodo Gibbs (1.000 iterazioni).	29
3.4	LDA: topic 1	34
3.5	LDA: topic 2	34
3.6	LDA: topic 3	35
3.7	LDA: topic 4	35
3.8	LDA: topic 5	36
3.9	LDA: topic 6	36
3.10	Heatmap dei tassi di penetrazione percentuale dei macro bisogni. . .	41
3.11	Heatmap dei tassi di penetrazione percentuale dei macro interventi. . .	42
4.1	Rappresentazione grafica del modello Hierarchical Dirichlet Process (HDP)	46
4.2	HDP: topic 0	52
4.3	HDP: topic 1	53
4.4	HDP: topic 2	53
4.5	HDP: topic 3	54

4.6	HDP: topic 4	54
4.7	HDP: topic 5	55
4.8	HDP: topic 6	55
4.9	HDP: topic 7	56
4.10	HDP: topic 9	56
4.11	Heatmap dei tassi di penetrazione percentuale dei macro bisogni. . .	59
4.12	Heatmap dei tassi di penetrazione percentuale dei macro interventi.	60

Elenco delle tabelle

2.1	Confronto delle misure descrittive del corpus (prima parte)	13
2.2	Confronto delle metriche di modello e lunghezza (seconda parte) . .	14
3.1	Confronto delle prestazioni del modello LDA ($K = 6$) al variare delle configurazioni e dei pacchetti.	31
3.2	Matrice di correlazione tra i sei topic del modello LDA.	37
3.3	Bisogni specifici e caratterizzanti per ciascun topic (LDA) stimati tramite Rischio Relativo bayesiano ($RR > 1$ e $HPD_{Lower} > 1$). . . .	39
3.4	Interventi specifici e caratterizzanti per ciascun topic (LDA) stimati tramite Rischio Relativo bayesiano ($RR > 1$ e $HPD_{Lower} > 1$). . . .	39
4.1	Presenza nei topic con probabilità maggiore o uguale al 5,0%	50
4.2	Matrice di correlazione tra le distribuzioni dei topic attivi nell'HDP.	57
4.3	Bisogni specifici e caratterizzanti per ciascun topic stimati tramite Rischio Relativo bayesiano ($RR > 1$ e $HPD_{Lower} > 1$).	58
4.4	Interventi specifici e caratterizzanti per ciascun topic stimati tramite Rischio Relativo bayesiano ($RR > 1$ e $HPD_{Lower} > 1$).	58
4.5	Sintesi comparativa degli indici prestazionali tra LDA e HDP. . . .	61
4.6	Profilazione socio-demografica dell'utenza distinta per topic HDP. .	63

Capitolo 1

Introduzione, ambito di applicazione e metodologie d'indagine

Nel Terzo Settore saper interpretare e prevedere i bisogni della popolazione è condizione imprescindibile per poter programmare ed erogare interventi di efficacia. La Caritas Diocesana di Treviso, ad esempio, è chiamata a svolgere quotidianamente un ruolo di presidio fondamentale sul territorio accogliendo persone e famiglie in stato di bisogno e in condizioni di vulnerabilità o vera e propria marginalità sociale ed economica.

La principale porta di accesso a questo sistema di supporto sono i Centri di Ascolto, nei quali operatori e volontari incontrano e dialogano con le persone che manifestano difficoltà. In questi contesti viene generata una grande mole di dati. I registri, infatti, seppur consentano di classificare in modo rigoroso i servizi offerti, come ad esempio la consegna di pacchi alimentari o l'erogazione di piccoli contributi economici a carattere temporaneo, raccontano solo una parte della storia. Le note di colloquio ne rappresentano invece la reale ricchezza e complessità. Si tratta di documenti liberi, diari di vita, note clinico-sociali, i racconti dei volontari che descrivono l'incontro e che ne custodiscono l'essenza.

Storicamente, questo tipo di informazione di natura qualitativa e non strutturata tende a restare nascosta o completamente ignorata nelle fasi di analisi aggregata e di programmazione territoriale. Ciò accade, principalmente, per le difficoltà connesse a trattare enormi volumi di testo libero e, inoltre, per l'estrema delicatezza delle questioni etico-legali legate alla protezione dei dati personali e che trovano nel Regolamento Europeo sulla Privacy (GDPR) un valido e imprescindibile punto di riferimento.

Il presente lavoro di tesi si colloca proprio in questo ambito con l'obiettivo di

sviluppare e applicare una metodologia basata sulle tecniche di *topic modeling* per dare valore all'informazione racchiusa nelle note di colloquio della Caritas Diocesana di Treviso trasformandola in conoscenza strutturata a supporto dei processi decisionali.

L'architettura è stata sviluppata affrontando una doppia sfida, una di metodo e una di merito. Da un lato c'era il problema della privacy e della sovranità del dato: il vincolo di garantire la privacy degli utenti ha imposto la rigorosa scelta di utilizzare solo infrastrutture che operano in locale, senza inviare alcuna informazione a servizi di terze parti. Un vincolo che ha stimolato la ricerca e la sperimentazione di modelli statistico-computazionali meno dispendiosi dal punto di vista computazionale, ma in grado di assicurare elevati livelli di accuratezza e interpretabilità. Dall'altro, si è posto il problema di svelare le cause e i fattori latenti: adottando ad esempio approcci probabilistici come l'LDA e la sua estensione non parametrica HDP, il lavoro di ricerca si propone di andare oltre la rigidità dei registri per far emergere i problemi che, come la precarietà abitativa, i blocchi burocratico-amministrativi, la solitudine o la fragilità familiare, spesso anticipano e generano la richiesta di aiuto materiale.

Sul fronte applicativo e di trasferibilità, l'infrastruttura tecnologica implementata dimostra come sia possibile coniugare la massima privacy con le informazioni da grandi dataset narrativi. I risultati prodotti rappresentano uno strumento ad alto valore informativo, potenzialmente estendibile ad altre realtà del Terzo Settore.

Capitolo 2

I Dati Caritas: Pre-processing e Profilazione

2.1 Il contesto operativo e l'origine dei dati

In questo capitolo viene presentato il contesto istituzionale di riferimento e viene descritta la struttura dei dati utilizzati nell'analisi, illustrando l'evoluzione storica del sistema di rilevazione e il processo di selezione del campione finale.

2.1.1 Il contesto istituzionale e i dati di registro

Caritas Italiana è l'ente pastorale della Conferenza Episcopale Italiana impegnato a promuovere la testimonianza cristiana della carità nella comunità ecclesiale e civile. Nata nel 1971 su iniziativa di Papa Paolo VI, opera attraverso una fitta rete di Caritas Diocesane distribuite su tutto il territorio nazionale, con il compito di animare le comunità locali, promuovere iniziative di solidarietà e gestire servizi di ascolto e accompagnamento per le persone in difficoltà. Secondo i rapporti annuali di povertà e di esclusione sociale editi dall'ente, lo scenario di bisogno che arriva dai Centri di Ascolto Diocesani rappresenta uno dei più affidabili e sistematici osservatori sulle povertà in Italia perché registra in modo longitudinale le storie di disagio di un bacino di utenza che altrimenti sarebbe più difficile intercettare da parte delle statistiche ufficiali [2]. La Caritas Diocesana di Treviso è una realtà che si colloca all'interno di questo contesto come struttura operativa territoriale della Diocesi di Treviso, con una rete di centri di ascolto e di servizi dislocati nel capoluogo e nei comuni della provincia, ci si focalizzerà sui dati della sede centrale di Treviso. Ogni persona che accede ai servizi dell'ente viene inserita in un unico sistema informativo: gli operatori inseriscono digitalmente una scheda anagrafica e sociale, che viene aggiornata ogni volta che il soggetto si presenta nelle strutture. Su

questa scheda vengono raccolte informazioni di tipo socio-demografico, come sesso, età, nazionalità, stato civile, titolo di studio, condizione professionale, composizione del nucleo familiare, ma anche informazioni relative ai bisogni, agli interventi, alle richieste ed infine, in forma di testo libero, le note di colloquio scritte dagli operatori al termine di ogni colloquio. È proprio a questo archivio e più in generale alle sue componenti testuali e socio-demografiche che fanno riferimento i dati analizzati in questa tesi.

2.1.2 Analisi temporale della qualità informativa e criteri di selezione

Il database totale della Caritas di Treviso conta, al momento del prelievo, 7.290 utenti in carico dal 2009. Di questi 4.344 hanno almeno una nota a testo associata al loro profilo; ma molte di queste sono note prodotte automaticamente dal gestionale, messaggi del tipo “Modifica valore variabile” seguiti da tecnicismi, che non hanno significato e che quindi inquinano il corpus di testo e che devono essere eliminate prima di ogni analisi. Una volta eliminati questi refusi delle applicazioni le note che contengono vero e proprio materiale informativo scendono a 3.412.

L’analisi temporale della qualità del dato evidenzia un profondo cambiamento di grande interesse dal punto di vista metodologico, che emerge in figura 2.1. Il periodo che va dal 2009 al 2019 è ancora caratterizzato da un’elevata presenza di valori non specificati o vaghi nelle principali variabili di interesse, come per esempio lo stato civile o il titolo di studio, e la quota di dati mancanti arriva al suo massimo intorno al 2018-2019. In questo contesto si osserva a partire dal 2018 una progressiva crescita della quota di note testuali presenti, segnale di una maggiore attenzione alla tracciabilità delle schede. Si verifica però un’interruzione con un nuovo picco di note mancanti tra il 2019 e il 2020, legato all’emergenza COVID-19 per la quale la Caritas non ha potuto effettuare i colloqui con gli assistiti in presenza. A partire dal 2020, grazie al cambio di gestione e alla maggiore accuratezza nella compilazione delle schede, la qualità informativa migliora: le variabili socio-demografiche sono molto più complete e le note testuali sono più frequenti e ricche di contenuto. L’intersezione tra questi due aspetti, ovvero la qualità delle variabili strutturate e la disponibilità di note testuali utili, ha indotto a individuare il sottoinsieme di analisi degli assistiti il cui ultimo passaggio è a partire dal 1° gennaio 2020, e che hanno almeno una nota testuale. Il campione finale risultante è di 2.053 casi.

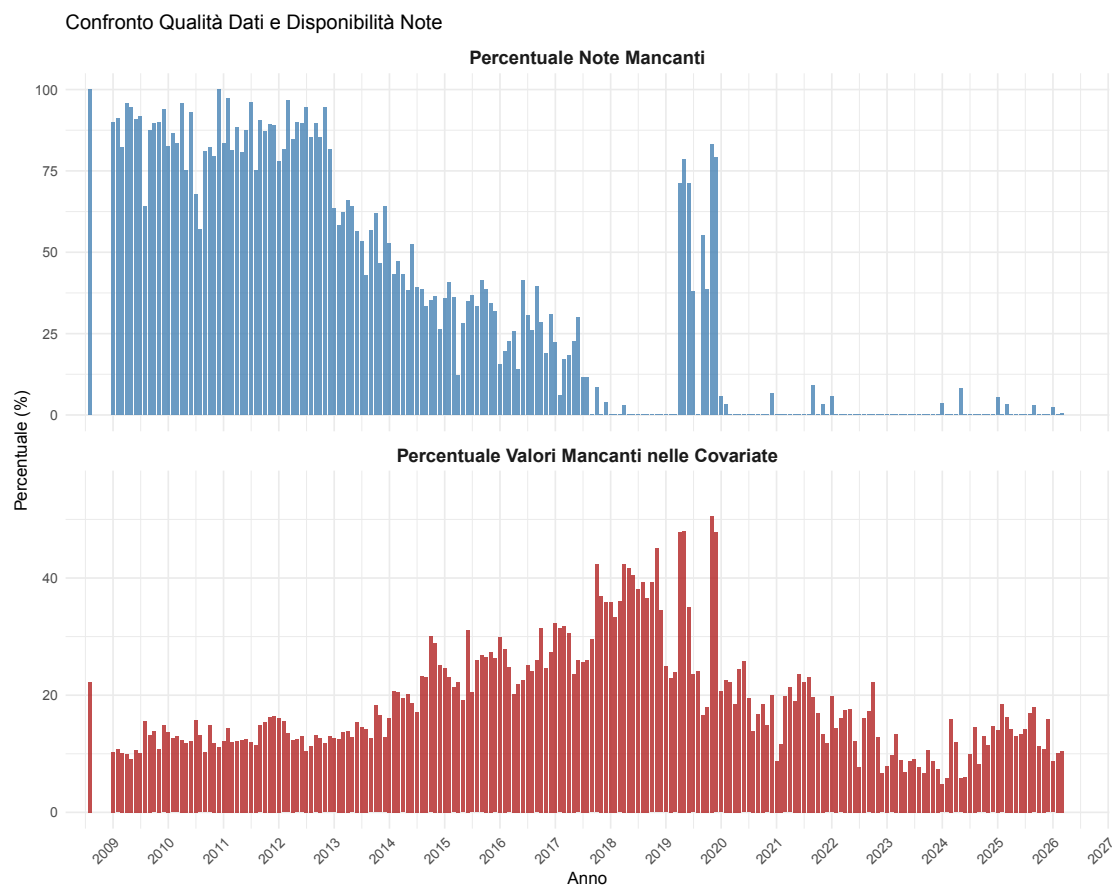


Figura 2.1: Confronto tra la qualità dei dati e la disponibilità di note narrative per anno di ultimo passaggio, dal 2009 al 2026.

2.2 Trattamento e pre-processing delle note

La fase di trattamento preliminare (*pre-processing*) dei testi, sia per l'estrazione dati sia per l'individuazione dei topic, è fondamentale per garantire la qualità delle analisi successive. Gli algoritmi di modellazione statistica applicati a testi necessitano infatti di un corpus di partenza ben pulito, perché la presenza di informazioni ridondanti può facilmente distorcere la stima delle componenti e rendere poco interpretabili i topic generati. La pulizia non è dunque una semplice formalità, ma una vera e propria impostazione di fondo per circoscrivere le variabili di interesse e ottenere un lessico compatto ma informativo.

La realizzazione di un vocabolario ridotto risponde a motivi ben precisi di natura metodologica e computazionale. In primo luogo un testo grezzo contiene elementi ripetitivi, forme flessive, secondarie e altri costrutti privi di valore informativo che sovraccaricano la matrice documenti-termini aumentando la dimensione dello spazio dati ma senza fornire alcun contributo esplicativo. In secondo luogo la scelta delle forme testuali impedisce che l'informazione si disperda tra vocaboli rari e privi di significato, consentendo al modello di fissare l'attenzione sui concetti principali. Gestendo il pre-processing si può cioè calibrarne la granularità in modo da mantenere la capacità di far emergere i topic latenti nascosti nel corpus.

2.2.1 Costruzione del vocabolario iniziale

La fase iniziale del lavoro riguarda la pulizia preliminare delle componenti testuali, un passaggio fondamentale per rimuovere anomalie e informazioni trascurabili prima di avviare le elaborazioni successive.

In un primo momento l'intervento si focalizza sulla fruibilità dei contenuti scritti. Tale operazione prevede la correzione dei difetti legati ai caratteri speciali e alla grafia accentata, dovuti a disallineamenti nella codifica originaria e risolti mediante la conversione nello standard ASCII con la contestuale eliminazione delle rimanenze tramite espressioni regolari. In seguito si procede con la cancellazione degli elementi di disturbo prodotti in automatico dai software gestionali, quali ad esempio le stringhe di registrazione del sistema e i tracciati delle variazioni anagrafiche.

Una volta che il testo è ripulito, si passa alla fase di normalizzazione testuale. In questa fase tutto il testo viene trasformato in caratteri minuscoli e vengono eliminate la punteggiatura e le cifre numeriche. Questa procedura impedisce che un singolo termine venga conteggiato più volte a causa della presenza di lettere

maiuscole o dell'adiacenza a segni di interpunzione. Ad esempio, le parole “Mensa”, “mensa” e “mensa.” dopo questa fase verranno considerate come la stessa parola.

L'operazione si conclude con la selezione delle forme grammaticali rilevanti per la ricerca. Vengono escluse le parole prive di significato autonomo come articoli o preposizioni tramite l'impiego di elenchi di stopwords, basati sui lavori originari di Martin Porter [14], unite alla rimozione delle forme brevi di lunghezza inferiore ai tre caratteri. Risultano inoltre scartati alcuni costrutti specifici del contesto organizzativo ma poco informativi per le finalità dello studio, come la dicitura relativa al cambiamento del recapito. La sistemazione finale della spaziatura interna restituisce un corpus ordinato, predisposto per i successivi passaggi.

2.2.2 Approcci metodologici di riduzione del lessico

A seguito della fase di pulizia preliminare e di unificazione della struttura, il materiale scritto rappresenta il punto di partenza per l'impiego di procedure avanzate nell'ambito dell'analisi linguistica. A partire da questa versione del testo, priva di incongruenze formali ma ancora grezza sotto il profilo semantico e morfologico, si passa all'adozione di differenti approcci per la selezione delle forme grammaticali. Tali interventi mirano a ridurre ulteriormente la dimensione del dizionario e ad accrescere il valore informativo delle componenti estratte. Le tecniche utilizzate spaziano dall'uso di strumenti classici dell'analisi testuale fino a soluzioni basate su modelli di intelligenza artificiale, mantenendo il rispetto della riservatezza dei dati. Nei paragrafi seguenti vengono analizzate nel dettaglio le diverse metodologie, evidenziando il contributo fornito alla costruzione del loro vocabolario definitivo.

Lo stemming

Lo stemming [14] è un procedimento pratico e veloce dal punto di vista computazionale per affrontare la variabilità morfologica dei testi. L'idea di fondo è che forme flessive diverse ma vicine sul piano semantico, ad esempio forme di un verbo o flessioni di genere e numero, condividano la stessa radice semantica. Riducendo ogni parola al proprio tronco originale si ottiene un dizionario più piccolo, in cui le diverse varianti di uno stesso concetto sono rappresentate da un unico elemento. Per capire come funziona questa trasformazione si consideri la frase “i clienti acquistano prodotti e hanno acquistato servizi”. Utilizzando l'algoritmo di stemming, le forme “acquistano” e “acquistato” vengono meccanicamente private della desinenza per risalire al tronco comune “acquist”, perdendo così le informazioni grammaticali

relative al tempo e alla persona. Analogamente, i termini al plurale come “prodotti” o “servizi” vengono riportati alla radice. Sebbene veloce ed efficace nel raggruppare forme simili, la tecnica ha una debolezza intrinseca dovuta alla mancanza di un vocabolario semantico di riferimento. In quanto si basa su regole fisse di privazione dei suffissi, la radice ottenuta non sempre corrisponde ad una parola valida nella lingua italiana ma è un semplice frammento. Tale rigidità può portare a raggruppare parole diverse perché accomunate da una vaga somiglianza grafica, creando ambiguità e perdendo le differenze di significato. Per questo i termini così tronchi sono spesso poco leggibili e difficili da interpretare, rendendo meno immediata e scorrevole la lettura dei risultati prodotti dal modello.

La lemmatizzazione

A differenza dello stemming, orientato a logiche di taglio meccanico, la lemmatizzazione costituisce una procedura di analisi linguistica e morfologica che riconduce ogni voce al proprio lemma, ovvero la forma base presente nel dizionario. Grazie all’ausilio di risorse grammaticali orientate alla dipendenza, il procedimento guarda anche il contesto grammaticale della frase per assegnare la corretta parte del discorso a ogni componente. Per osservare il funzionamento della procedura e la differenza rispetto allo stemming, è possibile considerare di nuovo l’espressione “i clienti acquistano prodotti e hanno acquistato servizi”. Mentre la tecnica di prima produceva delle parole tronche, incomplete, la lemmatizzazione ricondurrebbe la parola “acquistano” e la parola “acquistato” alla forma verbale all’infinito “acquistare”, e il sostantivo “prodotti” al singolare “prodotto”. Ogni termine viene quindi trasformato nella rispettiva forma canonica, preservando la corretta struttura della lingua. Il metodo garantisce due importanti vantaggi a livello analitico. In primo luogo si eliminano le imprecisioni e le ambiguità introdotte dalla meccanica rigida del precedente metodo e si ottiene un corpus composto interamente da parole vere. In secondo luogo, pur riducendo la dimensione del vocabolario attraverso la generalizzazione delle forme flesse, il metodo mantiene pienamente leggibile il testo, preparando il terreno per le successive fasi di modellazione statistica. Per l’applicazione di tale metodo sono stati sfruttati i modelli linguistici basati sulla libreria R `udpipe` [12].

L’uso di un Large Language Model locale

L’adozione di un large language model eseguito in locale è una possibile soluzione ai limiti delle metodologie di elaborazione del linguaggio naturale classico. Sebbene

le procedure precedenti abbiano permesso di ricondurre i termini alla forma base, l'analisi del testo mostra come tali strumenti non riescano a ridurre fortemente la dimensione del vocabolario. Nel testo delle note permangono infatti elementi di disturbo, come frasi di circostanza, un numero elevato di espressioni ridondanti, per cui la matrice documenti-termini risulta ancora troppo grande e troppo sparsa perché si possa procedere alle fasi di *topic modeling*.

La scelta di utilizzare una risorsa operativa in locale, nello specifico il modello open-weight Qwen nella versione da un miliardo e mezzo di parametri tramite l'infrastruttura Ollama, risponde a vincoli di natura normativa e di sicurezza dei dati. Nel trattamento di dati che contengono informazioni riservate o personali, infatti, l'utilizzo di un servizio in cloud o di un'API esterna comporterebbe l'invio dei dati al di fuori della propria sfera di controllo e, quindi, il rischio che il trattamento venga effettuato in violazione del GDPR (*General Data Protection Regulation*). Nel rispetto dei principi di tutela della riservatezza fin dalla progettazione, l'architettura tecnica è stata configurata per isolare l'elaborazione all'interno della postazione di lavoro, evitando il rischio di intercettazioni o memorizzazioni esterne e garantendo la piena titolarità del dato.

La fase di valutazione, prima di consolidare la scelta su Qwen 1.5B, ha previsto il test e il confronto di diverse alternative open-weight, tra cui modelli come LLaMA 3 (nelle versioni da 8B), Mistral 7B e Gemma 2 (9B). Le architetture caratterizzate da un maggior numero di parametri hanno evidenziato una notevole ricchezza espressiva nella gestione del testo, ma date le scarse risorse hardware a disposizione sulla postazione locale, in particolare i limiti di memoria VRAM e la scarsa potenza di calcolo, ne rendevano l'esecuzione lenta e insostenibile su ampi volumi di dati. La scelta finale è stata un modello da 1,5 miliardi di parametri, individuato come il punto di equilibrio ottimale tra sostenibilità computazionale, velocità di elaborazione e accuratezza del risultato.

Il Large Language Model è stato impiegato solo come strumento di compressione del significato e di pulizia testuale avanzata. L'efficacia di tale approccio dipende tuttavia dalla struttura delle istruzioni fornite. Durante la fase di prova sono state valutate differenti formulazioni del testo della richiesta (*prompt*). I comandi troppo brevi producevano risposte incomplete che tralasciavano aspetti rilevanti quali l'anonimizzazione e la rimozione delle registrazioni del software. D'altra parte, richieste troppo lunghe e articolate finivano per sovraccaricare il modello che tendeva così a trascurare alcune parti delle richieste o a produrre testi non fedeli al testo di partenza. Si è inoltre tentato di integrare alcuni esempi all'interno del prompt; tuttavia, ciò ha comportato la mera trascrizione dell'esempio stesso nel

testo di diverse note.

Si è definita un'istruzione strutturata basata su indicazioni precise e sintetiche. La procedura prevede l'oscuramento delle informazioni sensibili come nomi o recapiti, la cancellazione dei messaggi automatici per isolare il solo contenuto umano e la correzione delle imprecisioni ortografiche. Vengono inoltre effettuate la conversione in caratteri minuscoli con riduzione alla forma base, la rimozione degli elementi grammaticali deboli e la selezione esclusiva delle parole chiave portatrici del significato essenziale (Appendice).

L'impiego del modello su una risorsa hardware contenuta, caratterizzata da 8 gigabyte di memoria temporanea (RAM), ha reso necessaria la progettazione di una sequenza di calcolo specifica per evitare il blocco del sistema. La procedura suddivide ogni nota testuale in blocchi di ampiezza ridotta, evitando di superare la capacità di contesto della macchina e contenendo l'impronta di memoria durante la fase di elaborazione. I segmenti vengono processati uno alla volta e i termini risultanti vengono concatenati in un'unica lista per preservare l'ordine originale e la frequenza delle parole. A questo si affianca una gestione attiva delle risorse mediante la liberazione periodica della memoria volatile (RAM) ad intervalli regolari.

L'approccio modellistico linguistico ha consentito di comprimere il vocabolario generando un insieme di parole chiave compattato con poco rumore ed eliminando la maggior parte di dati personali. Il compromesso tra compressione semantica e risorse a disposizione ha però richiesto un tempo piuttosto lungo impiegando all'incirca due giorni di calcolo non stop per processare l'intero archivio. A fronte di questo sforzo, che evidenzia chiaramente il bilanciamento della metodologia, l'esecuzione era certamente più lenta di quella dei sistemi tradizionali ma la soluzione ha raggiunto un grado di pulizia, anonimizzazione e selezione di senso che non si sarebbe potuto ottenere con semplici strategie lessicali. Per una parte del corpus, infatti, l'elaborazione produce risultati anche di buona qualità e conformi alle regole di minimizzazione dei dati precedentemente descritte ma vi sono casi in cui emergono difficoltà da parte del modello che, infatti, non è così potente da generare sempre output ottimali o sufficientemente raffinati. A volte ciò si traduce nella generazione di testi senza valore costituiti dalla ripetizione di caratteri dell'alfabeto racchiusi tra parentesi (es: a e p c s d t r l u i o n m b g f h j k q w y z), altre volte nella restituzione di stringhe spezzettate formate da parole separate, da sequenze di punti e punti interrogativi (es: a?|sospesa|mensana|riattivato|servizio|tessera|senza|...). Si riscontrano altresì numerosi casi in cui il contenuto informativo della nota risulta quasi interamente compromesso, a favore di sequenze di parole casuali e prive di senso compiuto. Questi casi testimoniano il fragile bilanciamento che esiste

tra l'ambizioso obiettivo di ottenere una forte compressione del vocabolario e le limitazioni imposte dalla scarsa dotazione hardware.

Un approccio ibrido sequenziale

La procedura individuata per la preparazione del testo è un approccio ibrido e sequenziale, che si pone l'obiettivo di aggirare le limitazioni sia dei singoli metodi tradizionali sia dei modelli di linguaggio locali. In particolare questa procedura integra tecniche di elaborazione del linguaggio naturale basate sulla libreria **spaCy**, modelli di rappresentazione vettoriale effettuati tramite la struttura *SentenceTransformers* (strumento di intelligenza artificiale che trasforma intere frasi in elenchi di numeri (vettori) mantenendone il significato, in modo che frasi con lo stesso senso finiscano vicine tra loro.) e algoritmi di raggruppamento non supervisionato. La procedura si sviluppa in quattro livelli principali, accompagnati da verifiche finali sulla qualità della base dati.

Il primo livello è la selezione semantica a livello di frase, basata su idee di analisi testuale orientata. Utilizzando la trasformazione di frasi in vettori numerici tramite il modello all-MiniLM-L6-v2, ciascun documento viene suddiviso in segmenti elementari convertiti in vettori densi. Questi vengono quindi comparati, tramite misure di similarità, ad una istruzione in linguaggio naturale contenente i concetti di maggiore interesse, come ad esempio la condizione occupazionale, i servizi di supporto e le situazioni di disagio sociale. Questa operazione consente di scartare formule di cortesia, richieste burocratiche o parti di testo non rilevanti prima di procedere con l'analisi sintattica.

La seconda fase si concentra sulla lemmatizzazione avanzata, sul riconoscimento della categoria grammaticale delle parole e sulla protezione dei dati riservati mediante un modulo di identificazione delle entità. Con il modello statistico di grandi dimensioni per la lingua italiana fornito da **spaCy** il testo viene analizzato per identificare e rimuovere i nomi di persona, affiancando a questo procedimento un elenco manuale di parole da rimuovere per catturare eventuali elementi sfuggiti al filtro automatico. Il vocabolario viene poi filtrato mantenendo soltanto i sostantivi ed aggettivi, considerati i contenuti portatori di significato, e rimuovendo le forme di lunghezza inferiore a tre caratteri per eliminare acronimi e frammenti di testo.

La terza fase introduce un criterio di selettività basato sulla frequenza nei documenti, indirizzato a rimuovere i lemmi troppo rari che compaiono in un numero di documenti inferiore ad una soglia minima, fissata a due. Su un vocabolario iniziale di 16.450 lemmi l'imposizione di tale restrizione ne determina la rimozione

di 7.878 forme isolate o scarsamente rappresentate, diminuendo in maniera rilevante la sparsità della matrice e mantenendo un repertorio di 8.572 lemmi stabili.

La quarta fase riduce il vocabolario tramite il raggruppamento dei vettori associati ai termini rimasti. Attraverso l'algoritmo *MiniBatchKMeans* i vocaboli semanticamente vicini o sinonimi vengono aggregati in cluster coerenti nello spazio vettoriale e ad ogni cluster viene associato il lemma più frequente al suo interno. Con questa operazione il vocabolario è stato ridotto a 4.682 termini distinti. Il procedimento si chiude con un controllo sul corpus, eliminando le righe che sono risultate vuote a seguito dei filtri e i casi duplicati, per cui la dimensione finale del dataset è pari a 2.053 casi puliti e pronti per le successive analisi.

2.2.3 Valutazione comparativa tra le diverse tecniche utilizzate

Per valutare le prestazioni delle diverse strategie adottate e motivare la scelta della configurazione finale si è proceduto a un'analisi comparativa basata su indicatori quantitativi e metriche di valutazione dei modelli di analisi tematica. A tal fine sono stati stimati modelli *Latent Dirichlet Allocation* (LDA) sui diversi testi fissando a sei il numero di topic, un valore scelto in accordo con i risultati emersi nelle successive fasi dell'analisi, osservando l'evoluzione della struttura lessicale e delle prestazioni.

I parametri considerati si dividono in statistiche descrittive del testo e indicatori di aderenza del modello statistico. Il numero totale di parole è la quantità totale di parole contenute nei testi delle note dopo il trattamento, evidenziando la contrazione della massa testuale con l'applicazione di filtri più selettivi. Il vocabolario unico è il numero di parole univoche e misura la dimensione complessiva del dizionario. La sparsità percentuale esprime la proporzione di celle prive di informazione nella matrice dei documenti e dei termini, evidenziando una condizione tipica dell'analisi linguistica in cui la maggior parte dei vocaboli non compare nella totalità dei testi. La lunghezza media del documento calcola il numero medio di parole per nota e fornisce una misura della concisione raggiunta. La perplessità è una misura che valuta quanto bene il modello (*Latent Dirichlet Allocation*) riesce a spiegare i dati, dove valori più bassi indicano un miglior adattamento al testo. Infine la coerenza media misura la qualità semantica dei topic rilevati valutando quanto le parole chiave di un topic sono affini tra loro, dove valori più alti indicano gruppi di parole meglio interpretabili.

L'analisi approfondita dei dati mostra differenze importanti tra le strategie adottate. Il testo di partenza ha un volume complessivo di oltre due milioni di parole e un dizionario di oltre quarantamila vocaboli, con una sparsità di oltre il 99%. In queste condizioni, l'estrazione di topic diventa dispersiva e molto rumorosa. Applicando le prime fasi di pulizia e normalizzazione, si riduce di circa la metà il numero totale di parole, arrivando a poco più di un milione, mentre si mantiene un vocabolario ampio e un'elevata perplessità del modello.

Facendo un confronto tra le tecniche di riduzione del vocabolario, si osservano andamenti differenti. Lo stemming riduce il numero dei vocaboli a circa 18.000 e registra una perplessità più bassa rispetto al testo pulito, ma l'uso di stemmi impuri compromette la leggibilità. La lemmatizzazione, invece, offre un'interpretabilità più elevata con quasi 30.000 parole uniche, ma la dimensione del dizionario resta troppo alta per una modellazione compatta. L'uso del modello linguistico, invece, produce una riduzione radicale, in quanto la lunghezza media dei documenti passa da oltre 500 parole a sole 139. Purtroppo, questa forte riduzione genera una matrice sparsa per il 99,59%, con un peggioramento degli indici di coerenza e perplessità causato dalle rare fratture introdotte dal modello su un testo così limitato.

La soluzione ibrida risulta la più bilanciata sotto ogni profilo. Pur mantenendo una lunghezza media della nota di circa 210 parole, questa strategia riduce il suo lessico a 4.682 termini unici, ovvero un decimo circa rispetto al testo originale. Tale riduzione consente di abbassare la sparsità al 97,53%, una percentuale ragionevole per questa tipologia di testi. Dal punto di vista delle performance del modello, l'approccio ibrido raggiunge il valore di perplessità decisamente più basso, pari a 679,59, segno che il suo adattamento ai dati è migliore rispetto a tutte le altre versioni. Tali evidenze confermano la solidità del metodo combinato, garantendo la stabilità e la qualità dei topic necessari per le successive elaborazioni quantitative.

Tabella 2.1: Confronto delle misure descrittive del corpus (prima parte)

Versione	Num Parole Totali	Vocabolario Unico	Sparsità Pct
testo originale	2.177.889	41.735	99,07%
testo pulito	1.087.367	35.781	99,14%
testo stemming	1.087.367	18.166	98,61%
testo lemmatizzato	1.081.943	29.488	99,06%
testo LLM	285.928	19.903	99,59%
testo ibrido	430.240	4.682	97,53%

Tabella 2.2: Confronto delle metriche di modello e lunghezza (seconda parte)

Versione	Lunghezza Media Doc	Perplexità	Coerenza media
testo originale	1.060,83	1.443,64	0,031
testo pulito	529,65	2.460,54	0,135
testo stemming	529,65	1.150,45	0,123
testo lemmatizzato	527,01	1.608,37	0,112
testo LLM	139,27	2.073,56	0,130
testo ibrido	209,57	679,59	0,146

2.3 Analisi descrittiva del campione e trattamento delle assenze informative

A partire dai casi selezionati a seguito dei passaggi di pulizia testuale e del vincolo temporale, l'attenzione si sposta sull'analisi descrittiva del campione e sul trattamento delle informazioni non pervenute. Anche all'interno di questo gruppo di utenti registrati dal 2020, varie caratteristiche socio-anagrafiche presentano una non trascurabile quota di assenze o di risposte generiche, utilizzate dagli intervistatori per indicare aspetti non rilevati nel corso del colloquio. Per ragioni di massima trasparenza, tali assenze sono state mantenute come categoria a sé nelle fasi di sintesi, evitando la ricostruzione automatica dei dati che avrebbero inserito ipotesi non verificabili. In particolare, per le variabili utilizzate come fattori esplicativi nei modelli a classi latenti, come il genere, l'età, lo stato coniugale e il continente di nascita, le posizioni mancanti sono state considerate come classe a sé, in modo da mantenere la dimensione complessiva del gruppo senza influenzare la stima dei parametri.

2.3.1 Distribuzione per genere

Il profilo dei complessivamente 2.053 iscritti alla Caritas di Treviso nel post 2020, evidenzia una struttura sostanzialmente sovrapponibile a quella dell'utenza dei Centri di Ascolto diocesani del Nord-Est. In termini di composizione per genere, a spiccare maggiormente è l'elevata incidenza della componente maschile che rappresenta l'84,1% dei casi, rispetto al 15,9% registrato per quella femminile. Una differenza che, come emerge dai report nazionali dell'Ente, si spiega da un lato, con la natura dei flussi migratori, inizialmente costituiti da uomini giunti da soli in attesa di famiglie da ricongiungere, dall'altro, con la scelta più frequente da parte delle donne di rivolgersi ai servizi sociali pubblici del territorio, piuttosto che alle strutture di matrice ecclesiale.

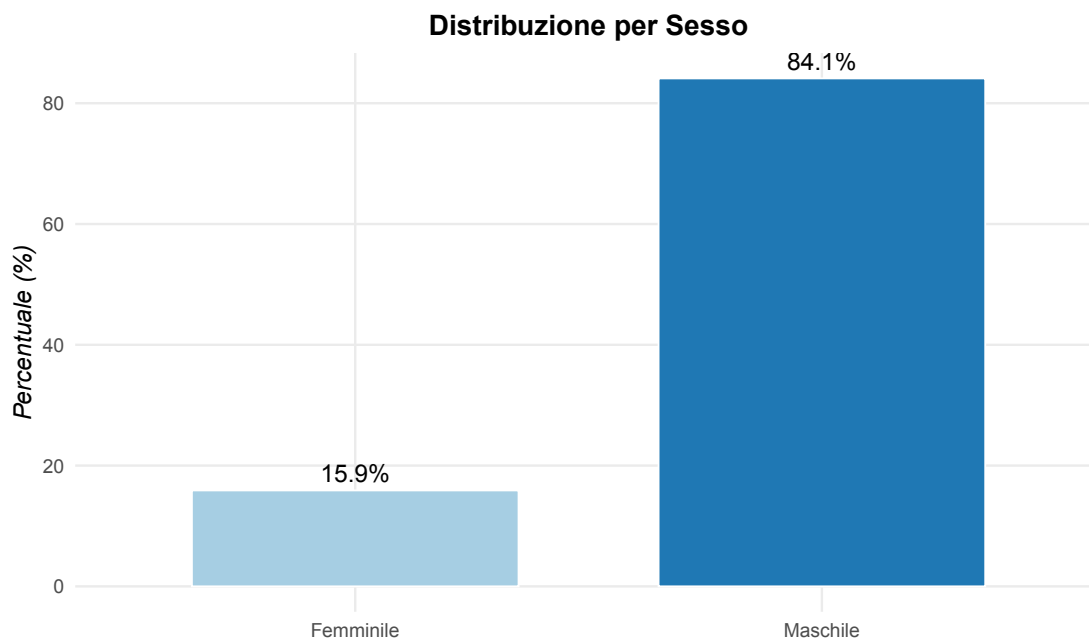


Figura 2.2: Distribuzione degli assistiti Caritas Treviso (post-2020, $n = 2053$) per genere.

2.3.2 Distribuzione per stato civile

Per quanto riguarda lo stato civile, la condizione di celibe o nubile è la più rappresentata, interessando il 46,2% del campione, seguita da quella di coniugato/a con il 28%. I separati o divorziati e i vedovi e le vedove rappresentano quote ridotte del campione, mentre in una quota significativa di casi, pari al 14,3%, lo stato civile non è noto. Prevale il celibato che, come abbiamo detto, è coerente con la giovane età e con la forte componente migratoria della popolazione che si sta esaminando, come vedremo nel paragrafo seguente.

2.3.3 Distribuzione per età

La distribuzione per classi di età ha disegnato il profilo di una popolazione che si trova per la maggior parte in età lavorativa. A guidare la classifica è infatti il gruppo composto dai giovani di età compresa fra i venticinque e i trentacinque anni che da solo contabilizza il 34,2% dei casi, seguito da quello dei trentacinque-quarantquattro anni con il 22,9%. I giovani adulti (diciannove-ventiquattro anni) rappresentano il 7,9% mentre i minori di diciotto anni e gli anziani di sessantacinque anni e oltre sono poco rappresentati, facendo registrare in entrambi i casi quote inferiori al 10%. Questa forte connotazione nei gruppi di età centrali

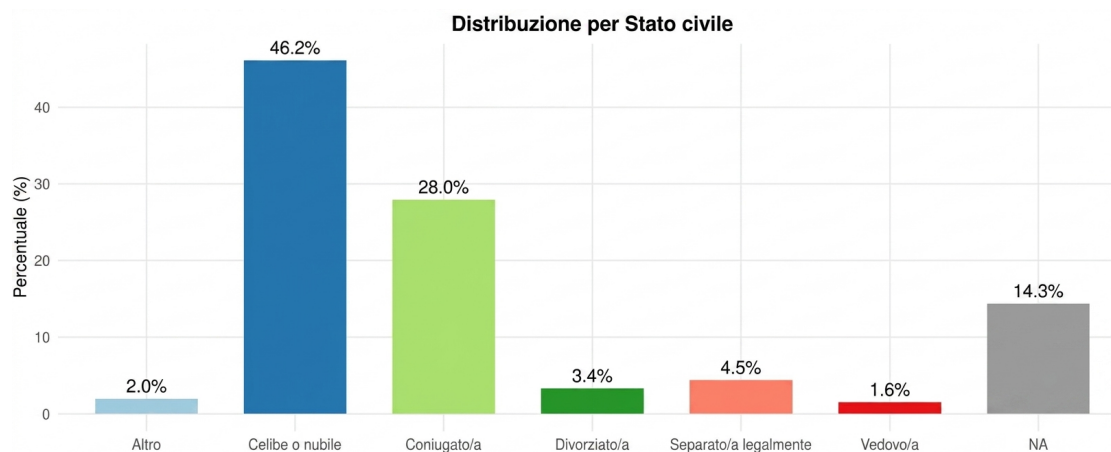


Figura 2.3: Distribuzione degli assistiti per stato civile.

dell'età lavorativa si traduce in un'età mediana che si attesta intorno ai 34 anni, a testimonianza di una popolazione nel pieno del proprio ciclo di formazione e riconoscimento familiare e professionale, più esposta ai rischi a livello di lavoro e di casa.

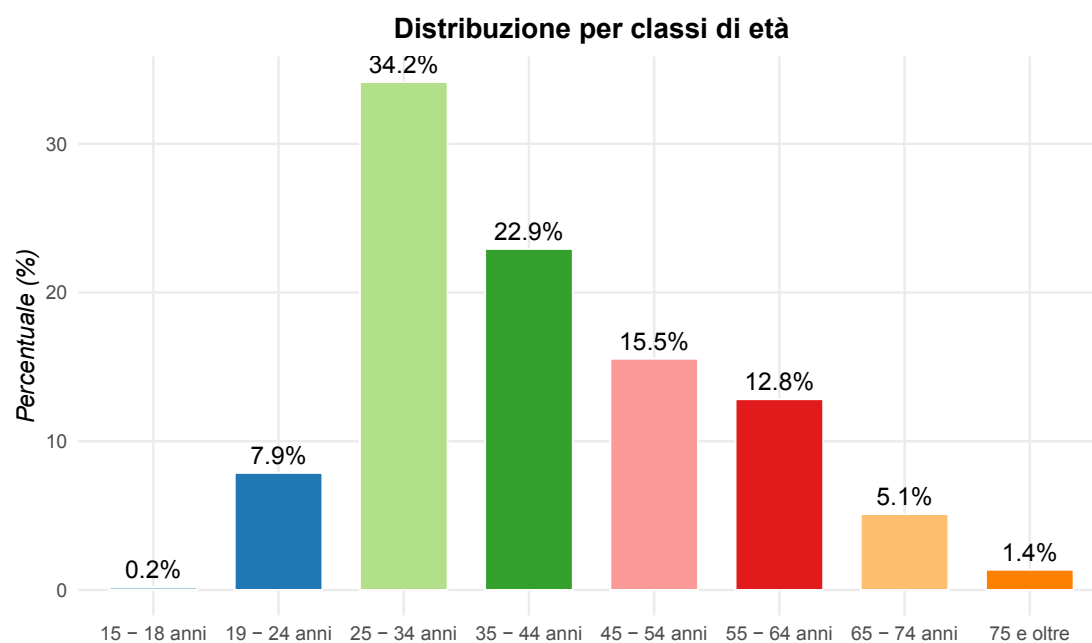


Figura 2.4: Distribuzione degli assistiti per classe di età.

2.3.4 Distribuzione per area geografica

Per l'intero campione analizzato la dimensione della provenienza geografica è l'aspetto che maggiormente lo caratterizza. Il primo gruppo per numerosità è

rappresentato dagli utenti originari dell'Asia meridionale che costituiscono il 31,5% del totale. Questo gruppo comprende Paesi come India, Pakistan, Bangladesh e Afghanistan. Seguono i cittadini provenienti dai Paesi dell'Unione Europea con il 22,6% e dell'Africa occidentale con il 22%. Il Nord Africa incide per il 14,7% mentre le altre aree geografiche, come ad esempio l'Europa centro-orientale, l'America Latina e le regioni dell'Africa orientale e meridionale, presentano complessivamente quote più contenute. Nel loro complesso le persone con cittadinanza non italiana rappresentano la maggioranza assoluta degli assistiti, confermando la componente migratoria come il principale tratto distintivo della popolazione che si rivolge alla struttura trevigiana.

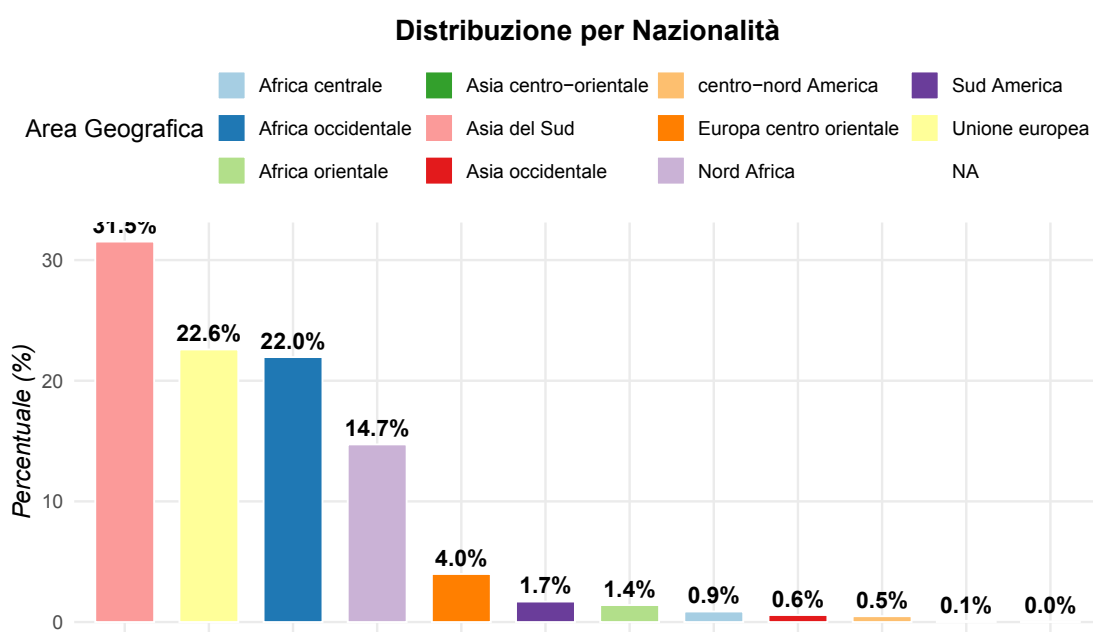


Figura 2.5: Distribuzione degli assistiti per area geografica di provenienza.

2.3.5 Mappatura dei bisogni degli assistiti

Per capire meglio i bisogni complessi che spingono le persone a rivolgersi ai servizi Caritas è importante ricordare che difficilmente ogni persona che si incontra in uno stato di difficoltà ha una sola problematica, ma quasi sempre ha molte difficoltà di fronte alle quali si trova in forte vulnerabilità. Su un totale di 2.053 persone che nel periodo hanno avuto almeno un bisogno rilevato, la distribuzione del numero di macro-bisogni diversi rilevati conferma la natura multidimensionale delle situazioni di disagio. Il primo quartile è 2, la mediana è 3, la media è 3,35, il terzo quartile è 4 e il valore massimo è 8. La mediana pari a 3, confermata da una media di

oltre tre bisogni per persona, è indice che le fragilità intercettate difficilmente sono monodimensionali; le persone incontrate si trovano infatti in media in presenza di tre diverse aree di privazione contemporaneamente.

Tra le singole macro-aree è l'abitare a raccogliere le problematiche più diffuse, in cima la mancanza di una casa stabile e accanto a essa la mancanza della residenza anagrafica, le situazioni di emergenza alloggio, gli arretrati condominiali, le richieste di sfratto e l'inadeguatezza abitativa, che confermano la casa come il punto di partenza fondamentale per ogni percorso di inclusione. L'area della povertà e dei problemi economici comprende le condizioni di povertà più gravi, legate a un reddito insufficiente, alla povertà e alla marginalità estrema delle persone senza dimora, e accanto a esse la mancanza eccezionale di liquidità, il debito, le difficoltà di gestione del bilancio familiare e i casi di usura.

I problemi legati al lavoro sono concentrati attorno alla disoccupazione che è la più frequente, seguita dalla precarietà, i licenziamenti, il lavoro nero, la cassa integrazione e la sottooccupazione. I bisogni legati all'immigrazione sono in linea con la presenza di stranieri molto numerosi e vedono predominare le richieste di asilo e le difficoltà amministrative, insieme alla condizione di profugo, irregolare, richiedente asilo, e la questione delle rimesse o del debito di viaggio.

L'area dell'educazione mette in luce difficoltà di natura culturale e linguistica in cui il fattore più rilevante sono le limitazioni linguistiche, affiancate da situazioni di analfabetismo. I problemi familiari sono espressione di difficoltà relazionali che si traducono nelle liti di coppia, nelle separazioni, nei conflitti generazionali, negli abusi, e nella gestione della genitorialità sola. Infine le difficoltà di natura sanitaria, legate alle dipendenze, alle vicende giudiziarie o alla disabilità, completano il quadro delle situazioni di grave marginalità presenti nell'archivio.

2.3.6 Mappatura degli interventi negli assistiti

L'insieme degli interventi proposti dalla rete dei servizi si articola in diverse aree di intervento che a loro volta si declinano in azioni volte a rispondere alle diverse esigenze dell'utenza. L'area dell'ascolto rappresenta l'ingresso principale e comprende colloqui di prima accoglienza, monitoraggi e progetti personalizzati e si estende in alcuni casi anche all'assistenza domiciliare e al sostegno relazionale.

Per quanto riguarda i beni e i servizi materiali, la risposta si focalizza sulle necessità di prima emergenza: la somministrazione di pasti presso la mensa, l'igiene personale, la distribuzione di alimenti e vestiario, i servizi di lavanderia e l'erogazione di buoni per l'acquisto di alimenti o per il trasporto. Nell'ambito abitativo, i

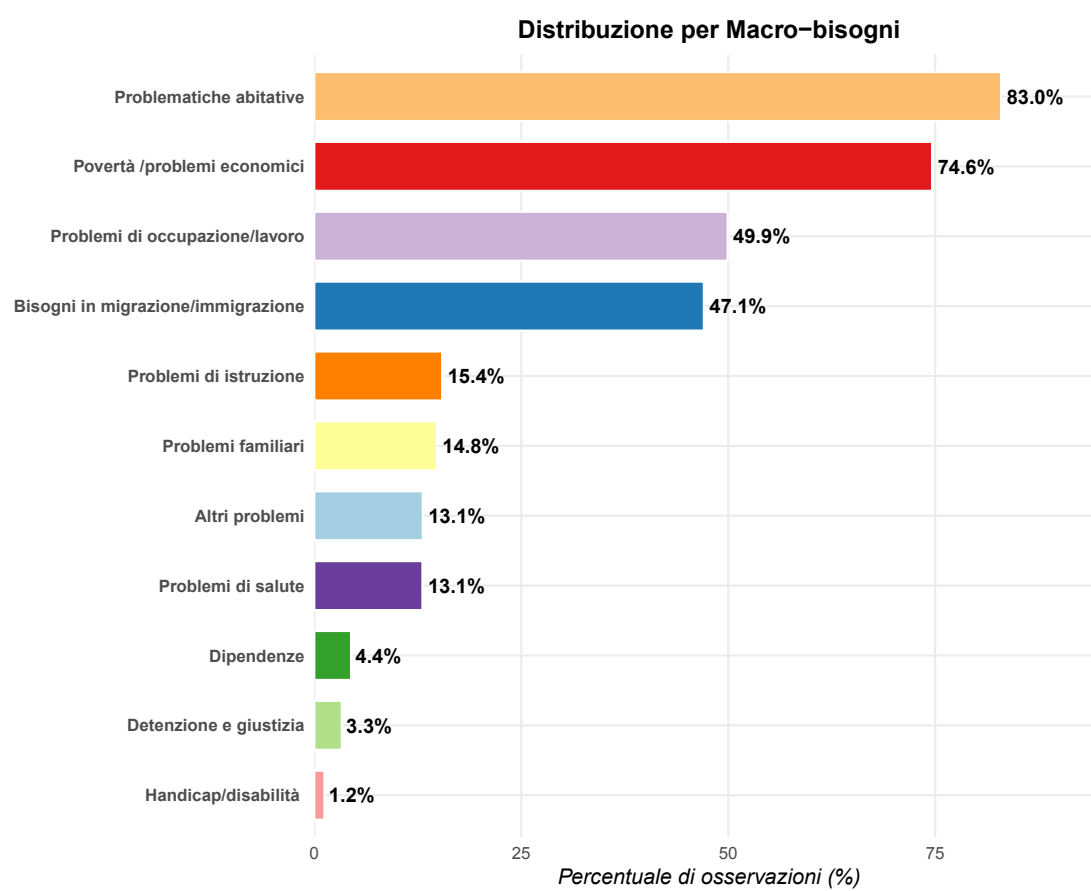


Figura 2.6: Distribuzione dei macro bisogni per utente.

principali servizi si indirizzano verso i servizi di prima accoglienza e i contributi per il pagamento dei canoni di locazione e delle utenze, accompagnati da soluzioni di accoglienza residenziale.

L'area dell'assistenza legale si occupa di fornire consulenza legale e supporto nelle pratiche amministrative, oltre che di contributi per la gestione delle pratiche amministrative. Nel settore sanitario, gli interventi maggiormente rilevanti sono quelli legati al pagamento delle spese mediche, la distribuzione di farmaci, l'accesso a visite specialistiche e la realizzazione di esami.

Le attività di orientamento riguardano soprattutto l'orientamento generale degli utenti, l'accompagnamento ai percorsi lavorativi o previdenziali e il supporto nell'interfaccia con gli uffici amministrativi. Infine, la formazione, l'assistenza socio-assistenziale e l'inserimento lavorativo sono ambiti a minor rilevanza. Si occupano principalmente di erogare corsi di lingua, aiuti per il materiale scolastico, trasporto accompagnato e l'attivazione di tirocini o borse lavoro.

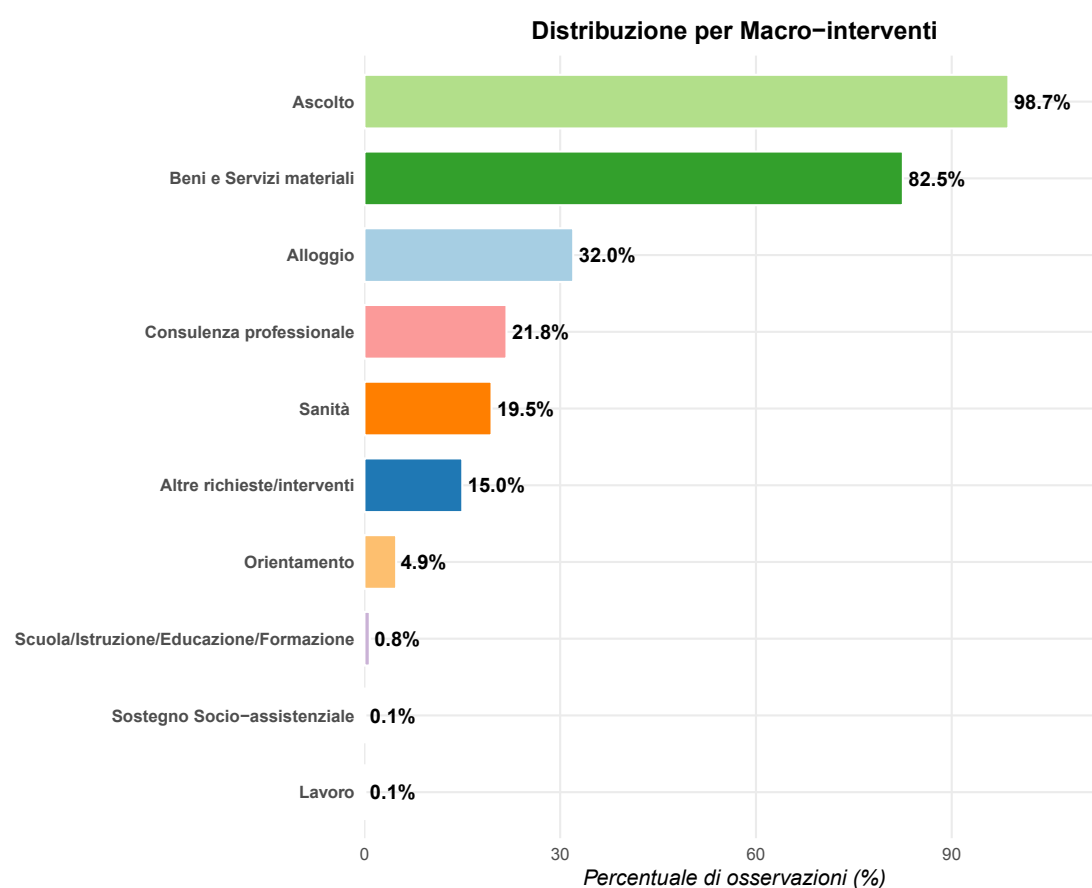


Figura 2.7: Distribuzione dei macro interventi per utente.

Capitolo 3

Modellazione Tematica tramite Latent Dirichlet Allocation

L'analisi dei racconti contenuti nei fascicoli degli assistiti viene effettuata con l'ausilio del modello *Latent Dirichlet Allocation (LDA)*, una tecnica probabilistica di *text mining* non supervisionato molto popolare, che permette di scovare le strutture tematiche latenti all'interno di estese raccolte di testi. Nei prossimi paragrafi si provvederà prima a inquadrare il modello teoricamente; poi si descriverà la sua applicazione alle note del dataset e infine verranno presentati e discussi i risultati ottenuti.

3.1 Fondamenti teorici della LDA

Nei paragrafi seguenti viene illustrata la struttura teorica del primo modello applicato, definendone il posizionamento all'interno dell'ambito del *topic modeling*, il processo di generazione dei testi, le specifiche matematiche e le procedure di inferenza statistica condotte mediante *Collapsed Gibbs Sampling*.

3.1.1 Inquadramento nel panorama del topic modeling

L'estrazione automatica di strutture tematiche da una collezione di documenti è da sempre uno dei principali problemi nel campo del trattamento automatico del linguaggio e dell'apprendimento non supervisionato. Prima dell'avvento dei modelli bayesiani gerarchici, la tecnica più diffusa era il *probabilistic Latent Semantic Indexing (pLSI)*, introdotto da Hofmann nel 1999 [9]. Se da un lato rappresentava un miglioramento rispetto agli approcci di matrice algebrica come la *Latent Semantic Analysis*, dall'altro pLSI soffriva di un problema strutturale non trascurabile: il numero di parametri cresceva linearmente con la dimensione del testo, portando a fenomeni di sovradattamento e rendendo il modello incapace di assegnare stime di

probabilità corrette a documenti non presenti nella fase di addestramento. Infatti, in pLSI la distribuzione dei topic veniva considerata un parametro fisso associato a ciascun documento, e non una vera e propria variabile latente governata da una distribuzione a priori, mancando così un meccanismo di generalizzazione su nuovi dati.

Per affrontare questa limitazione, Blei, Ng e Jordan proposero nel 2003 la *Latent Dirichlet Allocation (LDA)*, un modello bayesiano gerarchico a tre livelli che rivoluzionò il campo del *topic modeling* [1]. L'introduzione di una distribuzione di Dirichlet come a priori sulle proporzioni dei gruppi dei documenti consente di trattare la miscela dei topic come una vera e propria variabile latente, stimata dai dati osservati tramite procedure di inferenza probabilistica più rigorose. Questo modello generativo permette non solo di descrivere la composizione dei testi appartenenti al *corpus* in esame, ma anche di valutare e assegnare probabilità corrette a qualsiasi nuovo documento, risolvendo così il problema di generalizzazione che affliggeva pLSI [11].

3.2 Struttura Probabilistica e Processo Generativo

La *Latent Dirichlet Allocation (LDA)* concepisce ogni documento come una miscela probabilistica di K topic latenti, dove ciascun topic è descritto da una distribuzione di probabilità su un vocabolario di V parole distinte.

Iperparametri e distribuzioni di Dirichlet. Il comportamento probabilistico del modello è guidato da due iperparametri principali: α che governa la distribuzione dei topic all'interno dei documenti. Valori contenuti di α favoriscono composizioni caratterizzate da pochi topic dominanti (rappresentazione sparsa). E β che regola la distribuzione delle parole all'interno di ciascun topic. A partire da tali iperparametri vengono campionate le probabilità mediante la distribuzione di Dirichlet. Per ogni topic $k \in \{1, \dots, K\}$, si estrae la distribuzione delle parole sul vocabolario:

$$\phi_k \sim \text{Dir}(\beta)$$

Per ogni documento $d \in \{1, \dots, M\}$, si estrae la miscela dei topic presenti nel testo:

$$\theta_d \sim \text{Dir}(\alpha)$$

Generazione delle parole. Per ogni posizione n (con $n = 1, \dots, N_d$) del documento d , la singola parola viene generata attraverso una procedura sequenziale in due fasi basata su distribuzioni Multinomiali. In primo luogo viene campionata l'assegnazione tematica $z_{d,n}$ a partire dalla composizione del documento:

$$z_{d,n} \sim \text{Multinomiale}(\theta_d)$$

In secondo luogo viene estratta la parola osservata $w_{d,n}$ attingendo dalla distribuzione sul vocabolario associata al topic $z_{d,n}$ appena selezionato:

$$w_{d,n} \sim \text{Multinomiale}(\phi_{z_{d,n}})$$

Costruzione dei componenti del modello. Tali distribuzioni si integrano in modo gerarchico nel definire le tre entità fondamentali dell'architettura (3.1):

- topic (K): rappresentati dalle distribuzioni ϕ_k , definiscono la struttura semantica complessiva dell'archivio.
- Parole (N_D): le singole unità osservate $w_{d,n}$ derivano dall'interazione tra la scelta del topic $z_{d,n}$ e la relativa distribuzione sul vocabolario $\phi_{z_{d,n}}$.
- Documenti (D): costituiti dalla sequenza di parole generate (w_d), pesate dalla specifica proporzione tematica θ_d .

Probabilità congiunta del corpus. Indicando con W l'insieme delle parole osservate, Z l'insieme delle assegnazioni tematiche, $\Phi = \{\phi_1, \dots, \phi_K\}$ e $\Theta = \{\theta_1, \dots, \theta_M\}$, la probabilità congiunta dell'intero corpus si esprime come:

$$p(W, Z, \Phi, \Theta \mid \alpha, \beta) = \left(\prod_{k=1}^K p(\phi_k \mid \beta) \right) \left(\prod_{d=1}^M p(\theta_d \mid \alpha) \right) \left(\prod_{n=1}^{N_d} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \phi_{z_{d,n}}) \right) \quad (3.1)$$

Tale funzione si articola in tre componenti:

- $\prod_{k=1}^K p(\phi_k \mid \beta)$: individua la struttura globale dei topic all'interno dell'intero corpus.
- $\prod_{d=1}^M p(\theta_d \mid \alpha)$: esprime la propensione specifica di ciascun documento verso i differenti topic.

- $\prod_{n=1}^{N_d} p(z_{d,n} | \theta_d) p(w_{d,n} | z_{d,n}, \phi_{z_{d,n}})$: descrive la catena di campionamento multinomiale condizionata per ogni singolo token del dataset.

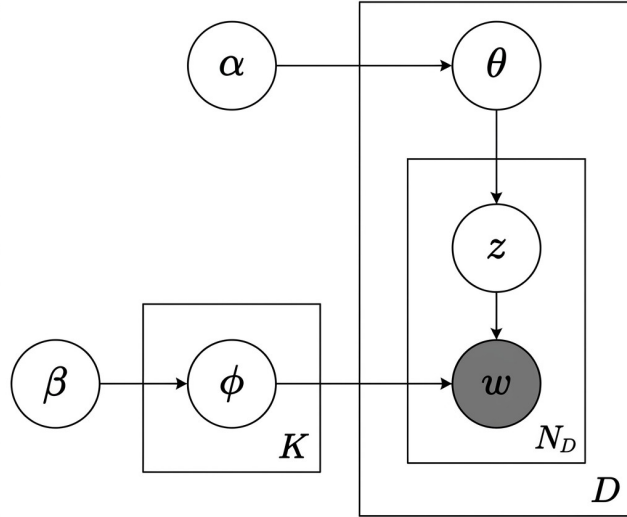


Figura 3.1: Rappresentazione grafica del modello LDA

Inferenza a posteriori L'obiettivo primario della LDA consiste nel determinare la distribuzione a posteriori delle variabili latenti, condizionatamente ai dati osservati e agli iperparametri:

$$p(Z, \Theta, \Phi | W, \alpha, \beta) = \frac{p(W, Z, \Phi, \Theta | \alpha, \beta)}{p(W | \alpha, \beta)} \quad (3.2)$$

Poiché la valutazione del denominatore (ovvero la verosimiglianza marginale del corpus) richiede l'integrazione su tutte le combinazioni di Θ e Φ , unitamente alla sommatoria su ogni possibile configurazione di Z , l'inferenza esatta risulta intrattabile dal punto di vista computazionale. Per la stima della distribuzione a posteriori si ricorre pertanto a tecniche di approssimazione stocastica, quali i metodi MCMC (*Markov Chain Monte Carlo*, in particolare l'algoritmo di *Gibbs Sampling*).

3.2.1 Inferenza statistica tramite Collapsed Gibbs Sampling

La distribuzione a posteriori congiunta esatta di variabili latenti e parametri, ovvero $p(\Theta, \Phi, Z | W, \alpha, \beta)$, non è di fatto calcolabile analiticamente. Il motivo è che le

variabili dei documenti θ_d e le distribuzioni dei topic ϕ_k sono dipendenti all'interno del processo generativo, infatti la verosimiglianza delle parole osservate dipende contemporaneamente sia dal peso del topic nel documento (θ_d) sia dalla probabilità che quel topic generi la parola (ϕ_k). Tale dipendenza incrociata non consente la fattorizzazione della funzione di verosimiglianza e rende il calcolo dell'integrale al denominatore impossibile.

Per affrontare questo problema, in letteratura sono state proposte due principali classi di metodi approssimati: i metodi variazionali che massimizzano un limite inferiore (*evidence lower bound*) della verosimiglianza marginale, e i metodi *Monte Carlo Markov Chain (MCMC)* che definiscono una catena di Markov la cui distribuzione stazionaria è la vera distribuzione a posteriori [1].

In questo lavoro si è scelto di utilizzare il *Collapsed Gibbs Sampling*, un metodo studiato a fondo da Griffiths e Steyvers [7] e descritto dettagliatamente da Heinrich [8]. L'idea alla base del Gibbs sampling, proposta da Geman [6], è quella di campionare a rotazione ciascuna variabile latente dalla sua distribuzione condizionata completa, data la configurazione corrente di tutte le altre variabili. Il termine «collapsed» indica che i parametri continui $\Theta = \{\theta_1, \dots, \theta_M\}$ e $\Phi = \{\phi_1, \dots, \phi_K\}$ vengono integrati prima del campionamento, sfruttando la coniugazione tra le distribuzioni di Dirichlet e Multinomiale. In tal modo si riduce lo spazio di stato del campionamento alle sole assegnazioni tematiche discrete Z , con evidenti vantaggi in termini di efficienza computazionale e velocità di convergenza della catena.

A seguito dell'eliminazione per integrazione delle distribuzioni θ_d e ϕ_k , la probabilità condizionale di assegnare la parola n -esima del documento d (rappresentata dal termine v) al topic k , condizionatamente alle assegnazioni di tutti gli altri token del corpus $z_{-(d,n)}$, è data da:

$$P(z_{d,n} = k \mid z_{-(d,n)}, w, \alpha, \beta) \propto \left(n_{d,k}^{-(d,n)} + \alpha_k \right) \frac{n_{k,v}^{-(d,n)} + \beta}{n_{k,\cdot}^{-(d,n)} + V\beta} \quad (3.3)$$

dove $n_{d,k}^{-(d,n)}$ rappresenta il numero di parole del documento d assegnate al topic k (escluso il token corrente), $n_{k,v}^{-(d,n)}$ indica la frequenza complessiva con cui il termine v è stato associato al topic k nell'intero archivio (escluso il token corrente), e $n_{k,\cdot}^{-(d,n)}$ è la quantità totale di parole assegnate al topic k nel corpus, sempre con l'esclusione del token in esame. Il primo fattore del prodotto quantifica la propensione del documento d verso il topic k , mentre il secondo fattore esprime la probabilità con cui il topic k genera il vocabolo specifico v [7, 8].

L'algoritmo opera in modalità iterativa: durante ciascun ciclo, visita sequenzial-

mente tutti i token dell'archivio e ne ricampiona l'assegnazione tematica sulla base della distribuzione condizionale appena definita. Le iterazioni iniziali, che costituiscono la fase di *burn-in*, consentono alla catena di affrancarsi dall'inizializzazione casuale e di stabilizzarsi nella regione ad alta probabilità a posteriori; tali elementi vengono scartati e non intervengono nelle stime finali. Una volta completata la fase di *burn-in*, le configurazioni estratte possono essere considerate campioni rappresentativi della vera distribuzione a posteriori, consentendo di ricavare le stime dei parametri d'interesse [7].

3.3 Implementazione e Configurazione della LDA

Dopo aver definito il quadro teorico di riferimento e aver terminato le procedure di pulizia del testo, con questo capitolo passiamo all'applicazione pratica delle tecniche di topic modeling al dataset messo a disposizione da Caritas Treviso. L'obiettivo è quello di far emergere in maniera non supervisionata e rigorosamente metodologica i topic trattati nelle note di colloquio, traducendo così il materiale qualitativo in una forma quantitativa che può essere utilizzata per successive analisi statistiche.

3.3.1 Dalle note processate alla matrice documenti-termini

Dopo aver completato la fase di pre-processing descritta nel capitolo precedente si è arrivati ad un corpus di 2.053 documenti costituito dai testi delle note degli assistiti del database Caritas Treviso registrati dopo il 2020. Il passo successivo è stato convertire la componente testuale in una rappresentazione numerica che poteva essere compresa dall'algoritmo appena descritto. Questa conversione è stata fatta utilizzando il pacchetto `tm` per R, realizzato da Feinerer, Hornik e Meyer [5], che fornisce un'infrastruttura di base per l'analisi testuale.

Si parte dalla creazione di un oggetto *Corpus* che rappresenta l'intera collezione di testi in un formato strutturato.

Dal corpus creato così si ricava la *Document-Term Matrix* (DTM) con la funzione omonima. Si tratta di una matrice di dimensione $D \times V$ dove D è il numero di documenti e V la dimensione del vocabolario, i cui elementi (d, v) indicano quante volte il termine v compare nel documento d . La DTM è l'input quantitativo fondamentale per il modello LDA in quanto rappresenta il contenuto qualitativo-

linguistico tradotto in una matrice di frequenze assolute su cui lavora l'algoritmo di campionamento.

3.3.2 Algoritmo di Stima e Gibbs Sampling

La stima del modello LDA è stata effettuata attraverso il pacchetto `topicmodels` per R, descritto e documentato da Grun e Hornik [13], che implementa in modo efficiente sia l'inferenza variazionale sia il Gibbs sampling per i modelli LDA. Nel caso applicativo a questo corpus è stato scelto di utilizzare l'algoritmo di Gibbs sampling.

Nella configurazione del modello, la lista dei parametri è stata dettagliata in ogni suo aspetto. Il parametro riguardante il numero di topic è stato impostato a $K = 6$; la motivazione teorica e metodologica di tale scelta verrà discussa nel prossimo paragrafo. Il numero totale di iterazioni dell'algoritmo di campionamento è stato fissato a 10.000: in ogni iterazione la procedura visita l'intero insieme dei token che compongono il corpus e ne ricampiona l'assegnazione tematica sulla base della regola condizionale di probabilità vista in precedenza, permettendo alla catena di esplorare lo spazio delle configurazioni latenti.

Le prime 1.000 iterazioni sono state considerate di burn-in, periodo durante il quale la catena si libera dall'inizializzazione casuale per avvicinarsi alla distribuzione stazionaria e i cui campioni vengono scartati e non contribuiscono alla stima finale dei parametri. Infine è stato fissato un valore del seme di generazione casuale pari a 1234 al fine di garantire la piena riproducibilità scientifica dei risultati: impostare lo stesso valore permette infatti di replicare esattamente la stessa sequenza di campionamento e le stesse stime finali, requisito fondamentale per il rigore dell'analisi empirica [7].

3.3.3 Ottimizzazione del Numero di topic (K)

Stabilire quanti topic K estrarre da un insieme di testi è uno degli snodi più delicati dell'intero percorso. In letteratura non esiste una regola formale valida in ogni occasione. Sul piano probabilistico, spingere in alto il numero dei topic porta quasi sempre a un adattamento migliore dei dati. Il rischio concreto, però, è quello di spaccare topic dotati di un senso compiuto in sotto-categorie troppo microscopiche e faticose da interpretare. Quindi è stato scelto di affiancare alla lettura critica dei testi alcuni indicatori numerici, in primo luogo la perplessità e la coerenza semantica.

La perplessità indica quanto il modello sia capace di generalizzare, ossia di prevedere l'andamento di nuovi testi mai osservati prima. Per un corpus $\mathcal{D}$, il calcolo parte dall'esponentiale della verosimiglianza logaritmica media per parola:

$$\text{Perplexity}(\mathcal{D}) = \exp \left(-\frac{\sum_{d=1}^M \log p(w_d)}{\sum_{d=1}^M N_d} \right)$$

dove M è il totale dei documenti e N_d indica la lunghezza del testo d [1]. Più il valore della perplessità è basso, più il modello possiede un potere predittivo elevato. Occorre fare attenzione, però: inseguire soltanto questo parametro porta spesso a scegliere soluzioni impeccabili da un punto di vista algebrico, ma confuse o del tutto opache sul piano del significato.

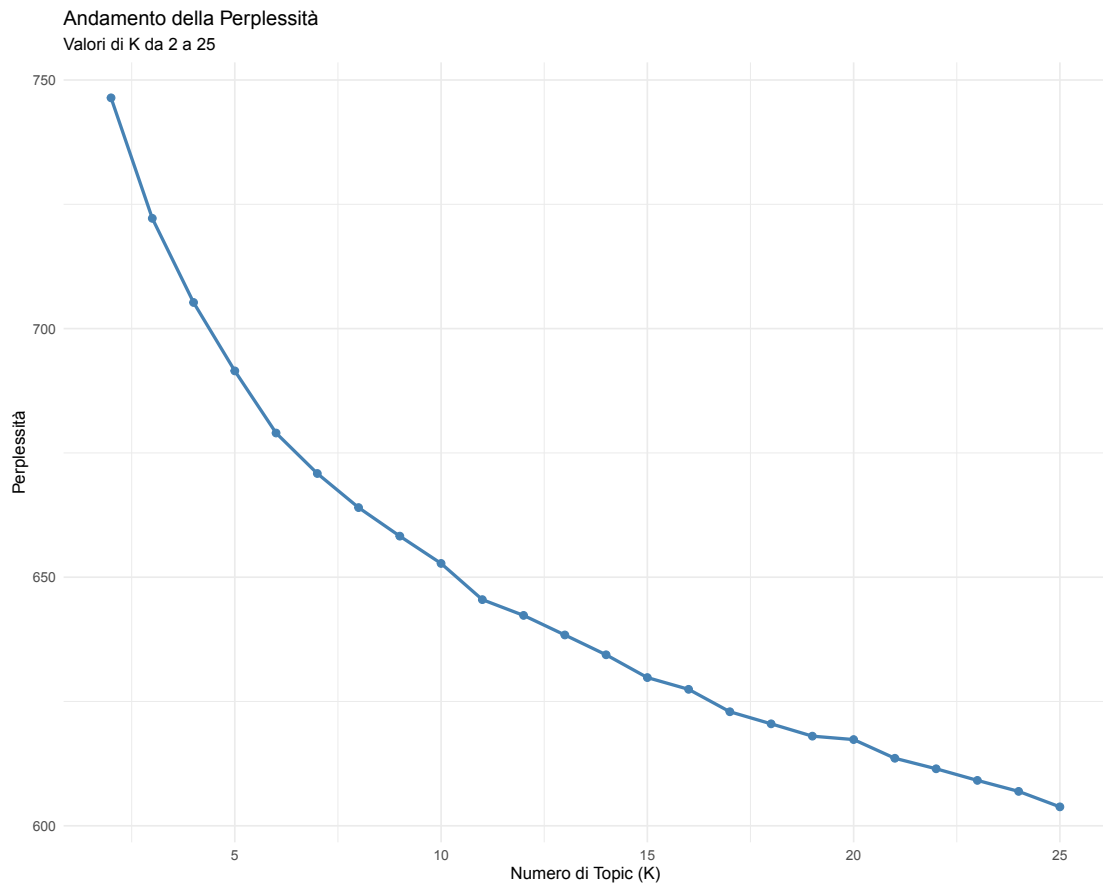


Figura 3.2: Andamento della perplessità in-sample al variare del numero di topic K da 2 a 25. Modelli LDA stimati con metodo Gibbs (1.000 iterazioni).

Per questa ragione si valuta parallelamente la coerenza semantica, un indicatore nato per verificare se i termini più rilevanti di un topic tendano o meno a comparire insieme all'interno dei testi. In questo lavoro si è utilizzata la *Probabilistic Coherence*, calcolata considerando i primi $M = 20$ termini più rappresentativi per ogni topic.

L'indice misura l'associazione tra le parole calcolando il rapporto logaritmico tra le probabilità congiunte e quelle marginali di compresenza nel corpus:

$$C(t, V) = \sum_{m=2}^M \sum_{l=1}^{m-1} \log \left(\frac{P(v_m, v_l) + \epsilon}{P(v_l)} \right)$$

in cui $P(v_m, v_l)$ rappresenta la probabilità empirica che i termini v_m e v_l compaiano nello stesso documento, $P(v_l)$ è la probabilità marginale del termine v_l ed ϵ è una frazione piccolissima usata per evitare logaritmi pari a zero [16]. Questa formulazione restituisce numeri positivi: più il punteggio è alto, più il topic risulta denso e chiaro nel suo significato.

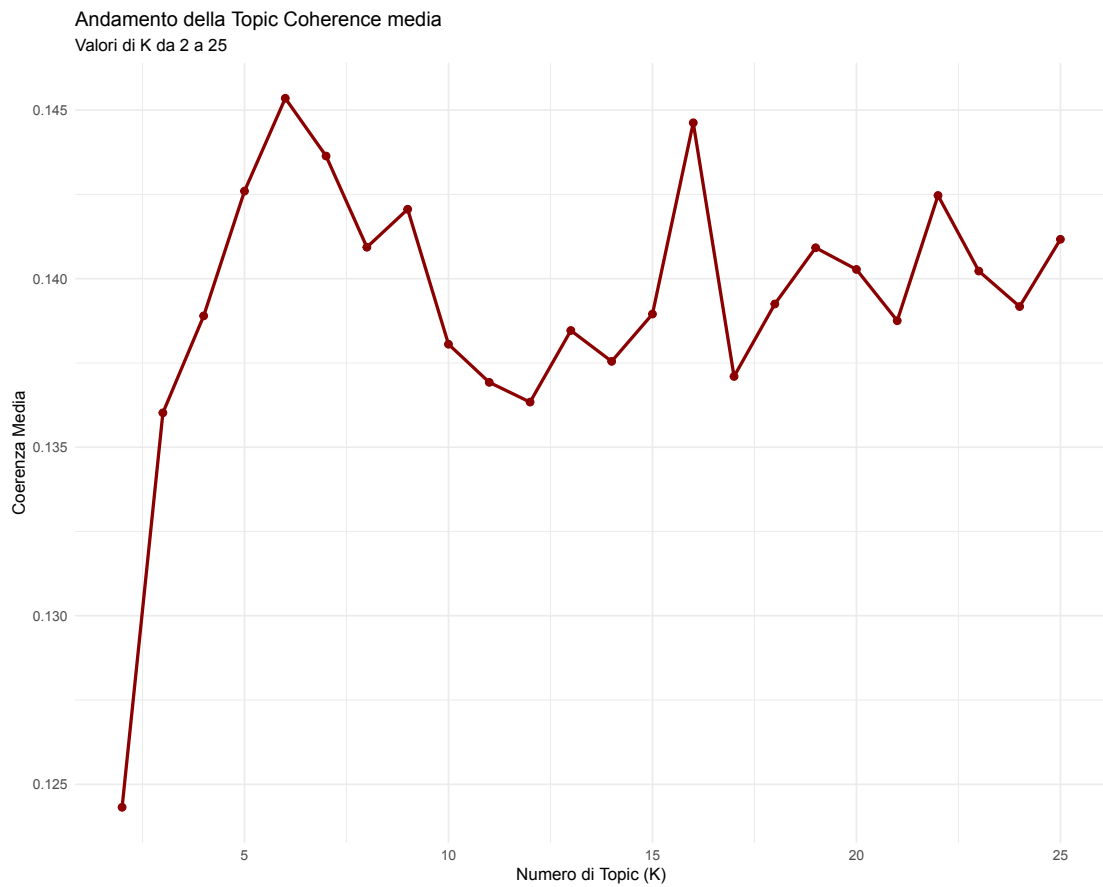


Figura 3.3: Confronto della topic coherence media (Probabilistic Coherence) al variare del numero di topic K . Modelli LDA stimati con metodo Gibbs (1.000 iterazioni).

L'incrocio tra queste due metriche offre un primo orientamento utile per la scelta. Per alleggerire il carico computazionale della fase esplorativa, la stima di perplessità e coerenza è stata condotta impostando un numero contenuto di iterazioni nel campionamento di Gibbs. Guardando all'andamento generale, la

perplexità segue il comportamento atteso e cala in modo progressivo all'aumentare di K . La coerenza media, dal canto suo, sale rapidamente con i primi gruppi per poi iniziare un sostanziale appiattimento subito dopo il sesto topic.

L'evidenza numerica da sola non esaurisce però la valutazione, rendendo necessario un esame diretto dei contenuti attraverso la lettura dei termini più rappresentativi per ciascun raggruppamento. Questa analisi qualitativa ha confermato la bontà della soluzione con $K = 6$, che restituisce topic ben distinti e facilmente leggibili. Riducendo ulteriormente il numero dei topic, diverse problematiche distinte finiscono per fondersi in categorie troppo generiche e poco informative. Al contrario, spingendosi oltre questo valore, si perde progressivamente l'interpretabilità di alcuni gruppi, che tendono a sovrapporsi o a mostrare sfumature poco chiare.

Alla luce di questi riscontri, la configurazione a sei topic è risultata il punto di equilibrio ideale, capace di combinare la tenuta degli indici con la chiarezza necessaria a rappresentare i contesti di supporto e i bisogni censiti dagli operatori della Caritas.

3.3.4 Taratura e Sensibilità degli Iperparametri

Nel modello bayesiano della *Latent Dirichlet Allocation* (LDA), la scelta degli iperparametri α e β non costituisce un dettaglio computazionale marginale, ma riflette l'ipotesi teorica e la struttura probabilistica con cui si intende rappresentare il corpus [1]. In particolare, α governa la distribuzione dei topic nei documenti e influisce sul loro grado di miscelazione. Il parametro β controlla invece la distribuzione delle parole tra i topic, regolando la densità del vocabolario tematico.

Per individuare la configurazione più idonea a descrivere le note di colloquio, si è stimato il modello con $K = 6$ topic, numero ritenuto ottimale alla luce delle successive analisi diagnostiche. La stima è stata condotta mettendo a confronto differenti approcci metodologici, algoritmi di inferenza e strumenti software. Si è valutata anzitutto la variante predefinita su R con il pacchetto `topicmodels`, stimata via campionamento di Gibbs imponendo le distribuzioni a priori simmetriche standard $\alpha = 50/K \approx 8,33$ e $\beta = 0,1$. A questa si affianca una configurazione sparsa, sempre su R, impostata con vincoli più stringenti sui testi e sui termini ($\alpha = 0,1$, $\beta = 0,001$). Un terzo confronto ha riguardato l'algoritmo di Massimizzazione dell'Aspettativa Variazionale (VEM), in cui α viene aggiornato ad ogni iterazione ($1/\alpha \approx 0,1115$), mentre la prior sul vocabolario β viene trattata come una matrice di parametri fissi stimata per Massima Verosimiglianza. Infine, si è testato un modello asimmetrico in Python mediante la libreria `tomotopy`. Quest'ultimo

sfrutta il campionamento di Gibbs ridotto e permette di stimare un vettore α asimmetrico per catturare la diversa prevalenza dei topic, assieme al parametro η per il vocabolario.

La Tabella 3.1 riassume i parametri utilizzati e confronta gli indicatori di valutazione: la Perplexità, che misura la bontà di adattamento sintattica ed esprime una maggiore capacità predittiva per valori bassi, e la Coerenza Media C_v , che valuta la leggibilità semantica dei topic e predilige punteggi elevati.

Tabella 3.1: Confronto delle prestazioni del modello LDA ($K = 6$) al variare delle configurazioni e dei pacchetti.

	Modello 1	Modello 2	Modello 3	Modello 4
Pacchetto	topicmodels	topicmodels	topicmodels	tomotopy
Linguaggio	R	R	R	Python
Metodo di Stima	Gibbs	Gibbs	VEM	Gibbs
Iperparametro α	8,33	0,1	0,11 (stimato)	Asimmetrico
Iperparametro β	0,1	0,001	Stimato (ML)	0,001
Perplexità	678,74	658,12	668,07	1261,88
Coerenza	0,1415	0,1343	0,1249	0,4223

I dati evidenziano come l'impostazione asimmetrica in Python mostri un guadagno netto nella coerenza semantica ($C_v = 0,4223$), pagando tuttavia il prezzo di una perplexità decisamente più alta. Per contro, la variante sparsa su R offre il valore migliore di perplexità (658,12), ma registra una coerenza piuttosto modesta (0,1343). La configurazione standard su R ($\alpha = 8,33$, $\beta = 0,1$) si colloca così nel mezzo, dimostrandosi un solido punto di equilibrio tra capacità di adattamento dei dati e stabilità diagnostica.

La scelta finale dell'impostazione di default non risponde solo a criteri quantitativi. Trova infatti una sponda fondamentale nella natura stessa del fenomeno sociale esaminato.

Impostare un valore di β a (0,1) spinge l'algoritmo a isolare topic, nei quali poche parole chiave concentrano gran parte della massa di probabilità e agevolano l'interpretazione dei contenuti. Sull'altro versante, un α pari a 8,33, nettamente superiore all'unità, formalizza l'idea che i documenti non affrontino un unico topic in modo isolato. Ogni testo viene inteso come un intreccio fluido di più topic.

Questa struttura rispecchia appieno la natura delle note redatte dagli operatori della Caritas di Treviso. La letteratura sociologica sui percorsi di impoverimento insegna che la fragilità è raramente un evento monolitico [3, 15]. Chi si rivolge ai servizi esprime spesso una sovrapposizione di bisogni materiali immediati, come la

mensa o l'alloggio, a cui si sommano problemi di salute o pratiche burocratiche complesse per i permessi di soggiorno. Fissare un valore di α vicino allo zero avrebbe forzato l'algoritmo verso una classificazione rigida, assegnando ogni colloquio a una sola dimensione prevalente e riducendo gli altri disagi a semplice rumore di fondo. Mantenere un α più alto preserva invece la complessità del quadro, lasciando emergere come i diversi elementi di vulnerabilità si combinino nella vita quotidiana delle persone accolte.

3.4 Caratterizzazione e Interpretazione dei topic

L'applicazione del modello LDA alle note di colloquio dell'ente Caritas di Treviso ha fatto emergere sei topic distinti, ognuno caratterizzato da uno specifico nucleo di parole chiave. Prima di esaminarli nel dettaglio, è opportuno sottolineare un aspetto di fondo: in tutte le categorie si nota la presenza costante di termini legati ai bisogni materiali primari. Si tratta di una caratteristica del tutto coerente con il lavoro quotidiano dell'organizzazione, che offre risposte immediate e ad ampia soglia a chiunque si presenti agli sportelli. Al di sotto di questo tratto comune, l'analisi ha saputo comunque cogliere sfumature molto precise, isolando per ciascun gruppo una specifica forma di disagio o un preciso percorso istituzionale.

Il primo topic descrive la vulnerabilità economica delle famiglie e dei soggetti a basso reddito. Termini come *famiglia*, *figlio*, *moglie*, *marito*, *bolletta*, *affitto*, *pagamento*, *appartamento*, *gas*, *luce* e *debito* portano l'attenzione sulla gestione delle uscite quotidiane e sui costi della casa. Nelle note emergono le fragilità finanziarie dovute a spese improvvise o entrate insufficienti, a cui l'ente risponde con contributi ed erogazioni straordinarie per evitare situazioni di morosità.

Il secondo topic riguarda invece la presa in carico e il sostegno relazionale prolungato nel tempo. Parole quali *colloquio*, *ascolto*, *relazione*, *storia*, *percorso*, *rapporto*, *persona* e *assistente* mettono al centro il contatto umano tra l'operatore e l'utente. Qui l'attività si concentra sull'inquadrimento delle storie personali e dei contesti familiari, alternando incontri e telefonate per definire risposte integrate assieme ai servizi sociali del territorio.

Un terzo versante è chiaramente orientato alla tutela della salute e all'accesso alle cure. La presenza di parole come *medico*, *visita*, *farmaco*, *sanitario*, *ospedale*, *ricetta*, *cura*, *esame* e *soccorso* evidenzia le barriere incontrate dalle persone nell'affrontare le spese mediche o le terapie. In molti casi gli interventi riguardano la copertura

dei ticket, la fornitura dei medicinali o il disbrigo delle pratiche per la residenza fittizia, indispensabile per accedere al servizio sanitario.

Il quarto topic riflette i percorsi amministrativi e legali vissuti in prevalenza dalla popolazione straniera. Vocaboli come *documento*, *questura*, *soggiorno*, *avvocato*, *asilo*, *protezione*, *passaporto*, *commissione*, *ricorso* e *traduttore* descrivono il lavoro necessario per districarsi nella burocrazia. In questo ambito la risposta dell'ente si traduce in orientamento specialistico per il rinnovo dei titoli di soggiorno, l'inoltro delle domande di protezione e il superamento dei gap linguistici.

Il quinto topic descrive i progetti dedicati all'accoglienza temporanea e all'emancipazione abitativa o lavorativa. Parole chiave quali *accoglienza*, *stanza*, *ospite*, *progetto*, *monitoraggio*, *alloggio*, *contratto*, *tirocinio* e *inserimento* rimandano alla gestione dei percorsi all'interno delle strutture ricettive. Le annotazioni registrano l'andamento della permanenza degli ospiti, la ricerca di una sistemazione stabile e l'attivazione di tirocini o contratti volti all'autonomia.

Il sesto e ultimo topic raccoglie infine gli interventi d'emergenza e a bassa soglia dedicati a chi vive in strada o in condizioni di grave emarginazione. Termini come *mensa*, *doccia*, *lavanderia*, *vestiario*, *dormitorio*, *scarpa*, *sacco*, *coperto*, *strada* e *stazione* si riferiscono direttamente alla risposta verso i bisogni di sopravvivenza immediata. Le note riassumono qui il supporto quotidiano offerto per l'igiene, il riparo dal freddo, la distribuzione di pasti o vestiti e l'accesso ai dormitori.

Nel loro insieme, queste sei dimensioni restituiscono una rappresentazione accurata del lavoro svolto. Il sostegno economico, l'ascolto relazionale, la tutela della salute, l'orientamento legale per gli stranieri, l'accompagnamento all'autonomia e i servizi d'emergenza costituiscono i pilastri attorno ai quali si articola l'impegno quotidiano della Caritas sul territorio.



Figura 3.4: LDA: topic 1

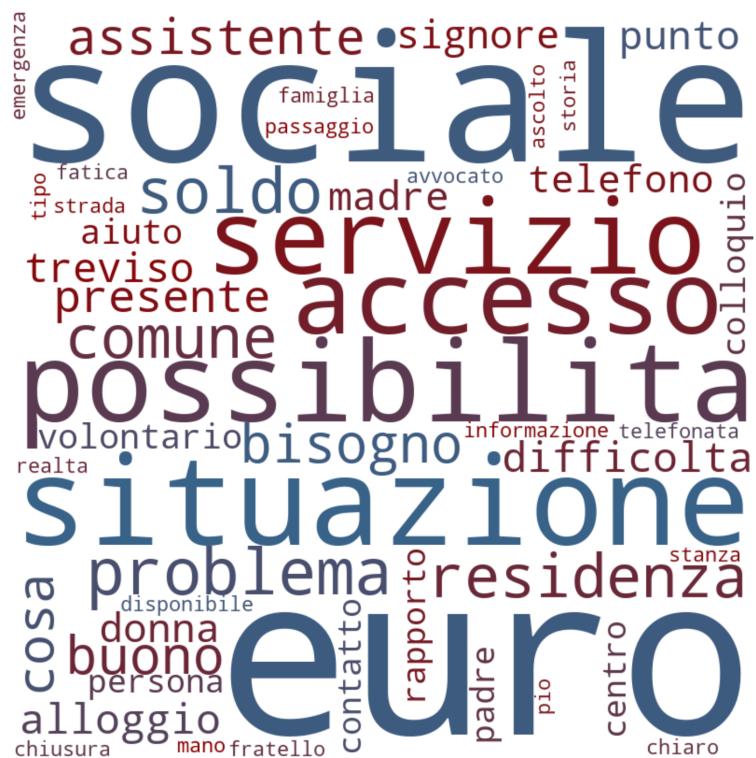


Figura 3.5: LDA: topic 2



Figura 3.6: LDA: topic 3

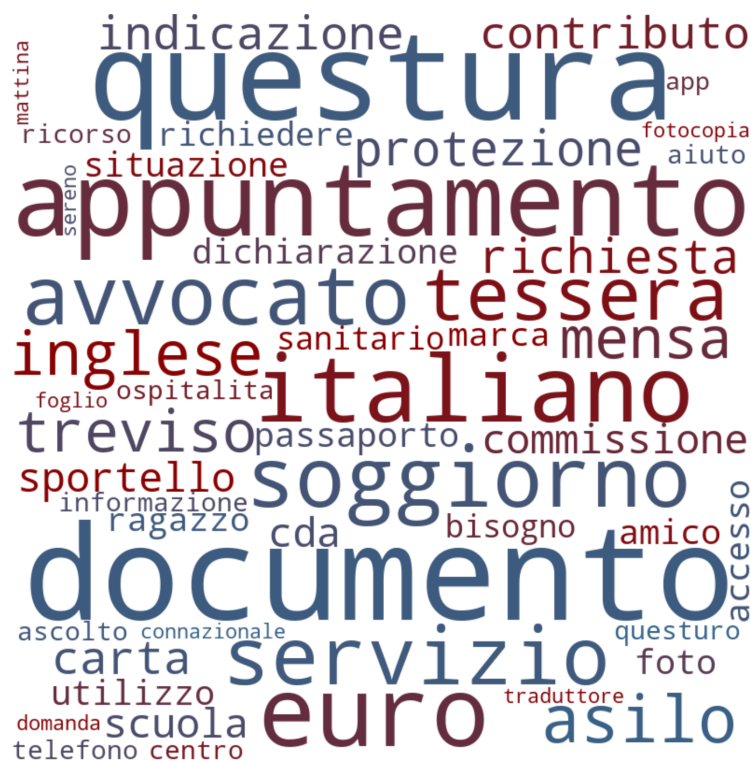


Figura 3.7: LDA: topic 4



Figura 3.8: LDA: topic 5

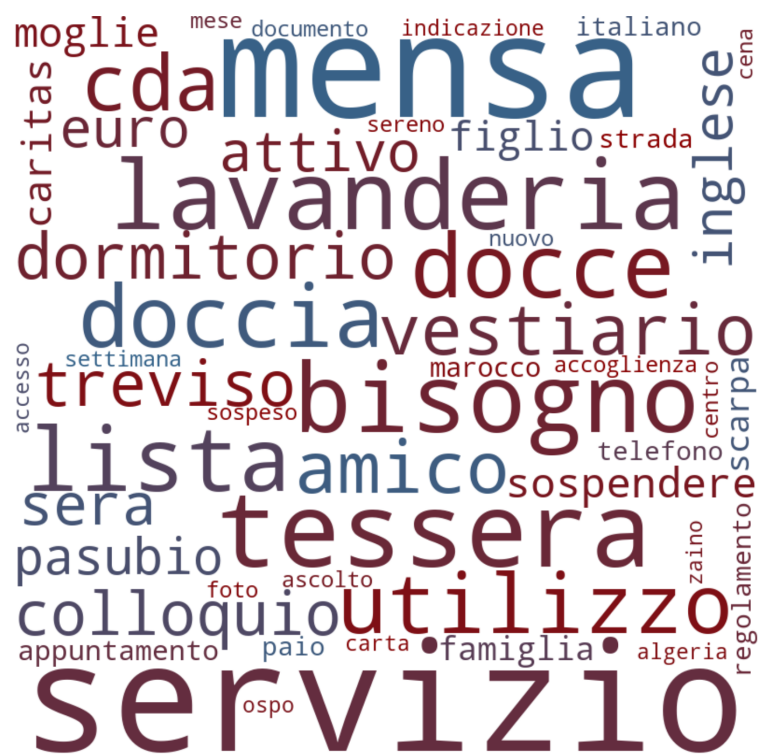


Figura 3.9: LDA: topic 6

Nota: Word cloud comparative dei sei topic estratti dal modello LDA applicato al corpus delle note Caritas Treviso (post-2020). La dimensione di ciascun termine è proporzionale alla sua probabilità stimata $\hat{\phi}_{kv}$ all'interno del rispettivo topic. La variazione cromatica ha funzione puramente estetica per differenziare meglio le parole.

3.4.1 Relazioni Latenti e Strutture di Co-occorrenza

Per approfondire le relazioni tra le tematiche emerse, è stata analizzata la matrice di correlazione di Pearson calcolata sulle colonne della matrice delle proporzioni documento-topic $\Theta_{N \times K}$ (ovvero i valori $\hat{\gamma}_{ik}$ stimati dal modello LDA per ciascun documento i e topic k). Ciascun elemento r_{jk} della matrice misura la covarianza lineare standardizzata tra le quote di presenza del topic j e del topic k all'interno del corpus. L'esame di questi coefficienti consente di identificare la struttura di co-occorrenza del modello: valori positivi indicano tematiche che tendono a presentarsi contestualmente nelle medesime note degli operatori, mentre valori negativi evidenziano dimensioni tematiche tendenzialmente alternative o mutuamente esclusive.

Tabella 3.2: Matrice di correlazione tra i sei topic del modello LDA.

	topic 1	topic 2	topic 3	topic 4	topic 5	topic 6
topic 1	1.000	0.043	-0.209	-0.440	-0.339	-0.422
topic 2	0.043	1.000	-0.159	-0.379	-0.147	-0.354
topic 3	-0.209	-0.159	1.000	-0.082	-0.127	-0.171
topic 4	-0.440	-0.379	-0.082	1.000	-0.049	0.076
topic 5	-0.339	-0.147	-0.127	-0.049	1.000	-0.146
topic 6	-0.422	-0.354	-0.171	0.076	-0.146	1.000

L'esame della matrice di correlazione mostra con chiarezza la struttura delle relazioni tra i domini tematici emersi dal modello. L'unico legame positivo di rilievo intercorre tra il quarto topic, relativo alla regolarizzazione dei cittadini stranieri, e il sesto, incentrato sulla bassa soglia. La convergenza tra questi ambiti evidenzia come la precarietà amministrativa si accompagni in modo sistematico a bisogni di sussistenza immediata quali mensa e riparo. All'opposto, si rileva una netta polarizzazione negativa tra il primo topic, dedicato al sostegno economico ai nuclei familiari, e i cluster legati alla marginalità e alla presenza straniera. Tale contrapposizione riflette la presenza di due bacini di utenza distinti all'interno dei servizi, con i nuclei stabili in difficoltà finanziaria da un lato e la popolazione in

grave marginalità abitativa o giuridica dall'altro. Per contro, le dimensioni della tutela sanitaria e dell'accoglienza abitativa presentano correlazioni pressoché neutre, confermandosi come ambiti di intervento trasversali rispetto ai diversi profili di bisogno.

3.4.2 Validazione dei topic con i Dati di Registro

La valutazione della validità concettuale e operativa dei topic estratti dal modello non supervisionato richiede un confronto sistematico con le informazioni strutturate raccolte dagli operatori in fase di colloquio. Tale verifica si articola su due livelli di analisi complementari. Da un lato, l'esame della penetrazione percentuale delle macrocategorie di bisogno e di intervento all'interno di ciascun gruppo permette di valutare la diffusione generale di tali dimensioni nell'insieme dei dati. Dall'altro, la stima della specificità statistica tramite un approccio inferenziale bayesiano consente di isolare la straordinarietà di ciascun topic. In particolare, vengono considerate caratterizzanti le voci per cui il Rischio Relativo (RR) risulta nettamente superiore a 1, con un intervallo di credibilità la cui soglia inferiore è strettamente maggiore di 1 ($HPD_{Lower} > 1$) [10], come sintetizzato nella Tabella 3.3 e nella Tabella 3.4. L'imposizione di questa soglia assicura una selezione rigorosa dei bisogni e degli interventi, focalizzando l'attenzione solo su quelle specifiche esigenze che mostrano un'associazione positiva e statisticamente robusta con il profilo considerato.

Un primo elemento trasversale che emerge con estrema chiarezza dall'analisi della penetrazione riguarda la presenza quasi universale dell'attività di ascolto e della risposta ai bisogni materiali primari. L'intervento legato all'ascolto caratterizza la quasi totalità dei documenti appartenenti a ciascuno dei sei topic, sfiorando ovunque la saturazione completa della distribuzione. Analogamente, la distribuzione di beni e servizi materiali mostra percentuali di penetrazione estremamente elevate in tutti i gruppi, con picchi particolarmente accentuati nel terzo, nel quinto e nel sesto topic. Sul versante dei bisogni, le problematiche abitative e le condizioni di povertà o difficoltà economica costituiscono lo strato di fondo comune dell'intero insieme di dati, registrando tassi di presenza diffusi che interessano la grande maggioranza delle note in quasi tutti i gruppi. Questa evidenza conferma come l'operatività dell'ente sia strutturalmente impregnata da un modello di accoglienza a bassa soglia, orientato alla ricezione della domanda e al soddisfacimento delle esigenze di sussistenza immediata.

Tuttavia, al di là di questo sfondo omogeneo, l'analisi del Rischio Relativo bayesiano dimostra come ciascun topic latente sia in grado di catturare una specifica

configurazione di bisogno e una risposta operativa nettamente differenziata.

Tabella 3.3: Bisogni specifici e caratterizzanti per ciascun topic (LDA) stimati tramite Rischio Relativo bayesiano ($RR > 1$ e $HPD_{\text{Lower}} > 1$).

topic	Voce di bisogno	Rischio Relativo bayesiano
topic 1	Problemi familiari	1.687 (1.219, 2.282)
	Povertà/problemi economici	1.093 (1.003, 1.183)
topic 2	Dipendenze	2.081 (1.279, 3.316)
	Problemi familiari	1.625 (1.247, 2.097)
	Altri problemi	1.478 (1.097, 1.964)
	Problemi di salute	1.381 (1.016, 1.850)
topic 3	Problemi di salute	2.348 (1.605, 3.308)
topic 4	Problemi di istruzione	1.907 (1.470, 2.447)
	Bisogni in migrazione/immigrazione	1.534 (1.374, 1.704)
	Problematiche abitative	1.083 (1.025, 1.140)
topic 5	Problematiche abitative	1.065 (1.002, 1.127)
topic 6	Bisogni in migrazione/immigrazione	1.242 (1.094, 1.403)
	Problematiche abitative	1.098 (1.044, 1.151)

Tabella 3.4: Interventi specifici e caratterizzanti per ciascun topic (LDA) stimati tramite Rischio Relativo bayesiano ($RR > 1$ e $HPD_{\text{Lower}} > 1$).

topic	Voce di intervento	Rischio Relativo bayesiano
topic 3	Sanità	2.465 (1.878, 3.161)
	Altre richieste/interventi	1.666 (1.099, 2.404)
	Consulenza professionale	1.534 (1.097, 2.067)
	Alloggio	1.356 (1.038, 1.721)
topic 4	Consulenza professionale	1.803 (1.462, 2.205)
topic 5	Alloggio	1.591 (1.338, 1.876)
topic 6	Beni e Servizi materiali	1.100 (1.045, 1.154)

Nel primo topic, l'analisi della penetrazione mostra una prevalenza marcata delle situazioni di povertà economica e di disagio abitativo. La stima bayesiana evidenzia una sovra-rappresentazione specifica per i problemi familiari ($RR = 1,69$) e, in misura minore ma significativa, per le condizioni di povertà e problemi economici. Questo gruppo definisce i nuclei in cui il disagio economico si intreccia in modo caratterizzante con le fragilità e le tensioni delle relazioni familiari.

Il secondo topic isola in modo evidente le forme complesse di devianza e marginalità sociale. A fronte di elevati livelli di penetrazione per i bisogni materiali, il modello bayesiano attribuisce a questo gruppo la maggiore specificità dell'intero

insieme di dati per la voce dipendenze, che registra un Rischio Relativo raddoppiato rispetto alla media ($RR = 2,08$). Inoltre, emergono con significatività i problemi familiari, gli altri problemi e le criticità legate alla salute, delineando una presa in carico multilivello per utenti con marcate fragilità personali e sociali.

Nel terzo topic si osserva una convergenza eccezionale tra la dimensione testuale latente e le rilevazioni degli operatori in ambito sanitario. L'analisi della penetrazione mostra che oltre due terzi delle note registrano un intervento di natura sanitaria, dato confermato dalle stime bayesiane. Il gruppo registra infatti sia una forte specificità per i problemi di salute tra i bisogni ($RR = 2,35$), sia una marcata saturazione sul versante operativo. Si tratta dell'unico gruppo capace di associare una pluralità di risposte caratterizzanti, tra cui sventa l'intervento in sanità ($RR = 2,47$), affiancato da altre richieste, consulenza professionale e soluzioni di alloggio.

Il quarto topic definisce con precisione l'ambito della popolazione straniera e dell'orientamento. Dal punto di vista dei bisogni, si registra un Rischio Relativo nettamente elevato per i problemi di istruzione ($RR = 1,91$) e per i bisogni in migrazione e immigrazione ($RR = 1,53$). Sul piano operativo, la caratterizzazione è univoca. Spicca in modo esclusivo l'intervento di consulenza professionale ($RR = 1,80$), isolando nettamente le attività di supporto burocratico, formativo e legale.

Il quinto topic è dominato dalla dimensione dell'accoglienza e dell'autonomia abitativa. Sul piano delle prestazioni erogate, questo gruppo mostra la più alta penetrazione dell'intervento d'alloggio, dato confermato dall'analisi bayesiana in cui la voce alloggio registra la principale specificità operativa ($RR = 1,59$). Tra i bisogni sottostanti, la risposta si lega in via esclusiva alle problematiche abitative, delineando la gestione delle strutture ricettive e dei progetti di housing temporaneo.

Nel sesto topic, infine, si concentra la risposta per la grave marginalità e la prima accoglienza d'emergenza. L'analisi della specificità evidenzia una combinazione tra i bisogni in migrazione o immigrazione e le problematiche abitative. A questa configurazione risponde un orientamento operativo fortemente sbilanciato sul soddisfacimento immediato dei bisogni primari, confermato dal Rischio Relativo caratterizzante per l'erogazione diretta di beni e servizi materiali.

L'approccio basato sul testo non supervisionato non soltanto rispecchia con fedeltà le categorie amministrative predefinite, ma riesce a ricomporle in configurazioni tematiche complesse che descrivono con maggiore accuratezza la realtà multidimensionale dei percorsi di assistenza e di intervento sul territorio.

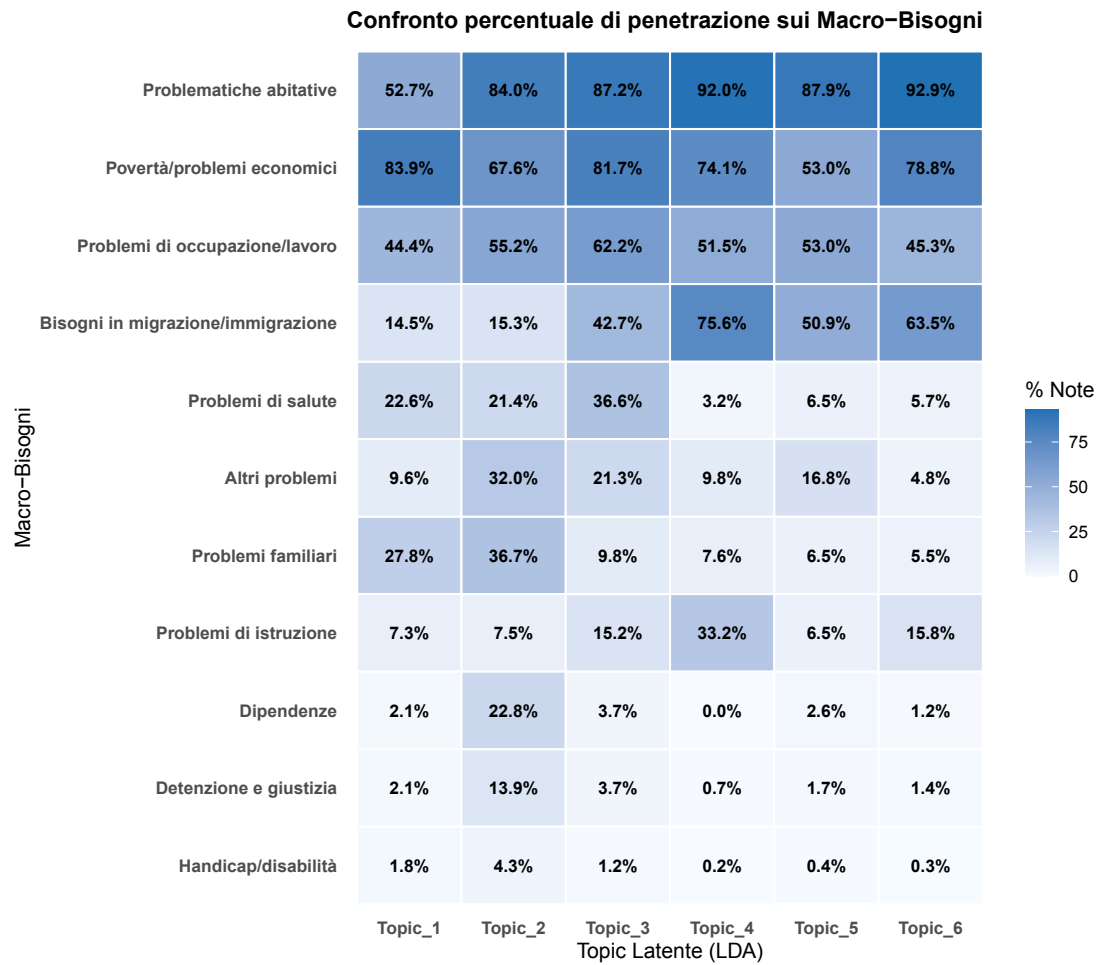


Figura 3.10: Heatmap dei tassi di penetrazione percentuale dei macro bisogni.

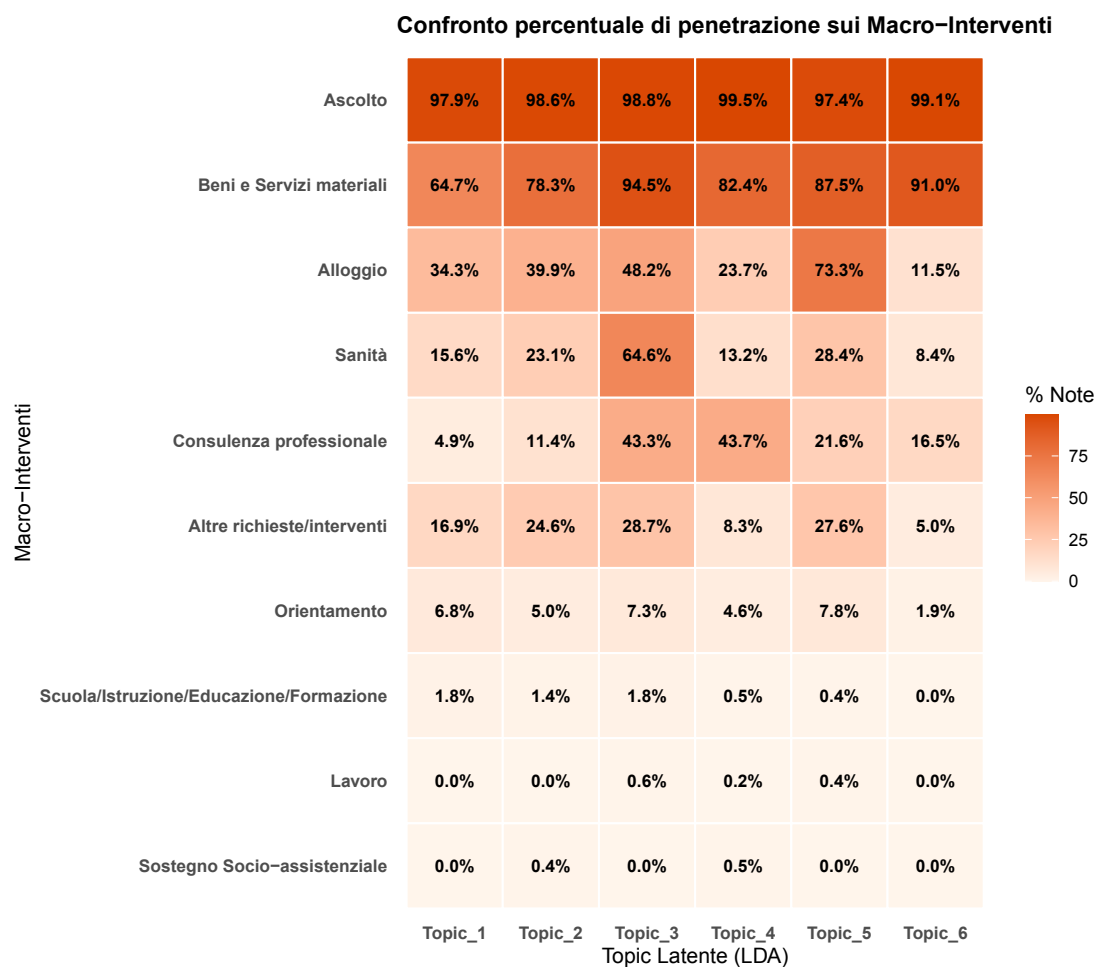


Figura 3.11: Heatmap dei tassi di penetrazione percentuale dei macro interventi.

Capitolo 4

Modellazione Non-Parametrica: Hierarchical Dirichlet Process

Questo capitolo approfondisce l'analisi delle note di colloquio di Caritas Treviso applicando l'*Hierarchical Dirichlet Process (HDP)* per superare i limiti parametrici della LDA e lasciare che la struttura latente emerga dai dati in modo completamente non supervisionato. Dopo un approfondimento dei principi teorici del modello gerarchico e del processo di *Chinese Restaurant Franchise*, vengono discusse le alternative implementative, l'algoritmo di campionamento di Gibbs e la valutazione diagnostica di convergenza degli iperparametri. Infine, nella parte applicativa si discute dei nove topic attivi rinvenuti confermando la diversità delle due classi principali e delle sette sottoclassi tramite l'analisi di correlazione e la validazione incrociata con i dati amministrativi.

4.1 Dal Modello Parametrico al Non-Parametrico: L'Hierarchical Dirichlet Process

Nonostante la solidità teorica della Latent Dirichlet Allocation (LDA), l'utilizzo di tale modello su testi reali impone un vincolo piuttosto forte: dover indicare *a priori* il numero K di topic latenti. Per ovviare a questo problema, è stato proposto l'Hierarchical Dirichlet Process (HDP) [17] [18], un'estensione non parametrica bayesiana che permette al numero di topic di emergere ed adattarsi ai dati in modo flessibile.

Iperparametri e Processi di Dirichlet. A differenza della distribuzione di Dirichlet classica, che associa una misura di probabilità ad un simpleso di dimensione

finita $K - 1$ (limitando quindi in modo rigido il numero di classi), il Processo di Dirichlet (DP) è un processo stocastico che associa una distribuzione di probabilità su misure di probabilità discrete potenzialmente infinite.

Nell'HDP, la condivisione dei topic tra tutti i documenti dell'archivio è modellata da una gerarchia a due livelli di Processi di Dirichlet. Un livello globale (G_0) che definisce la misura di probabilità globale del corpus, modellata come $G_0 \sim \text{DP}(\gamma, H)$ dove H è la misura di base continua (solitamente una distribuzione di Dirichlet sul vocabolario V) che funge da distribuzione “a priori” da cui vengono estratti gli atomi tematici ϕ_k (ossia le distribuzioni delle parole per ciascun topic) e γ è il parametro di concentrazione globale, che governa la varianza complessiva del corpus e controlla la velocità con cui a livello globale vengono creati nuovi topic. Valori di γ più elevati favoriscono la comparsa di un maggior numero di topic.

Poiché G_0 è campionata da un DP, essa è per definizione una misura discreta, che può essere espressa come una somma (con infinità di termini) di misure di massa:

$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k} \quad (4.1)$$

dove gli atomi $\phi_k \sim H$ rappresentano i topic, e i pesi aleatori $\beta = (\beta_1, \beta_2, \dots)$ definiscono la distribuzione globale di frequenza di ciascun topic tramite una costruzione *stick-breaking*, ovvero un procedimento ricorsivo che ottiene i pesi β_k spezzando iterativamente uno stick di lunghezza unitaria in parti di lunghezza $V_k \sim \text{Beta}(1, \gamma)$, in modo che la somma di infinite parti sia pari a 1. Il secondo è un livello locale (G_d): per permettere a più documenti di condividere gli stessi topic senza rinunciare alla continuità di H , la misura discreta G_0 appena ottenuta viene utilizzata essa stessa come misura di base per i Processi di Dirichlet locali relativi ai singoli documenti:

$$G_d \mid G_0 \sim \text{DP}(\alpha_0, G_0), \quad \forall d \in \{1, \dots, M\}$$

In questo secondo livello G_0 rappresenta una misura di base discreta condivisa, che assicura che i Processi di Dirichlet locali G_d scelgano i loro atomi solo all'interno dell'insieme di topic già individuati a livello globale in G_0 . α_0 è il parametro di concentrazione locale, che controlla quanto la miscela tematica del singolo documento G_d possa discostarsi dalla distribuzione media globale G_0 . Un valore di α_0 piccolo costringe il documento a concentrare la propria massa su pochi topic prevalentemente presenti (*sparsity*), mentre un valore elevato rende la costituzione del documento più equilibrata.

Generazione delle parole. Per ogni parola in posizione n (con $n = 1, \dots, N_d$) all'interno del documento d , la singola parola viene generata secondo una procedura sequenziale condizionata dalla misura locale. Si campiona la variabile latente tematica $\theta_{d,n}$ prelevando da G_d , la distribuzione discreta locale del documento:

$$\theta_{d,n} \mid G_d \sim G_d$$

Per ogni parola in posizione n (con $n = 1, \dots, N_d$) all'interno del documento d , la singola parola $w_{d,n}$ viene generata condizionatamente alla misura locale G_d . Equivalentemente la variabile latente $\theta_{d,n} \sim G_d$ sceglie uno degli atomi ϕ_k , e associando al token l'indice del topic $z_{d,n} \sim \pi_d$, da cui viene estratta la parola osservata:

$$w_{d,n} \mid z_{d,n}, \{\phi_k\} \sim \text{Multinomiale}(\phi_{z_{d,n}})$$

Interpretazione intuitiva: Il Chinese Restaurant Franchise (CRF) Per comprendere in modo intuitivo il meccanismo di condivisione e crescita dinamica dei topic, l'HDP viene formalizzato tramite la metafora del *Chinese Restaurant Franchise* [17]. In questa rappresentazione, l'intero corpus è visualizzato come una catena di ristoranti in franchising dove ogni documento d corrisponde a un singolo ristorante e le parole $w_{d,n}$ rappresentano i clienti che vi entrano. All'interno di ciascun ristorante, i clienti si accomodano a diversi tavoli che fungono da cluster locali intermedi. Quando un cliente entra nel locale, decide di sedersi a un tavolo già occupato con una probabilità proporzionale al numero di persone che vi sono già sedute, ereditando automaticamente il piatto già servito a quel tavolo, secondo un principio di auto-consolidamento (*“rich-get-richer”*). In alternativa, con una probabilità proporzionale all'iperparametro di concentrazione locale α_0 , il cliente può inaugurare un nuovo tavolo vuoto. È a questo punto che interviene la condivisione globale della franchigia: nel momento in cui viene aperto un nuovo tavolo, il cameriere deve ordinare un piatto dal menu condiviso dell'intera catena G_0 . Il tavolo si sintonizzerà su un piatto esistente (un topic ϕ_k) con una probabilità proporzionale a quanti tavoli in tutti i ristoranti della catena stanno già servendo quel medesimo piatto. Tuttavia, con una probabilità proporzionale all'iperparametro globale γ , il tavolo può ordinare un piatto completamente inedito dal catalogo generale H , che viene così inserito nel menu comune e reso disponibile per il resto del franchise.

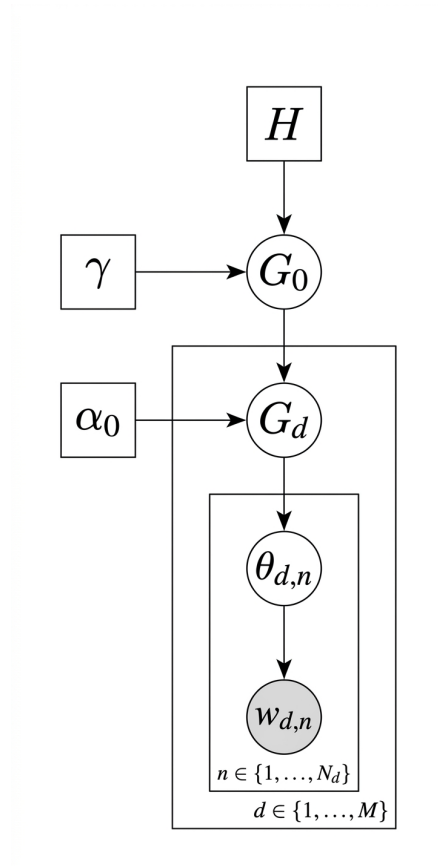


Figura 4.1: Rappresentazione grafica del modello Hierarchical Dirichlet Process (HDP)

4.1.1 Inferenza: Collapsed Gibbs Sampling

L'inferenza esatta sull'HDP risulta analiticamente impossibile. La stima della distribuzione a posteriori viene quindi approssimata tramite tecniche MCMC, nello specifico con l'algoritmo di Collapsed Gibbs Sampling [18].

Sfruttando la coniugazione delle distribuzioni, l'approccio *collapsed* integra analiticamente le variabili continue, come le miscele tematiche dei documenti e del corpus, riducendo lo spazio di campionamento alle sole assegnazioni discrete dei topic per ciascuna parola. A differenza della LDA, dove il numero K di topic è fisso, l'HDP gestisce una dimensionalità dinamica basata sulla metafora del *Chinese Restaurant Franchise (CRF)*.

L'assegnazione della parola n -esima del documento d al topic $z_{d,n}$ avviene condizionatamente a tutte le altre assegnazioni $z_{-(d,n)}$ valutando due opzioni. Da un lato, la parola può essere associata a un topic k già presente nel documento, con una probabilità proporzionale al numero di termini già collegati a quel topic e alla frequenza del vocabolo nel topic. Dall'altro, con una probabilità legata al parametro di concentrazione α_0 , la parola può aprire un nuovo topic locale; questo a sua volta può collegarsi a un topic già attivo nel corpus oppure crearne uno del tutto nuovo, regolato dal parametro globale γ . L'algoritmo ricampiona iterativamente le assegnazioni di ogni token e aggiorna i relativi conteggi empirici, consentendo alla catena di convergere sul numero effettivo di topic attivi presenti nei dati.

4.2 Implementazione e Configurazione dell'HDP

Dopo aver delineato il quadro teorico dell'Hierarchical Dirichlet Process (HDP), la presente sezione illustra l'architettura applicativa e le scelte metodologiche adottate per l'addestramento del modello sul corpus testuale in esame. L'intero processo di modellazione è stato sviluppato in ambiente Python, avvalendosi della libreria `tomotopy` [4].

4.2.1 Iperparametri Iniziali e Setup del Modello

Diversamente da un modello Latent Dirichlet Allocation (LDA) in cui il numero di topic K deve essere scelto a priori, un modello HDP considera K come una variabile latente la cui cardinalità viene stimata a partire dai dati. La gerarchia del processo impone però la scelta di alcuni iperparametri di concentrazione e la definizione di una politica per la selezione dei vocaboli.

Per la rappresentazione dei termini si è scelto di utilizzare la frequenza assoluta unitaria mantenendo quindi la natura discreta dei conteggi necessaria per il campionamento di Gibbs. L'iperparametro di concentrazione locale a livello di documento, α , è stato posto inizialmente pari a 0.1: tale valore esprime un'ipotesi a priori di elevata sparsità supponendo, cioè, che ciascuna nota sia composta da pochi topic predominanti. L'iperparametro di concentrazione globale a livello di corpus, γ , è stato fissato inizialmente pari a 1.0 in modo da bilanciare la nascita di nuovi cluster nel processo del *Chinese Restaurant Franchise*.

La distribuzione a priori di Dirichlet sul vocabolario V per ciascun topic ($\phi_k \sim \text{Dir}(\eta)$) è stata impostata ponendo $\eta = 0.01$. In tal modo si induce una forte concentrazione della distribuzione multinomiale delle parole su un piccolo insieme di vocaboli, aspetto che favorisce la chiarezza e l'interpretabilità semantica dei topic estratti. Infine, la selezione del vocabolario è stata completata introducendo una soglia di frequenza documentale minima pari a tre presenze in modo da escludere i vocaboli troppo rari e attenuare il rumore statistico; non è stata invece utilizzata alcuna soglia per i vocaboli ad alta frequenza affidandone la gestione alle fasi preliminari di pre-processing. Inoltre, è stato fissato il *seed* del generatore di numeri pseudo-casuali in modo che i risultati siano pienamente riproducibili.

4.2.2 Ottimizzazione degli Iperparametri

Una delle criticità principali nei modelli non parametrici bayesiani risiede nella sensibilità dei risultati rispetto a una scelta rigida o arbitraria degli iperparametri di concentrazione. Fissare i valori di α e γ per tutta la catena rischia infatti di imporre a priori un determinato grado di saturazione del corpus e di distorcere la stima del numero di topic effettivi. Per superare questo limite è stata attivata l'opzione di ottimizzazione automatica degli iperparametri, che prevede che i valori di α e γ vengano stimati empiricamente e aggiornati ogni 10 iterazioni del campionamento di Gibbs.

Questo aggiornamento continuo viene effettuato stimando i parametri che massimizzano la verosimiglianza marginale dei dati. Durante il training l'algoritmo valuta la complessità effettiva dei singoli documenti: se le note tendono a concentrarsi su un numero molto ridotto di topic, il valore di α verrà riadattato verso valori più bassi per riflettere una maggiore sparsità; viceversa se il corpus è più eterogeneo tematicamente α crescerà di conseguenza. Allo stesso tempo la revisione di γ modula l'inclinazione del modello a generare nuovi topic a livello globale, equilibrando il numero di osservazioni che vengono attribuite ai cluster già esistenti

e la probabilità di istanziare una nuova componente latente. In questo modo la crescita di K si stabilizza e viene evitata la generazione di topic duplicati o inutili.

4.2.3 Stima del Modello e Diagnostica di Convergenza

Una volta completato il caricamento dei documenti e rimosse le sequenze vuote, il processo di stima è stato eseguito lanciando una catena di campionamento di Gibbs per diecimila iterazioni. Per sicurezza, i primi mille cicli sono stati considerati come periodo di *burn-in*, come per il modello LDA. I campioni prelevati in questo intervallo iniziale sono quindi stati scartati dalla stima finale, consentendo alla catena di distaccarsi dalle condizioni iniziali casuali e di stabilizzarsi nella regione di massima probabilità dello spazio dei parametri.

L'andamento dell'addestramento è stato controllato periodicamente ogni cento iterazioni tramite quattro statistiche chiave. La log-verosimiglianza per parola ha permesso di valutare la bontà di adattamento complessiva del modello. La sua tendenza a stabilizzarsi ha garantito la convergenza della catena MCMC, attestando di aver raggiunto un risultato valido dal punto di vista statistico. In secondo luogo, il numero di topic attivi, dopo una prima fase di fluttuazioni, ha progressivamente trovato un bilanciamento intorno al numero di topic latenti ideale per catturare la struttura semantica del corpus. Infine, il monitoraggio costante degli iperparametri di concentrazione locale (α) e globale (γ) ha mostrato una traiettoria regolare di aggiustamento verso i valori ottimali, pari rispettivamente a $\alpha = 0,02180$ e $\gamma = 1,14984$. Ciò assicura che il modello possa aderire alla vera complessità dei dati senza forzature introdotte a priori.

4.3 Caratterizzazione e interpretazione dei topic

L'applicazione del modello ha fatto emergere in totale nove topic attivi. La struttura tematica è fortemente sbilanciata. Infatti si possono individuare due soli grandi macro-topic che riescono a contenere quasi tutto il materiale informativo e la totalità dell'attività dell'associazione, ai quali si aggiungono sette piccoli gruppi che, pur registrando frequenze assai più basse, risultano molto specifici e ben circoscritti a precise nicchie di intervento.

Si noti che la numerazione degli topic spazia da 0 a 9, escludendo l'8. Tale comportamento è tipico della stima tramite *Chinese Restaurant Franchise*: il modello crea e chiude continuamente i raggruppamenti. In fase di convergenza, la

componente 8 si è svuotata ed è stata dismessa, lasciando attivi solo i 9 argomenti effettivamente popolati.

4.3.1 Prevalenza dei topic nel corpus

L'ampiezza del topic nei gruppi viene individuata attraverso una tecnica di assegnazione sfumata (*soft assignment*). Più precisamente vengono considerati tutti i documenti nei quali un certo topic raggiunge una probabilità superiore al 5,0%. Nella Tabella 4.1 si può vedere la presenza di ciascun topic nel corpus.

Tabella 4.1: Presenza nei topic con probabilità maggiore o uguale al 5,0%

ID topic	Documenti sopra soglia	Percentuale sul corpus
1	1648	80,27%
0	613	29,86%
9	6	0,29%
7	4	0,19%
2	3	0,15%
5	3	0,15%
6	3	0,15%
3	2	0,10%
4	2	0,10%

I numeri riportati illustrano bene la struttura del corpus e il tipo di interventi registrati nelle annotazioni. La somma delle percentuali è superiore al 100% proprio perché l'approccio probabilistico permette a ogni documento di toccare più topic contemporaneamente e superare la soglia in più di un topic.

Da un lato si vede chiaramente come sono distribuiti i valori dei due soli grandi topic che dominano la scena. Il topic 1 è infatti il topic centrale dell'attività dell'Ente e viene toccato nell'80,27% delle annotazioni. In seconda posizione troviamo il topic 0 che è presente nel 29,86% dei documenti come secondo topic più importante. Dall'altro lato vediamo gli altri sette topic (topic 9, 7, 2, 5, 6, 3 e 4) che compaiono come molto piccoli e con presenze che vanno dallo 0,10% allo 0,29% del totale. Una distribuzione a coda lunga come questa indica che il modello ha ben separato l'attività ordinaria dalle risposte a bisogni molto settoriali o a situazioni particolari che coinvolgono solo una piccola parte degli utenti.

4.3.2 Interpretazione semantica dei nove topic

L'interpretazione dei nove topic viene effettuata a partire dall'esame delle parole chiave con più alta probabilità in ciascun topic, che consente di ricostruire il senso delle attività e delle situazioni descritte.

I due topic maggiori rappresentano la spina dorsale dell'intervento sul territorio. Il topic 1 si riferisce alla fase di primo contatto e alle risposte di emergenza, come suggerisce la presenza di parole quali *servizio, mensa, accoglienza, appuntamento, tessera, colloquio, documento, lavanderia, questura, soggiorno, dormitorio, docce, medico, avvocato* e *strada*. Questo topic riassume l'orientamento iniziale più burocratico, la risposta alla fruizione dei servizi essenziali per l'igiene e il nutrimento e la prima accoglienza delle marginalità. Il topic 0 descrive invece i percorsi di presa in carico sociale continuativa e di supporto alla famiglia, essendo caratterizzato da parole quali *sociale, figlio, situazione, assistente, aiuto, comune, spesa, economico, contributo, bolletta, marito, moglie, difficoltà, famiglia, affitto, alloggio, residenza, madre, padre* e *pensione*. Qui si parla dunque di gestione delle fragilità economiche e abitative della famiglia in stretta collaborazione con i servizi sociali e le istituzioni comunali.

I restanti sette topic minori individuano alcuni ambiti di intervento. La salute si declina in due aspetti. Il topic 2 intercetta i bisogni di salute di tipo specialistico e oculistico con parole quali *occhio, visita, medico, farmaco, indicazione, sanitario, oculistico, farmacia, trapianto, retina* e *oculista*. Il topic 6 si riferisce invece all'ambito clinico e all'oncologia grave, con parole quali *ritiro, terapia, tumore, clinico, cartella, ospedale, radioterapia, oncologico, équipe, invalidità, infermiere, tiroide, chemio, benigno* e *oncologo*.

L'ambito della tutela legale e della protezione internazionale è tratteggiato dal topic 3, che si focalizza sulle tematiche amministrative e giudiziarie più complesse attraverso parole quali *avvocato, ricorso, musulmano, carta, paura, procedura, diniego, tunisino, allontanamento* e *questore*.

In tema di sostegno economico e abitativo mirato, il topic 4 racconta di interventi di aiuto finanziario diretto e di acquisto di beni essenziali o farmaceutici attraverso parole come *euro, signora, assistente, acquisto, economico, bolletta, spesa, farmaco, condominiale, ricetta, parrocchia, invalidità* e *medicina*. Il topic 7 si incentra invece in modo molto specifico sul pagamento delle utenze e sui debiti della casa con parole quali *bolletta, aiuto, economico, soldo, gas, pagamento, assicurazione, mutuo, ISEE, canone* e *trasloco*.

Infine, il sostegno materiale e relazionale trova spazio nel topic 5, che si collega

ai contatti e alle comunicazioni per i casi di disagio economico o familiare mediante parole quali *euro*, *moglie*, *comune*, *aiuto*, *economico*, *telefonata*, *soldo*, *giuridico*, *alimentare*, *biglietto* e *disperato*. Il topic 9 si focalizza invece sulla distribuzione di abiti e aiuti per la famiglia e l'infanzia intorno a parole quali *tessera*, *viveri*, *figlio*, *vestiario*, *moglie*, *ritiro*, *rinnovo*, *distribuzione*, *bimbo*, *ricorso* e *gruppo*.



Figura 4.2: HDP: topic 0



Figura 4.3: HDP: topic 1



Figura 4.4: HDP: topic 2



Figura 4.9: HDP: topic 7

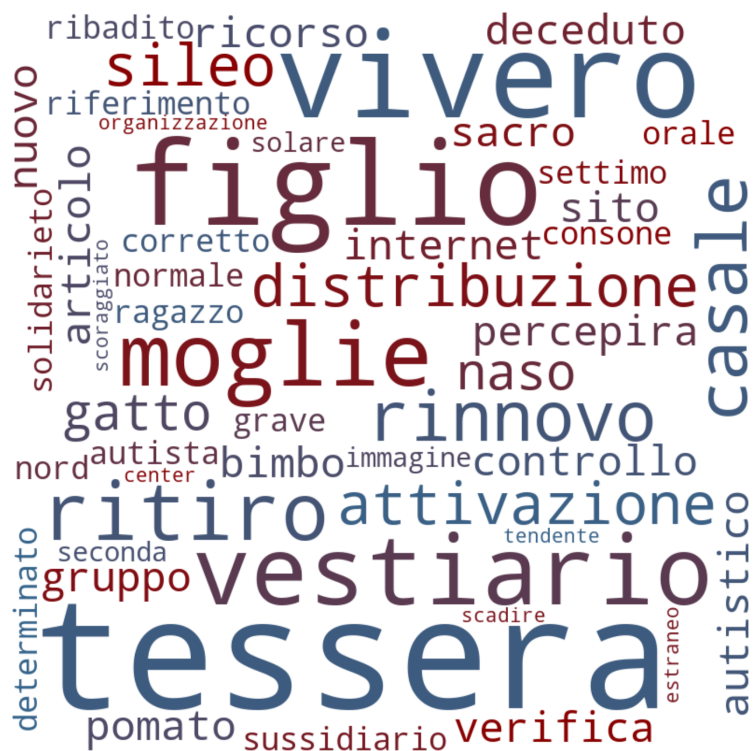


Figura 4.10: HDP: topic 9

Nota: Wordcloud comparativa dei nove topic estratti dall'HDP applicata al corpus delle note Caritas Treviso (post-2020). La dimensione di ciascuna parola è proporzionale alla sua probabilità stimata $\hat{\phi}_{kv}$ all'interno del rispettivo topic. La variazione cromatica ha funzione puramente estetica per differenziare meglio le parole.

4.3.3 Analisi di correlazione tra i topic

Come accaduto per il modello LDA, anche l'analisi della matrice di correlazione di Pearson aiuta a capire la struttura di dipendenza e le eventuali relazioni lineari che intercorrono tra i diversi topic latenti individuati dal processo di stima.

Tabella 4.2: Matrice di correlazione tra le distribuzioni dei topic attivi nell'HDP.

topic	topic 0	topic 1	topic 2	topic 3	topic 4	topic 5	topic 6	topic 7	topic 9
topic 0	1,000	-0,999	0,001	-0,019	0,034	0,047	0,030	0,059	0,060
topic 1	-0,999	1,000	-0,017	0,007	-0,052	-0,063	-0,042	-0,078	-0,097
topic 2	0,001	-0,017	1,000	-0,001	-0,001	-0,001	-0,001	-0,001	-0,002
topic 3	-0,019	0,007	-0,001	1,000	-0,001	-0,001	-0,001	-0,001	-0,002
topic 4	0,034	-0,052	-0,001	-0,001	1,000	-0,001	-0,001	-0,001	-0,001
topic 5	0,047	-0,063	-0,001	-0,001	-0,001	1,000	-0,001	-0,001	-0,002
topic 6	0,030	-0,042	-0,001	-0,001	-0,001	-0,001	1,000	-0,002	-0,002
topic 7	0,059	-0,078	-0,001	-0,001	-0,001	-0,001	-0,002	1,000	-0,002
topic 9	0,060	-0,097	-0,002	-0,002	-0,001	-0,002	-0,002	-0,002	1,000

Dai coefficienti emergono alcune considerazioni interessanti. Il fatto più significativo ai fini dell'interpretazione è sicuramente la quasi perfetta correlazione negativa che intercorre tra il topic 0 e il topic 1 ($-0,999$). Un valore del genere è indicativo della presenza di un vero e proprio asse di bipolarità nei dati. I testi si distribuiscono infatti in maniera quasi perfettamente dicotomica, dove ad una forte presenza del topic 0 fa seguito quasi sempre un'assenza, anche se non totale, del topic 1 e viceversa. Dal punto di vista del contenuto testuale, la fortissima contrapposizione tra i due topic porta a pensare che gli stessi siano due macro-topic distinti o due modi alternativi di parlare e operare nei colloqui.

A parte la forte interazione tra i primi due topic, la matrice mostra valori di correlazione prossimi allo zero per tutte le altre coppie. Un risultato che conferma come il modello non parametrico HDP sia riuscito a identificare due topic verticali e ben separati in grado di non sovrapporre i loro significati. I restanti topic rappresentano dei topic specifici trasversali o delle piccole nicchie presenti in modo indipendente rispetto al resto dei testi.

4.3.4 Validazione dei topic con i dati di registro

L'analisi comparata delle mappe di penetrazione e della stima del Rischio Relativo bayesiano (RR) evidenzia una totale concordanza metodologica e sostantiva, confermando la robustezza della partizione vettoriale del modello HDP.

Come avvenuto per il modello LDA, si è adottato un approccio bayesiano basato sull'intera matrice delle probabilità Θ . La presenza di ciascun bisogno o intervento è stata pesata direttamente tramite le masse di probabilità $\theta_{i,k}$, modellando la prevalenza a posteriori e stimando i relativi intervalli di credibilità al 95% (HPD).

A garanzia di una stima solida, sono stati applicati due vincoli metodologici fondamentali. In primo luogo si è inserito un filtro sulla massa minima del gruppo ($n_t = \sum \theta \geq 10$), in modo da garantire che un topic possedesse una dimensione campionaria pesata sufficiente prima di valutarne la specificità, escludendo così i sette microgruppi. In secondo luogo si è definita una soglia di specificità bayesiana ($\text{HPD}_{\text{Lower}} > 1$), selezionando esclusivamente le associazioni la cui sovra-rappresentazione fosse accertata con una certezza probabilistica del 95%.

Tabella 4.3: Bisogni specifici e caratterizzanti per ciascun topic stimati tramite Rischio Relativo bayesiano ($RR > 1$ e $\text{HPD}_{\text{Lower}} > 1$).

topic	Voce di bisogno	Rischio Relativo bayesiano
topic 0	Problemi familiari	2.076 (1.639, 2.616)
	Problemi di salute	1.753 (1.331, 2.288)
topic 1	Bisogni in migrazione/immigrazione	1.245 (1.140, 1.359)
	Problematiche abitative	1.092 (1.053, 1.131)

Tabella 4.4: Interventi specifici e caratterizzanti per ciascun topic stimati tramite Rischio Relativo bayesiano ($RR > 1$ e $\text{HPD}_{\text{Lower}} > 1$).

topic	Voce di intervento	Rischio Relativo bayesiano
topic 1	Consulenza professionale	1.244 (1.056, 1.465)
	Beni e Servizi materiali	1.064 (1.024, 1.106)

Si osserva una perfetta sovrapposizione tra la dimensione descrittiva (Figura 4.11, 4.12) e quella inferenziale bayesiana. Nel topic 1, le stime del Rischio Relativo identificano una sovra-rappresentazione per i bisogni in migrazione o immigrazione ($RR = 1,245$) e le problematiche abitative ($RR = 1,092$), affiancati sul versante operativo da consulenza professionale ($RR = 1,244$) e beni e servizi materiali ($RR = 1,064$). Questo quadro rispecchia fedelmente il profilo lessicale del gruppo (ad esempio: servizio, mensa, accoglienza, appuntamento, tessera, colloquio, documento, questura, soggiorno, dormitorio), confermando come il topic 1 intercetti l'area della prima accoglienza e del supporto amministrativo tempestivo.

Nel topic 0, la specificità bayesiana si sposta in modo netto sui problemi familiari ($RR = 2,076$, con un'incidenza più che raddoppiata rispetto alla media dell'insieme di dati) e sui problemi di salute ($RR = 1,753$). Tale convergenza concorda appieno

con il lessico caratterizzante (ad esempio: sociale, figlio, situazione, assistente, aiuto, comune, spesa, economico, contributo, bolletta, difficoltà, famiglia, affitto), convalidando l'interpretazione del topic 0 come il polo dedicato alla stabilizzazione della vulnerabilità relazionale e complessa.

La totale convergenza tra l'approccio probabilistico e la mappa di penetrazione dimostra che la separazione dei topic non è un artefatto analitico, ma riflette la reale dualità operativa della rete dei servizi: da un lato l'aggancio e l'emergenza immediata (topic 1), dall'altro l'accompagnamento e l'inclusione profonda (topic 0).

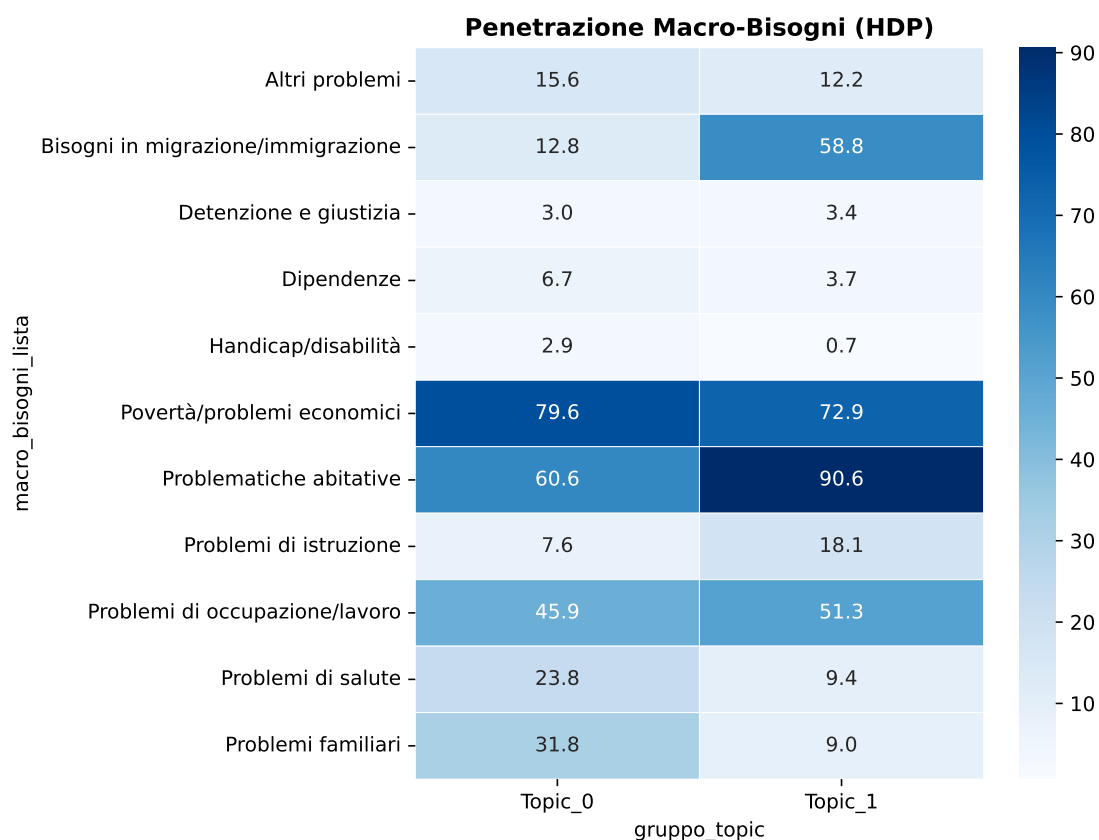


Figura 4.11: Heatmap dei tassi di penetrazione percentuale dei macro bisogni.

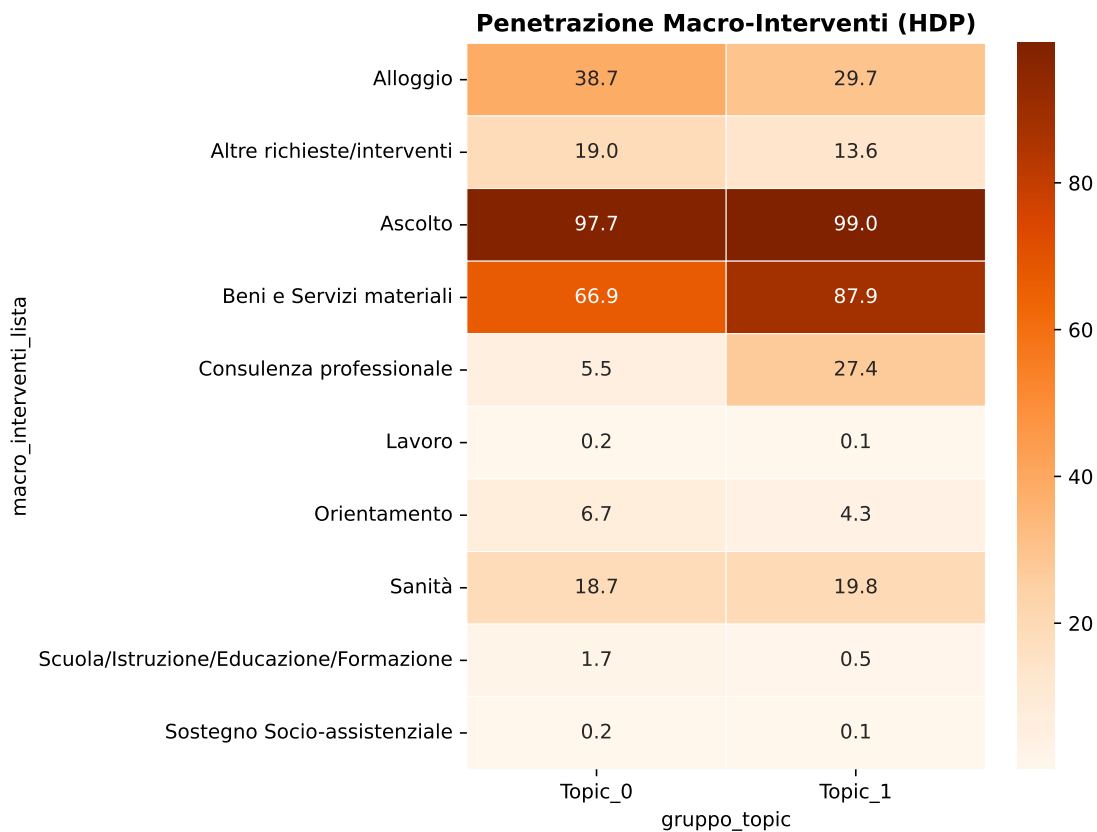


Figura 4.12: Heatmap dei tassi di penetrazione percentuale dei macro interventi.

4.4 Valutazione Comparativa tra LDA e HDP

A conclusione dell'analisi non parametrica condotta tramite *Hierarchical Dirichlet Process* (HDP), si rende opportuno un confronto con i risultati ottenuti dalla modellizzazione parametrica classica sviluppata mediante *Latent Dirichlet Allocation* (LDA). La valutazione comparativa non si limita alla sola dimensione quantitativa degli indici di bontà di adattamento e coerenza semantica, ma si estende alla capacità interpretativa dei vettori di probabilità e alla loro attitudine nel descrivere la struttura informativa contenuta nelle note di colloquio della Caritas di Treviso.

Al fine di offrire una sintesi organica, la Tabella 4.5 riporta il quadro complessivo degli indici calcolati su entrambi i modelli. Alle classiche misure di perplessità e coerenza semantica si affiancano gli indicatori di diversità lessicale e gli indici di entropia relativi alla distribuzione tematica dei singoli documenti.

Tabella 4.5: Sintesi comparativa degli indici prestazionali tra LDA e HDP.

Metrica di valutazione	LDA	HDP
Perplexità	678,7431	892,4988
Coerenza Media	0,1415	0,3816
Diversità dei topic (<i>topic diversity</i>)	0,7333	0,0375
Entropia Media del Documento	2,4006	0,0956
Entropia Normalizzata	0,9287	0,0258
Quota del topic dominante ($\max \theta$)	0,3174	0,9650

4.4.1 Analisi Quantitativa delle Metriche di Performance

Dall'esame dei risultati emerge un comportamento chiaramente divergente tra le due metodologie, a testimonianza della differente struttura teorica sottostante.

In termini di perplexità e coerenza semantica, la *Latent Dirichlet Allocation* registra un valore di perplexità pari a 678,74, inferiore rispetto a 892,50 rilevato per l'HDP, indicando una maggiore fluidità previsionale dell'LDA sul testo non visto. Tuttavia, sul piano della coerenza semantica, l'HDP segna un incremento sostanziale, raggiungendo un indice pari a 0,3816 contro lo 0,1415 dell'LDA, a dimostrazione che i nuclei tematici estratti dal modello non parametrico risultano significativamente più coesi dal punto di vista concettuale.

Per quanto riguarda la diversità dei topic (*topic diversity*), ovvero la misura che esprime la percentuale di parole uniche presenti tra le prime venti parole chiave di tutti i topic estratti, l'LDA presenta un valore pari a 0,7333, a indicare che le parole chiave dei sei topic sono ampiamente distinte. Al contrario, l'HDP mostra una diversità contratta a 0,0375. Tale contrazione deriva dal meccanismo gerarchico del processo di Dirichlet: condividendo una base comune di vocabolario caratteristica del contesto assistenziale Caritas, l'HDP tende a ripartire i termini ad alta frequenza tra più sotto-topic, concentrando la specificità lessicale solo su pochi vocaboli caratterizzanti.

L'analisi dell'entropia, che misura il grado di concentrazione o dispersione dei topic all'interno della singola nota, evidenzia la differenza più profonda nella rappresentazione dei colloqui. L'LDA genera un'entropia normalizzata di 0,9287 e una quota media del topic dominante pari al 31,74% ($\max \theta = 0,3174$), indicando che la probabilità di ogni documento viene ripartita in modo continuo su più gruppi. L'HDP opera invece in maniera quasi discreta, con un'entropia normalizzata che crolla a 0,0258 e una quota del topic dominante che sale al 96,50% ($\max \theta = 0,9650$), assegnando la quasi totalità della massa di probabilità a un singolo topic prevalente per ciascun testo.

4.4.2 Discussione Critica e Sintesi dei Modelli

L'analisi dei due modelli evidenzia un'approfondita differenza di prospettiva in merito a come ciascun approccio fotografi la realtà dei colloqui.

Per il modello *Latent Dirichlet Allocation* il valore aggiunto è costituito dalla capacità di far emergere topic che rappresentano nuclei di bisogno ben delineati (la casa, il supporto amministrativo, la salute o le difficoltà di reddito). Ogni documento viene interpretato come una composizione continua di più topic, aderendo in modo fedele alla naturale concomitanza di fragilità diverse che caratterizza ogni singola storia. Per contro, quest'approccio paga una minore coesione semantica e una distribuzione delle probabilità che tende a renderle piatte, con entropie elevate, per cui in alcuni casi, attribuire un'etichetta sintetica ed univoca ad una singola nota risulta poco agevole.

Per il modello *Hierarchical Dirichlet Process* il punto di forza si sposta sull'ottima coerenza dei topic e sulla capacità di disegnare un profilo chiaro dell'utenza. Guardando al corpus, questo modello individua una vera e propria dualità di fondo che divide la popolazione afferente ai servizi in due grandi cluster di lavoro, restituendo una ripartizione chiara e senza ambiguità per la gestione dei casi. I punti di debolezza stanno invece in una perplessità più elevata, in una bassa diversità lessicale dovuta alla condivisione dei termini di uso comune e nella tendenza a comprimere le sottotracce presenti all'interno di uno stesso colloquio.

4.5 Analisi delle covariate demografiche sui topic HDP

L'analisi svolta mette in luce come il modello HDP sia in grado di tracciare due raggruppamenti ben distinti. Risulta di particolare interesse verificare se le differenze tra tali gruppi riguardino unicamente la tipologia di bisogni emersi durante i colloqui, oppure se esse siano riconducibili a differenti condizioni socio-demografiche di partenza. L'esame delle covariate consente infatti di definire i profili d'utenza specifici che afferiscono a ciascun topic. La Tabella 4.6 descrive la distribuzione delle principali variabili all'interno dei due topic principali.

I dati rivelano l'esistenza di due gruppi di utenti molto polarizzati e profondamente diversi per caratteristiche di base.

Il topic 0 rappresenta utenti con un'età media che si aggira intorno ai cinquant'anni e con una distribuzione per genere praticamente equa. Si tratta principalmente di persone sposate o con alle spalle un matrimonio (separate, divorziate o vedove), che vivono in un'abitazione di proprietà e per la maggior parte provenienti da paesi dell'Unione Europea. Queste caratteristiche rispecchiano fedelmente la natura dei problemi che emergono dalle

Tabella 4.6: Profilazione socio-demografica dell'utenza distinta per topic HDP.

Variabile / Caratteristica	topic 0 ($N = 525$)	topic 1 ($N = 1528$)
Età (anni)		
Media	50,2	37,2
Mediana	50,0	34,0
Genere (%)		
Femminile	50,9%	3,9%
Maschile	49,1%	96,1%
Stato Civile (%)		
Celibe / Nubile	29,8%	62,7%
Coniugato/a	42,9%	28,9%
Separato/a legalmente	9,7%	3,6%
Divorziato/a	8,5%	2,3%
Vedovo/a	4,0%	1,1%
Altro	5,1%	1,4%
Condizione Abitativa (%)		
Casa	42,5%	1,7%
Senza tetto	20,2%	61,9%
Alloggio Inadeguato	19,0%	8,8%
Alloggio Temporaneo	0,8%	1,6%
Altro / Non specificato	17,5%	26,0%
Area Geografica (%)		
Unione Europea	53,3%	12,1%
Asia del Sud	3,4%	41,1%
Africa occidentale	16,2%	24,0%
Nord Africa	11,2%	16,0%
Europa centro-orientale	9,5%	2,1%
Sud America	4,0%	0,9%
Africa orientale	0,6%	1,7%
Africa centrale	0,8%	0,9%
Altre aree (<1%)	1,0%	1,2%

note, dove a maggior associazione statistica troviamo le difficoltà legate alla gestione familiare e al bilancio domestico. Questa corrispondenza trova un riscontro diretto nel lessico a più alta probabilità, che vede la presenza di parole chiave come *figlio*, *spesa* e *bolletta*.

D'altra parte il topic 1 intercetta un profilo molto più giovane, composto per la stragrande maggioranza da uomini celibi e di nazionalità straniera, provenienti per la maggior parte dall'Asia del Sud (soprattutto Pakistan e Bangladesh) e dall'Africa (tra i paesi più rappresentati Marocco, Nigeria e Gambia). Questo gruppo si caratterizza per condizioni di fortissima vulnerabilità e grave marginalità abitativa, risultando per la maggior parte senza dimora o inserito in situazioni abitative precarie. Ed è proprio questo il quadro dei bisogni che rileva il modello, fortemente focalizzati su sportello amministrativo per pratiche di soggiorno, emergenza abitativa e primo ascolto. Specificità di questo insieme che si riflette molto chiaramente nel vocabolario caratterizzante, dominato da parole quali *accoglienza*, *appuntamento* e *questura*.

Capitolo 5

Considerazioni Finali

5.1 Sintesi del Lavoro e Obiettivi Raggiunti

Il presente lavoro di tesi nasce dall'esigenza di valorizzare il patrimonio informativo contenuto nei dati testuali non strutturati raccolti dai Centri di Ascolto della Caritas di Treviso. Nello specifico, l'attenzione si è focalizzata sulle note narrative redatte dagli operatori durante i colloqui con gli utenti. Si tratta di un materiale ricco di sfumature umane e sociali, ma storicamente complesso da analizzare in modo sistematico ed aggregato con i classici strumenti di statistica descrittiva o con le sole informazioni di registro.

L'obiettivo principale della ricerca è stato quindi definire, implementare e valutare una procedura metodologica completa capace di trasformare il testo libero in informazioni strutturate, interpretabili e direttamente utilizzabili per la comprensione delle fragilità dell'utenza e per il supporto ai processi decisionali dell'ente. Per raggiungere tale scopo, il percorso di analisi si è articolato lungo alcune fasi fondamentali.

In primo luogo si è lavorato sulla preparazione e pulizia del testo, progettando una sequenza di trattamento specifica per il linguaggio colloquiale e burocratico delle note Caritas. Attraverso passaggi di normalizzazione, eliminazione degli elementi di disturbo, riduzione alle forme base delle parole e selezione delle componenti grammaticali più rilevanti, si è ottenuto un insieme di dati ottimizzato per i successivi algoritmi di analisi.

Successivamente si è passati alla modellizzazione tematica, applicando e mettendo a confronto due approcci probabilistici per l'estrazione dei topic latenti. Da un lato si è impiegato il modello parametrico Latent Dirichlet Allocation (LDA), valutato tramite un'analisi sistematica delle metriche di coerenza e perplessità al variare del numero di topic K . Dall'altro si è applicato il modello non parametrico Hierarchical Dirichlet Process (HDP), sfruttando una flessibilità bayesiana capace di determinare in autonomia il numero e la struttura dei topic sulla base della natura dei dati.

Infine, si è curata l'integrazione delle informazioni e la profilazione dei bisogni. Le quote tematiche estratte dai testi sono state incrociate con le variabili socio-demografiche e di registro degli assistiti. Questo passaggio ha permesso di stimare i profili e le associazioni

tra le caratteristiche personali (come genere, cittadinanza o stato occupazionale) e specifici raggruppamenti di bisogno (tra cui l'emergenza abitativa, il supporto economico, il disagio lavorativo o la tutela della salute).

I risultati ottenuti dimostrano come l'integrazione tra le tecniche avanzate di trattamento del linguaggio naturale e la modellizzazione statistica bayesiana permetta di superare i limiti delle rilevazioni tradizionali, restituendo una fotografia nitida, multidimensionale e quantitativa delle necessità latenti espresse dal territorio.

5.2 Risultati Chiave e Confronto Metodologico: LDA vs HDP

Il nucleo della tesi ha riguardato la modellizzazione tematica delle note narrative di colloquio, mettendo a confronto l'approccio parametrico *Latent Dirichlet Allocation* (LDA) con l'estensione non parametrica *Hierarchical Dirichlet Process* (HDP) per valutarne l'accuratezza statistica, l'interpretabilità semantica e la flessibilità nell'adattarsi alla complessità del linguaggio naturale. L'applicazione di LDA ha richiesto una ricerca su griglia per individuare il numero ottimale di topic K tramite il monitoraggio delle metriche di coerenza tematica e perplessità; il modello ha mostrato le prestazioni migliori con $K = 6$, sintetizzando efficacemente le macro-categorie di intervento dell'ente, ma rivelando al contempo i limiti della rigidità parametrica, come la tendenza ad accorpare sotto lo stesso topic bisogni differenti e una sensibilità al rumore lessicale che richiede una preparazione del testo molto aggressiva. Al contrario, l'approccio bayesiano non parametrico HDP ha superato il vincolo della scelta a priori del numero di topic, consentendo alla struttura gerarchica basata sul processo di Dirichlet di far emergere autonomamente una configurazione più netta. HDP ha così isolato raggruppamenti specifici legati a principalmente due nicchie di disagio che in LDA venivano assorbite dalle macro-aree, dimostrando un'eccellente capacità di filtrare il rumore lessicale e le difformità stilistiche tra operatori. Integrando le distribuzioni tematiche dei documenti (θ_d) come covariate nei modelli di regressione socio-demografica e valutandole tramite il Rischio Relativo bayesiano, l'approccio HDP ha fornito rappresentazioni più nitide, migliorando la capacità predittiva per la profilazione dell'utenza e facendo emergere chiaramente dinamiche specifiche, come la sovra-rappresentazione delle problematiche abitative tra gli utenti stranieri di recente arrivo a fronte di una prevalenza di richieste di supporto economico e sanitario nella popolazione anziana italiana.

5.3 Implicazioni Pratiche e Impatto Operativo

Al di là del valore metodologico e statistico, i risultati di questo lavoro offrono ricadute applicative dirette sui processi gestionali, organizzativi e decisionali della Caritas Diocesana di Treviso. La trasformazione delle note di colloquio narrative in indicatori quantitativi interpretabili permette di superare i limiti delle sole rilevazioni di registro tradizionali, consentendo all'ente di valorizzare appieno la ricchezza informativa generata quotidianamente dagli operatori nei Centri di Ascolto.

I principali ambiti di impatto operativo individuati riguardano in primo luogo la rilevazione precoce e multidimensionale dei bisogni. I sistemi classici tendono a categorizzare manualmente gli interventi; al contrario, la modellizzazione tematica sviluppata (in particolare tramite l'approccio HDP) permette di portare alla luce la multidimensionalità del disagio, identificando le cause latenti che precedono la richiesta di aiuto economico e offrendo un supporto alla profilazione temporale tra povertà congiunturale e marginalità.

Un secondo ambito fondamentale è rappresentato dall'ottimizzazione della programmazione delle risorse e dei servizi. L'integrazione delle proporzioni tematiche con le variabili socio-demografiche offre alla direzione dell'ente uno strumento basato sull'evidenza per riallocare o potenziare i servizi sul territorio, calibrando l'offerta degli sportelli specialistici in base alla reale incidenza dei topic emersi ed evidenziando l'insorgere di nuove vulnerabilità.

Il modello permette inoltre una sostanziale valorizzazione del lavoro degli operatori. La conoscenza dei termini a maggior valore informativo guida la definizione di linee guida per la compilazione omogenea delle schede, mentre la capacità di sintetizzare storici narrativi complessi facilita il passaggio di consegne tra operatori e volontari nella presa in carico degli utenti.

Infine, l'architettura sviluppata si configura come un modello altamente scalabile e versatile, potenzialmente applicabile in molteplici contesti e organizzazioni. Il suo principale elemento di forza risiede nella capacità di valorizzare ed estrarre conoscenza da dati sensibili non strutturati, un patrimonio informativo di straordinario valore che spesso viene trascurato a causa della sua complessità e dei vincoli di riservatezza.

5.4 Limiti della Ricerca

Nonostante i significativi contributi metodologici e operativi emersi dallo studio, la presente ricerca presenta alcuni limiti intrinseci, riconducibili principalmente alla natura dei dati trattati, alle stringenti esigenze di tutela della riservatezza e alle caratteristiche dei modelli adottati.

Un primo vincolo cruciale riguarda le implicazioni del trattamento dei dati e della privacy in conformità al GDPR (Regolamento UE 2016/679). Le note dei Centri di Ascolto contengono informazioni altamente sensibili sulla salute, sulla situazione economica e sulla sfera familiare degli utenti. I necessari processi di anonimizzazione e rimozione delle informazioni identificative hanno comportato una fisiologica perdita di dettaglio informativo. Inoltre, la stringente necessità di tutelare la privacy ha imposto l'obbligo di operare esclusivamente tramite sistemi informatici locali, senza poter ricorrere a servizi di terze parti. Sebbene questo vincolo abbia rappresentato una significativa sfida tecnica e operativa, ha stimolato la ricerca e la messa a punto di soluzioni ad architettura "leggera". Si è così dimostrato come sia possibile ottenere risultati di elevata accuratezza ed efficacia analitica mantenendo un basso costo computazionale, garantendo al contempo la totale sicurezza dei dati sensibili e la piena sostenibilità tecnologica.

Un secondo elemento di complessità deriva dalla disomogeneità e dalla soggettività della redazione narrativa. Il materiale analizzato raccoglie le annotazioni redatte da decine di operatori e volontari differenti, ciascuno caratterizzato da un proprio stile di scrittura, diverso livello di sintesi e uso di abbreviazioni o gerghi interni. Tale variabilità introduce un rumore linguistico che può ostacolare l'elaborazione automatica. Inoltre, il testo non costituisce una trascrizione letterale delle parole dell'utente, ma rappresenta la mediazione culturale e la rielaborazione effettuata dall'operatore. Di conseguenza, i topic emersi riflettono sia i bisogni reali della persona, sia la lente interpretativa di chi ha compilato la scheda.

Infine, emergono alcuni limiti legati all'architettura dei modelli di *topic modeling* adottati. Basandosi sull'assunzione della "borsa di parole" (*Bag-of-Words*), sia l'LDA che l'HDP non considerano la sequenza sintattica dei termini, perdendo elementi di contesto legati alle negazioni, alle sfumature verbali o al tono emotivo.

5.5 Sviluppi Futuri

Il naturale passo successivo sul piano metodologico riguarda l'evoluzione verso architetture avanzate basate su *Large Language Models*. L'adozione dei modelli di *topic modeling* tradizionali ha fornito una solida base quantitativa, ma l'impiego di tecniche di rappresentazione semantica consentirebbe di superare definitivamente i limiti dell'approccio *Bag-of-Words*, catturando le relazioni sintattiche, il contesto delle frasi e le sottili sfumature di significato tipiche del linguaggio sociosanitario. Parallelamente, l'esplorazione di modelli linguistici di ultima generazione, eseguiti all'interno di ambienti locali protetti e pienamente conformi alle normative sulla riservatezza dei dati, permetterebbe di generare sintesi automatiche degli storici clinico-narrativi e di assistere gli operatori durante la redazione delle schede di colloquio.

Un'altra estensione fondamentale consiste nell'introduzione esplicita della dimensione temporale all'interno del processo di modellizzazione. L'applicazione di varianti dinamiche dei modelli tematici consentirebbe di tracciare l'evoluzione dei bisogni nel tempo, monitorando come l'incidenza di specifiche problematiche vari in risposta a shock socio-economici o mutamenti normativi locali e nazionali. Questo approccio dinamico si rifletterebbe anche nell'analisi longitudinale dei singoli percorsi di assistenza, permettendo di osservare come le tematiche associate all'utente si trasformino dal primo contatto fino alla conclusione della presa in carico, offrendo una misura diretta dell'efficacia dei progetti di reinserimento.

Appendice

1 Prompt completo utilizzato per l'addestramento

Di seguito il prompt completo utilizzato per la pulizia testuale.

Sei un assistente specializzato nel text mining e nella riduzione del vocabolario (text minimization) per dati sensibili in italiano.

Elabora questa porzione di testo seguendo RIGIDAMENTE queste regole:

1. **ANONIMIZZAZIONE:** Rimuovi qualsiasi dato sensibile (nomi, cognomi, numeri di telefono, codici fiscali, indirizzi).
2. **PULIZIA LOG DI SISTEMA:** Elimina completamente stringhe di log automatizzate, modifiche di sistema o pattern ripetitivi come “Modifica valore variabile X. Vecchio valore: Y”. Tieni solo il testo umano.
3. **CORREZIONE:** Correggi eventuali errori grammaticali o refusi prima di ridurre le parole.
4. **NORMALIZZAZIONE E LEMMATIZZAZIONE ESTREMA:** Converti tutto in minuscolo. Riduci ogni parola rimasta alla sua forma base/radice (un mix tra stemming e lemmatizzazione, es. “pazienti” → “paziente”, “operato” → “operare”).
5. **RIMOZIONE STOPWORDS E VERBI VUOTI:** Elimina articoli, preposizioni, congiunzioni, pronomi e verbi ausiliari/ausiliari deboli come “essere” e “avere”.
6. **MINIMIZZAZIONE:** Mantieni solo i termini/concetti chiave essenziali per non perdere il significato informativo della nota.

Rispondi SOLO con le parole chiave risultanti, tutte in minuscolo, separate da un singolo spazio. Non aggiungere commenti o introduzioni.

TESTO DA ELABORARE: blocco

2 Codice e materiali di supporto

Il codice utilizzato per le procedure di *pre-processing* e per l'implementazione dei modelli *Latent Dirichlet Allocation* (LDA) e *Hierarchical Dirichlet Process* (HDP) è disponibile nel repository GitHub del progetto:

`https://github.com/matteobrognera/thesis-topic-modeling`

Per garantire la riservatezza e la tutela dei dati personali degli assistiti, i dati originali forniti dalla Caritas Diocesana di Treviso non sono stati pubblicati nel repository. È stato invece inserito un file di prova, contenente dati dimostrativi, che consente di comprendere la struttura richiesta e di verificare il funzionamento del codice senza esporre informazioni reali o sensibili.

Bibliografia

- [1] David M. Blei, Andrew Y. Ng e Michael I. Jordan. «Latent Dirichlet Allocation». In: *Journal of Machine Learning Research* 3 (2003), pp. 993–1022.
- [2] Caritas Italiana e Fondazione Migrantes. *XXXIV Rapporto Immigrazione 2025: Giovani, testimoni di speranza*. Consultabile sul sito ufficiale di Caritas Italiana. Roma: Tau Editrice, 2025. URL: https://www.caritas.it/wp-content/uploads/sites/2/2025/10/XXXIV-RAPPORTO-IMMIGRAZIONE-CARITAS-ITALIANA-MIGRANTES-_definitivo_compressed.pdf.
- [3] Robert Castel. *Inquadrare la disoccupazione: l'esclusione sociale e i problemi dell'assistenza*. Roma: Carocci, 2004.
- [4] Heewon Cho. *Tomotopy: Python Package of Tomoto (Topic Modeling Tool)*. Ver. 0.14.0. 2026. URL: <https://github.com/bab2min/tomotopy>.
- [5] Ingo Feinerer, Kurt Hornik e David Meyer. «Text Mining Infrastructure in R». In: *Journal of Statistical Software* 25.5 (2008), pp. 1–54.
- [6] Stuart Geman e Donald Geman. «Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images». In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-6.6 (1984), pp. 721–741.
- [7] Thomas L. Griffiths e Mark Steyvers. «Finding scientific topics». In: *Proceedings of the National Academy of Sciences* 101.Suppl. 1 (2004), pp. 5228–5235.
- [8] Gregor Heinrich. *Parameter estimation for text analysis*. Technical Report. University of Leipzig, 2005.
- [9] Thomas Hofmann. «Probabilistic Latent Semantic Indexing». In: *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1999, pp. 50–57.
- [10] John K Kruschke e Torrin M Liddell. «The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective». In: *Psychonomic bulletin & review* 25.1 (2018), pp. 178–206.

- [11] Hang Li. «Latent dirichlet allocation». In: *Machine learning methods*. Springer, 2023, pp. 439–471.
- [12] Joakim Nivre et al. «Universal dependencies v1: A multilingual treebank collection». In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 2016, pp. 1659–1666.
- [13] Martin Ponweiser. «Latent Dirichlet allocation in R». In: (2012).
- [14] Martin F Porter. «An algorithm for suffix stripping». In: *Program* 14.3 (1980), pp. 130–137.
- [15] Jean-François Ravaud e Michel Fardeau. «Disability and social exclusion: the contribution of visual impairment research». In: *International Journal of Rehabilitation Research* 23.1 (2000), pp. 33–45.
- [16] Michael Röder, Andreas Both e Alexander Hinneburg. «Exploring the space of topic coherence measures». In: *Proceedings of the eighth ACM international conference on Web search and data mining*. 2015, pp. 399–408.
- [17] Yee W Teh et al. «Sharing clusters among related groups: Hierarchical Dirichlet processes». In: *Advances in Neural Information Processing Systems*. 2005, pp. 1385–1392.
- [18] Yee Whye Teh et al. «Hierarchical Dirichlet Processes». In: *Journal of the American Statistical Association* 101.476 (2006), pp. 1566–1581.