Politecnico di Torino

Politecnico di Torino

Corso di Laurea Magistrale in Ingegneria Gestionale

Anno accademico 2022/2023

Sessione di Laurea Luglio 2023

Tesi di Laurea Magistrale

Digital VoC analysis: migliorare l'identificazione delle determinanti latenti di qualità di un prodotto con strumenti di topic modeling semi-supervisionato

Relatore:
Prof. Federico Barravecchia
Co-relatore:
Prof. Luca Mastrogiacomo

Candidato:
Mattia Vanin

INDICE

1 INTRODUZIONE

1.1 Premessa

In un mondo in cui i dati sono sempre più fondamentali e generano sempre più valore per le aziende, è importante riuscire ad estrapolare le giuste informazioni da essi.

Internet è un valido archivio di dati da cui attingere, infatti oggigiorno, attraverso il web si riescono ad ottenere moltissime recensioni sui prodotti o servizi che le imprese forniscono al pubblico.

L'utilizzo di moderni approcci che includono algoritmi, permettono di non disperdere queste informazioni e di sfruttarle ottenendo un valore aggiunto.

A questi metodi vengono poi affiancati gli approcci tradizionali, come ad esempio le interviste personali, i focus group, le tecniche qualitative strutturate e le tecniche di analisi del prodotto, i quali riescono ad ottenere il meglio quando utilizzati direttamente con il cliente.

Essi però trascurano le quantità di dati presenti online e non permettono di raggiungere un'elevata popolazione, non garantendo una visione ampia dei bisogni dei clienti e delle determinanti di qualità dei prodotti o servizi che le aziende propongono.

A questo scopo, le aziende che forniscono servizi o prodotti manifatturieri si rivolgono sempre di più a tecniche di data mining che permettono di estrarre una elevata quantità di informazioni dalla Voice of Customer, per intercettare le esigenze dei clienti e capire quali sono le determinanti di qualità del proprio prodotto/servizio, per poter intervenire con un focus più mirato ed evitare di sprecare tempo e risorse dove non è necessario.

In particolare, queste tecniche di elaborazione dei dati impiegano algoritmi di topic modeling che permettono di comprendere il pensiero dei consumatori partendo dalle numerose recensioni che si possono raccogliere in internet (ad esempio social media, forum, aggregatori di recensioni).

Per raggiungere l'obiettivo non esiste un unico approccio metodologico, ma diversi che possono avere una differente efficienza e onerosità di tempo e risorse, quindi le aziende utilizzeranno il metodo più adatto alle proprie esigenze.

Nel presente elaborato verranno prese come riferimento due diverse metodologie di topic modeling:

- non supervisionata: la cui caratteristica principale è l'assenza di parole d'ancoraggio che facilitino l'algoritmo ad individuare determinanti di qualità più coerenti, ed è quella maggiormente utilizzata dalle aziende;
- semi-supervisionata: al contrario della precedente, utilizza delle parole riferite a specifici argomenti che dovrebbero garantire un miglior output.

In riferimento a queste due tipologie di topic modeling, la domanda di riferimento di questo elaborato a cui si dovrà dare una risposta è:

- l'utilizzo di strumenti di topic modeling semi-supervisionato permette di ottenere performance migliori rispetto a quelle ottenibili con strumenti di topic modeling non supervisionato?

La domanda è interessante perché grazie alla continua ricerca nell'ambito della qualità dei prodotti/servizi, si sta tentando di migliorare quanto già si sta facendo nelle aziende: si applicano quindi nuovi algoritmi semi-supervisionati, i quali sono innovativi perché permettono di eliminare le lacune derivanti da una cattiva comprensione del testo da parte del software che potrebbe raccogliere informazioni non inerenti agli argomenti individuati.

In particolare, per raggiungere l'obiettivo della tesi verranno applicate entrambe le tecniche, successivamente verranno analizzati gli output e confrontati i valori ottenuti con diversi indicatori, per stabilire infine qual è il più efficiente e di conseguenza se alle aziende conviene applicare la novità del topic modeling semi-supervisionato al proprio processo di analisi.

Infatti, nel caso i risultati conseguiti siano positivi per questa metodologia, si otterrebbero ricadute positive per le aziende che decidono di applicarlo, traendo un vantaggio competitivo sui competitors in quanto avrebbero informazioni più accurate e di conseguenza saprebbero con maggiore precisione dove intervenire nei casi in cui:

- ci fossero malfunzionamenti o lamentele dei clienti sui prodotti/servizi offerti;
- ci fossero caratteristiche dei propri prodotti o servizi da mantenere perché molto apprezzate dai clienti;
- i competitors offrano prodotti/servizi con caratteristiche migliori o più apprezzate dal pubblico.

1.2 Organizzazione del lavoro

Il presente studio, citato nella premessa, verrà strutturato nel seguente modo:

- Capitolo 2: vengono definiti i concetti chiave utili alla comprensione dell'intero lavoro.

  Infatti, nelle varie sottosezioni vengono riportate le definizioni e le caratteristiche generali delle task svolte per il raggiungimento dell'obiettivo: dal linguaggio di programmazione utilizzato, a cos'è la digital Voice of Customer, i metodi tradizionali e innovativi per la sua analisi, e infine al web scraping;
- Capitolo 3: vengono descritte in maniera dettagliata tutte le attività svolte per giungere all'obiettivo finale, quindi viene spiegato il vero e proprio metodo utilizzato per effettuare lo studio.

  Si procede partendo da cosa è stato applicato per la creazione di un database attraverso il processo di web scraping, passando per i vari task del topic modeling non supervisionato e poi semi-supervisionato, per giungere infine al metodo utilizzato per la validazione e l'analisi dei risultati;

- Capitolo 4: vengono esposti e commentati tutti gli output che si presentano lungo le varie fasi metodologiche dei due casi di studio analizzati.
  Per facilitarne la comprensione vengono utilizzati grafici e tabelle che riassumono i risultati;
- Capitolo 5: vengono esposte le conclusioni che si traggono dal lavoro, le quali comprendono un riassunto degli obiettivi che si volevano raggiungere e dei risultati ottenuti, un'analisi critica di quanto conseguito, delle riflessioni sul processo e le implicazioni che questa ricerca potrebbe avere nella pratica;
- Appendice: contiene una parte più tecnica, principalmente di descrizione dell'algoritmo di topic modeling semi-supervisionato, e si concentra sui comandi che lo compongono.

2 LITERATURE REVIEW

In questo capitolo verranno riportati e spiegati concetti che si presenteranno nel corso della presente tesi, in modo tale da facilitare la comprensione del lavoro.

2.1 Linguaggio di Programmazione R

R è un linguaggio di programmazione nato negli anni ’90 per essere utilizzato nel mondo della statistica, infatti ancora oggi è largamente impiegato in ambiti scientifici e statistici, specialmente quando bisogna analizzare grandi quantità di dati.

Prende spunto da un linguaggio più vecchio di circa 20 anni chiamato S, il quale è commercialmente sottoposto a una licenza.

Il grosso vantaggio di R è proprio il fatto che è stato pensato per essere open-source, permettendogli di diffondersi nel mondo e di avere una enorme comunità che gli consente un continuo sviluppo. Inoltre possiede una sintassi molto flessibile e una natura modulare.

R fornisce un’ampia varietà di utilizzi nel campo della statistica (modelli lineari e non lineari, test statistici, analisi di serie temporali, classificazioni, clustering e altro) e della grafica.

La distribuzione di R possiede 8 pacchetti di partenza che permettono il suo utilizzo più generico.
Essi però possono essere facilmente ampliati attraverso i siti CRAN, che ne contengono sempre di nuovi per coprire un ampio range di statistiche moderne.

Secondo il Tiobe Programming Community Index, ovvero un famoso indicatore che classifica la diffusione dei vari linguaggi di programmazione, a Novembre 2022 R risulta essere al 12° posto.

2.2 Digital Voice of Customer (VoC)

La digital Voice of Customer è un termine usato nel business per descrivere il processo di cattura dei bisogni dei clienti (Griffin A. e Hauser J., 1991).

Nel loro paper proseguono sostenendo che la VoC produce un set dettagliato di cosa i clienti hanno bisogno o vogliono e le loro aspettative.
Queste informazioni sono poi organizzate in strutture gerarchiche in base alla loro importanza e soddisfazione rispetto alle alternative.

La digital Voice of Customer può essere estratta da blogs, da posts sui social media e da recensioni online, ed ha avuto un enorme successo come fonte di informazione perché è gratuita, affidabile e facilmente accessibile a chiunque (Özdağoğlu *et al.*, 2018).
Tutto ciò è reso maggiormente praticabile dal Web 2.0, ovvero dalla seconda fase di sviluppo e di diffusione di internet, il quale è caratterizzato da una sempre più crescente interazione tra il sito web e l'utente.
Per questa ragione gli utenti sono essenzialmente più coinvolti, diventando spesso autori (attraverso blog, chat, forum), condividendo in maniera più efficiente le informazioni attraverso sistemi peer to peer o con sistemi di condivisione multimediale, e con l'affermazione dei social networks (Tirunillai e Tellis, 2014).

La Voice of Customer permette di ottenere una comprensione dettagliata dei bisogni dei consumatori, un linguaggio comune per il team che elabora queste informazioni, input di specifiche per la realizzazione di nuovi prodotti/servizi e anche una spinta per l'innovazione del prodotto o servizio già sul mercato (Griffin A. e Hauser J., 1991).

Le funzioni di un prodotto/servizio che un'azienda decide di garantire e la soddisfazione delle esigenze di un cliente che lo utilizza sono due facce della stessa medaglia (Franceschini F., 2001) (vedi figura 1).

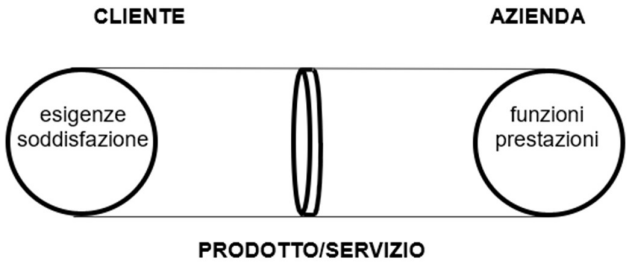


Figura 1 – Le due facce della "medaglia prodotto/servizio": il punto di vista del cliente e dell'azienda (Franceschini, 2001)

Non sempre però i bisogni dei clienti e le effettive funzioni di un prodotto/servizio si incontrano (vedi figura 2), infatti conta molto anche la percezione della qualità di un prodotto o servizio che essi ricevono.

In riferimento a quest'ultima affermazione, un esempio che richiamerà un servizio che verrà analizzato nei prossimi paragrafi dell'elaborato potrebbe essere la valutazione assegnata dagli utenti di Uber alla pulizia del veicolo sulla quale viaggiano: infatti, a parità di pulizia, se arrivasse un vecchio modello di automobile potrebbero avere una percezione di minore igiene e maggior trascuratezza rispetto al caso in cui viaggiassero su un mezzo più moderno, e di conseguenza il voto assegnato al servizio sarebbe inferiore.

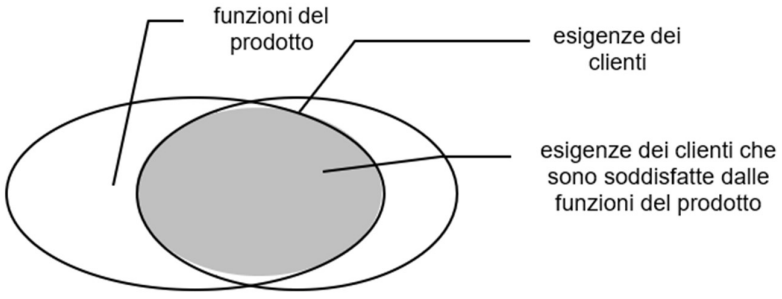


Figura 2 – Funzioni del prodotto ed esigenze dei clienti: solo un sottoinsieme di esigenze viene soddisfatto dalle funzioni (Franceschini, 2001)

La percezione può essere paragonata ad una lente che i clienti utilizzano per filtrare tutti i complicati stimoli che ricevono dal prodotto/servizio, dalle loro

caratteristiche e dal canale comunicativo/persuasivo (in generale dalla pubblicità) (Urban, Hauser, 1993; Franceschini F., 1998).

È proprio grazie alla Voice of Customer e ai più recenti strumenti per analizzarla (topic modeling semi-supervisionato e non, per citare ciò che verrà utilizzato in questa tesi) che le aziende riescono sempre di più ad avvicinare le funzioni del prodotto alle esigenze dei clienti.
Infatti il data mining e le tecniche di machine learning rendono possibile l'analisi di queste informazioni, permettendo di estrapolare quelle più rilevanti ed evitando il compito impossibile per gli umani di lettura e interpretazione dell'enorme quantità di dati estratti.

2.3 Web scraping

Il web scraping è definito come una tecnica per estrarre dati da internet e salvarli in un file system o in un database per una successiva attività di recupero o di analisi (Bo Zhao, 2017).

Questa tecnica è riconosciuta come estremamente efficace e potente per la raccolta di big data (Mooney *et al.* 2015; Bar-Ilan 2001).

I software utilizzati per svolgere l'attività di web scraping simulano l'essere umano che naviga nel browser e raccoglie dati e informazioni nei diversi siti web, ma permettono di farlo velocemente e in maniera automatica (Diouf R. *et al.*, 2019).

Essi si fondano su applicazioni basate sul formato HTML.

In generale, questo processo viene eseguito da un software (conosciuto anche come Web robot) che imita l'interazione che ci sarebbe tra i server web e l'essere umano attraverso un browser (Glez-Peña *et al.*, 2013).

Principalmente l'attività di web scraping procede passo dopo passo attraverso il software che accede ai siti web di cui si ha bisogno, analizza i loro contenuti, estrae i dati d'interesse e li "struttura" come l'utente finale desidera.

In particolare le fasi che vengono svolte in quest'attività sono:

1. La fase di accesso ai siti, che consiste nel software che comunica con la pagina web contenente le informazioni desiderate attraverso il protocollo HTTP, ovvero un protocollo Internet che coordina le transazioni di richiesta-risposta tra client (generalmente un web server e un browser);
2. La fase di HTML parsing e di estrazione dei contenuti, ovvero una volta che i documenti in formato HTML sono recuperati, il software web data scraper estrae i contenuti d'interesse;
3. Il task di costruzione dell'output, il cui principale obiettivo è di trasformare i dati estratti grezzi in informazioni più strutturate con il fine di poterle poi utilizzare per effettuare analisi o semplicemente di immagazzinarle.

Sempre secondo lo studio effettuato da Glez-Peña *et al.* (2013) esistono tre principali categorie web data scraper:

1. Libraries: i software vengono costruiti utilizzando il linguaggio di programmazione più conosciuto dal soggetto che lo realizza. Tendenzialmente, in questo caso, vengono implementate nel codice librerie di terze parti. Questa categoria può avere alcuni inconvenienti. Spesso le librerie hanno bisogno di essere integrate oppure i software realizzati non stanno al passo con i cambiamenti del formato HTML, e quindi richiedono una continua manutenzione;
2. Frameworks: la presente categoria tenta di risolvere i problemi della precedente presentando soluzioni maggiormente integrate nei software oppure presentano un domain-specific languages (DSL), ovvero un linguaggio di programmazione specifico per una particolare situazione;
3. Desktop based environments: è la categoria di software che verrà utilizzata in questa tesi per l'estrazione del database di recensioni dei servizi presi in esame, in quanto è utilizzabile anche da chi non è esperto di programmazione. Infatti è uno strumento che possiede al suo interno un browser integrato attraverso il quale un utente può navigare nel sito web d'interesse e selezionare gli elementi della pagina che dovranno poi essere

estratti. Inoltre, sono software interessanti perché permettono di ottenere output in vari formati, come ad esempio CSV, Excel o XML, che possono poi essere inseriti in un database.

La tecnica del web scraping può essere utilizzata in un'ampia varietà di situazioni, come ad esempio la raccolta di recensioni di prodotti/servizi, il monitoraggio o la comparazione dei prezzi, la raccolta di annunci immobiliari o di dati di monitoraggio del meteo, per tenere traccia dei cambiamenti dei siti web e altro ancora (Bo Zhao, 2017).

I software di web scraping (di tipologia Desktop based environments) che verranno impiegati per l'estrazione dei dati sui quali verranno processati gli algoritmi di topic modeling saranno "Octoparse" e "ParseHub".

2.4 Approcci tradizionali di analisi della Voice of Customer

La caratteristica principale di questa tipologia di approccio è quella di essere più diretta e limitata con il consumatore.

Ne esistono di diverse tipologie, e qui di seguito verranno elencate e descritte le principali:

1. Intervista personale: al cliente viene chiesto direttamente cosa pensa del prodotto/servizio.

   Possono essere somministrati dei semplici questionari da compilare oppure poste a voce le domande ritenute più utili ai fini dell'individuazione del pensiero del cliente.

   Molteplici studi hanno dimostrato che raggiunto un numero di circa 30 intervistati l'efficienza raggiunge il valore massimo, come mostrato in Figura 3;

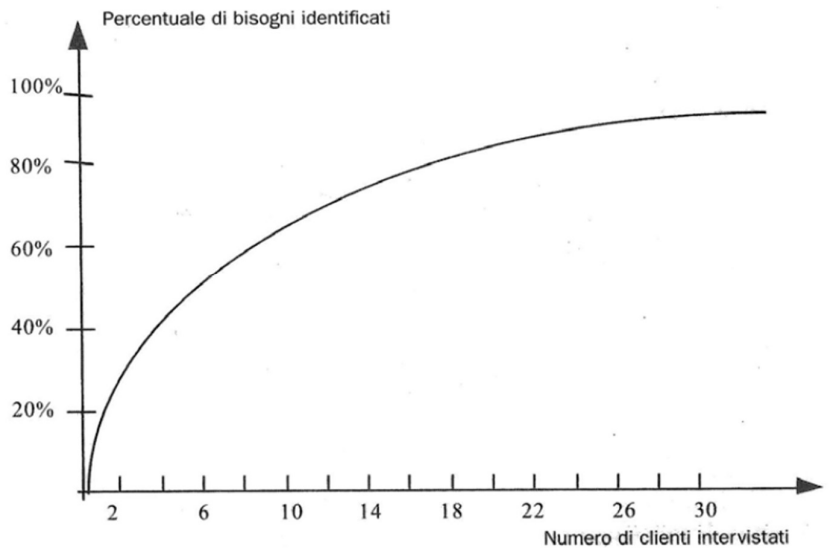


Figura 3 – Grafico rappresentante l'andamento della percentuale di identificazione dei bisogni rispetto al numero dei clienti intervistati

2. Focus group: gruppi di 6/8 clienti a cui viene richiesto di dialogare riguardo un prodotto/servizio, concentrandosi su pregi, difetti, modalità di utilizzo, eventuali modifiche da apportare, in modo tale da poter recepire informazioni dalla discussione;
3. Tecniche qualitative strutturate: ai clienti viene richiesto di classificare e valutare tendenzialmente tre diversi prodotti, con il fine di comprendere gli aspetti positivi e negativi della concorrenza e poter valutare con dei benchmark il prodotto/servizio che viene offerto al pubblico;
4. Tecniche di analisi del prodotto: viene chiesto ai clienti di spiegare i motivi per il quale compra e utilizza un certo prodotto/servizio.

2.5 Approccio innovativo di analisi della Voice of Customer: Topic Modeling

Il topic modeling è una delle più comuni e potenti tecniche di text mining, per la scoperta di dati latenti e per trovare relazioni tra dati e documenti testuali (Jelodar *et al.*, 2019).

È un processo che viene utilizzato moltissimo per elaborare il linguaggio naturale per trovare topic e semantic mining da documenti che non hanno un'origine comune (Jelodar *et al*., 2019).

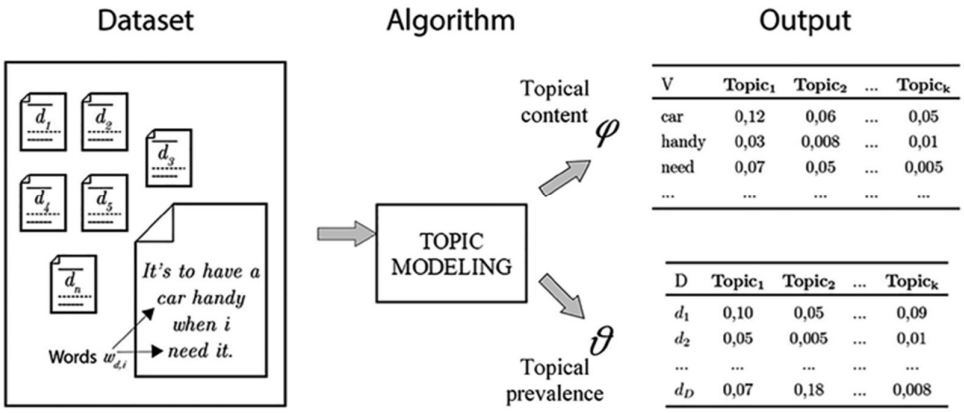
Il metodo può essere applicato in un'ampia varietà di campi molto diversi tra loro, come ad esempio nell'estrazione di recensioni per scovare le determinanti latenti di qualità (argomento della presente tesi), oppure nelle scienze politiche, linguistiche, mediche o nel ramo del software engineering, e molti altri ancora.
In particolare, è molto sfruttato per analizzare la Voice of Customer di un prodotto o servizio e stabilire quindi le rispettive determinanti latenti di qualità, ovvero tutti quegli aspetti di cui discutono e si concentrano i consumatori, per cui è necessario che le aziende comprendano e valutino adeguatamente queste tematiche in modo tale da poter intervenire e migliorare quanto offrono.

La principale idea delle tecniche di topic modeling è che i documenti sono rappresentati da un insieme casuale di topic latenti, i quali sono caratterizzati da una distribuzione di parole.
I vocaboli che compaiono con una più elevata frequenza in ogni topic, tendenzialmente rappresentano l'idea centrale di quest'ultimo.

Questi algoritmi di text mining di solito si svolgono su cinque fasi distinte:

1. La raccolta dei documenti per la creazione di un dataset;
2. La fase di pre-processamento (o in inglese pre-processing) del testo;
3. La preparazione e la selezione dei dati;
4. L'estrazione delle informazioni;
5. La valutazione e l'interpretazione dei risultati.

Nella figura sottostante (figura 4) vengono sintetizzate e rese più chiare le fasi e quindi il funzionamento di un algoritmo di topic modeling, in cui d_i sono i documenti estratti da fonti diverse con il software di web scraping che formano il dataset della Voice of Customer.



V	$Topic_1$	$Topic_2$	...	$Topic_k$
car	0,12	0,06	...	0,05
handy	0,03	0,008	...	0,01
need	0,07	0,05	...	0,005
...	...	...	...	...

D	$Topic_1$	$Topic_2$	...	$Topic_k$
d_1	0,10	0,05	...	0,09
d_2	0,05	0,005	...	0,01
...	...	...	...	...
d_D	0,07	0,18	...	0,008

Figura 4 – Rappresentazione grafica del funzionamento degli algoritmi di topic modeling (Barravecchia et al., 2022)

L'output che si ottiene dall'applicazione dell'algoritmo di topic modeling comprende due insiemi di valori di probabilità:

1. Topical content φ: probabilità che le parole trovate siano usate in un topic;
2. Topical prevalence ϑ: distribuzione multinomiale di probabilità che spiega quanto di un documento è associato ad un topic.

La topical prevalence è utile per effettuare una fase finale del modello, definita di validazione, che permette di capire quanto sono accurati i risultati ottenuti attraverso degli indicatori.

Come descritto da Roberts *et al.* nel 2019, dato un insieme di documenti, i principali problemi degli algoritmi di topic modeling sono:

1. Identificare un set di topics che descriva un corpus testuale (ad esempio la raccolta di documenti testuali da fonti diverse);
2. Associare un set di parole chiave per ciascun topic (φ, topical content);
3. Definire per ogni documento un insieme dei topic trovati (ϑ, topical prevalence) (Blei *et al.*, 2003).

Esistono diversi approcci per implementare l'attività di topic modeling, ma tutti hanno in comune che si basano su algoritmi di machine learning.

Uno di questi è il Latent Dirichlet Allocation (LDA) da cui nascono molte varianti, tra cui lo Structural Topic Models (STM), che verrà utilizzato per effettuare lo studio di questa tesi.
È spesso implementato per elaborare il linguaggio naturale nel text mining, per fare analisi sui social media e per recuperare informazioni.
STM è innovativo perché permette di incorporare dei metadati arbitrari, ovvero informazioni contenute all'interno di ciascun documento, nel topic model (Roberts *et al.*, 2014).
Alcuni esempi di metadati possono essere l'autore del documento, la data, i ratings, la nazione, e molto altro ancora.
I topic models statistici (come quello citato sopra) molto spesso sono riferiti a un metodo non supervisionato, in quanto deducono il contenuto dei topics oggetto di studio invece di comprenderlo (Blei, Ng, and Jordan 2003;Grimmer 2010; Quinn *et al.* 2010;Wang and Blei 2011).

Al contrario, nei modelli supervisionati (come verrà spiegato nel successivo paragrafo), l'analista definisce ex-ante dei topics, tendenzialmente codificando lui stesso a mano un set di documenti e inserendoli in categorie prestabilite (Laver, Benoit, Garry, 2003).

2.6 Topic Modeling semi-supervisionato

Il topic modeling semi-supervisionato sarà l'elemento centrale su cui si svilupperà la tesi, perché sarà attraverso questo strumento che si capirà se vale la pena per le aziende e per i ricercatori effettuare quest'ulteriore step per l'analisi dei dati, tentando di verificare se si avrà un significativo miglioramento delle performance.

Si parte dal presupposto, già in parte citato nel precedente paragrafo, che il Latent Dirichlet Allocation (LDA) ha il problema di essere un metodo di generazione probabilistica non supervisionata (Jelodar *et al*., 2019), cioè in estrema sintesi non

riceve come input dei documenti etichettati in precedenza (non c'è una fase di training) per orientare al meglio l'algoritmo durante la ricerca dei topics.

La tecnica semi-supervisionata invece consiste nel poter fornire al modello dei semi (seed), ovvero delle parole che rappresentano al meglio un determinato topic, che fungono da ancoraggio per trovare vocaboli sempre più inerenti a quest'ultimo, garantendo quindi una sorta di training per l'algoritmo.
Infatti, attraverso un topic modeling non supervisionato, potrebbe accadere che vengano restituiti degli output confusi o non interpretabili che potrebbero fuorviare l'utente.

Un esempio di algoritmo di topic modeling semi-supervisionato è il Seeded-LDA, la cui caratteristica principale è quella di permettere di salvare in una variabile i seed riferiti a ciascun argomento, in modo tale da poterlo elaborare attraverso un comando che restituisce quindi l'output richiesto.

L'implementazione di questo processo verrà effettuato tramite il linguaggio di programmazione R prendendo spunto da algoritmi realizzati da ricercatori che hanno già provato ad utilizzare questo strumento.
L'algoritmo utilizzerà delle librerie sviluppate da quest'ultimi, in modo tale da poter usufruire di comandi specifici che permettono di elaborare le informazioni fornite in input, ovvero le recensioni dei clienti dei prodotti/servizi presi in esame, e restituire gli output richiesti (la topical prevalence) in comune con l'algoritmo non supervisionato, in modo tale da poter confrontare i risultati e stabilire la migliore tra le due tipologie.

In particolare, in questa tesi verrà usato l'algoritmo proposto nello studio "Seeded-LDA for Topic Modeling" il 14 Ottobre 2022, utilizzando il quanteda package e la libreria GibbsLDA++.

3 METODOLOGIA

Nella presente tesi, per perseguire il suo obiettivo, verranno attuate le seguenti fasi per i due casi di studio che verranno analizzati:

1. Raccolta dati dalla Digital Voice of Customer (web scraping);
2. Prima depurazione del database dalle recensioni non utili;
3. Fase di pre-processing;
4. Applicazione topic modeling non supervisionato:
    a. Identificazione numero ottimale di topics;
    b. Applicazione dello Structural Topic Modeling (STM);
    c. Analisi dei topics (labelling);
    d. Estrazione dei parametri del modello;
5. Applicazione topic modeling semi-supervisionato:
    a. Identificazione numero ottimale di topics;
    b. Analisi dei topics (labelling);
    c. Perfezionamento dei topics elaborati;
    d. Definizione dei seed;
    e. Estrazione dei parametri del modello;
6. Validazione dei risultati dei modelli di topic modeling;
7. Analisi e confronto dei risultati.

Qui di seguito (figura 5) vengono schematizzate tutte le task appena identificate:

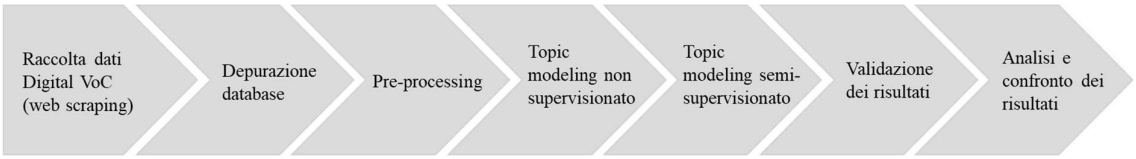


Figura 5 - Fasi metodologiche dello studio

Nei prossimi sottoparagrafi verranno esposte nei dettagli tutte le task della metodologia appena citate.

3.1 Web Scraping

Per effettuare l'estrazione del dataset su cui verranno eseguite le analisi, sono stati presi come riferimento siti web in lingua inglese in quanto i software di web scraping utilizzati elaborano meglio le parole in quella lingua, inoltre permette anche di ottenere recensioni che provengono da tutto il mondo, in modo tale da avere un database più ampio e completo.

I database sono stati ricavati utilizzando prevalentemente il software Octoparse, e in parte anche ParseHub.

Le informazioni che sono state ritenute utili per lo svolgimento dell'elaborato sono la data in cui la recensione è stata inserita, il testo della recensione e il rating.
Per agevolare le successive fasi di analisi sono anche state aggiunti manualmente un codice identificativo ID per ogni recensione e la fonte di provenienza.

3.2 Depurazione del database

Una volta ottenuto il database contenente tutti i metadati necessari per affrontare il lavoro, si procede con una sua prima pulitura, ovvero vengono eliminate tutte quelle recensioni che possiedono anche solo una tra le seguenti caratteristiche:

- Recensioni troppo datate che quindi non rappresentano più adeguatamente il prodotto o servizio analizzato;
- Reviews che possiedono vocaboli inerenti ad altri prodotti/servizi offerti dalla stessa azienda, i cui clienti magari hanno erroneamente inserito;
- Recensioni che posseggono meno di 35 caratteri, perché povere di informazioni;
- Recensioni con più di 3500 caratteri, perché troppo ricche di contenuti e quindi con il rischio di avere informazioni che si allontanano dal prodotto o servizio;
- Recensioni che non rispettano il valore di caratteri ottenuto da una soglia dinamica che viene calcolata con la seguente formula:

$$DT_i = Q3_i + (1{,}5 \cdot IQR_i)$$

dove Q3 è il terzo quartile e IQR è il range interquartile;

Una volta ottenuto il database ripulito e con le effettive recensioni che verranno prese in esame nelle successive task, si procede con la fase di pre-processing che verrà analizzata nel prossimo paragrafo.

3.3 Pre-processing

Il pre-processing permette di effettuare una ulteriore pulitura del database, ma questa volta non verranno eliminate le recensioni bensì si lavora su di esse per rimuovere tutte quelle informazioni che non risultano utili al fine dello studio.

Le operazioni che vengono svolte in questa fase sono:

- La conversione del testo delle recensioni in minuscolo, che ha lo scopo di eliminare le possibili ambiguità causate da eventuali lettere maiuscole o da errori di battitura;
- L'eliminazione dei simboli di punteggiatura e dei caratteri numerici perché ai fini della successiva analisi dei topics del testo sono ritenuti non significativi;
- La rimozione delle stopwords, ovvero tutte quelle parole che hanno un'alta frequenza perché sono funzionali al linguaggio, ma allo stesso tempo sono prive di informazioni (ad esempio gli articoli o le preposizioni). Sono anche state rimosse tutte quelle parole che l'autore della tesi ha ritenuto prive di significato per l'analisi (stopwords personalizzate);
- L'eliminazione delle parole che si ripetono con una bassa frequenza nei vari documenti, in quanto possono derivare da errori di battitura o fare riferimento ad argomenti poco comuni;
- L'applicazione del processo di stemming, ovvero un procedimento che ha lo scopo di ottenere solamente le radici delle parole, per il semplice fatto che le parole che derivano da una stessa radice esprimono concetti simili;
- La rimozione delle parole non connesse al contenuto discusso dal topic;

- La sostituzione di tutti gli n-grammi, i quali sono sequenze di n elementi adiacenti che formano un'unica espressione (ad esempio "credit card" è stata rimpiazzata dal termine "creditcard").

Le funzioni utilizzate per eseguire ciascun punto sopracitato, appartenenti alla libreria dell'algoritmo STM, sono "textProcessor()" e "prepDocuments()", dove la prima ha permesso di preparare i dati, effettuare lo stemming e rimuovere le stopwords, mentre la seconda è stata utilizzata per strutturare e indicizzare i dati per poterli utilizzare nel topic model.

Ultimata questa fase, si ha il database definitivo sul quale poi verranno applicate le due diverse tipologie di modelli di topic modeling prese in esame nel presente studio.

3.4 Topic Modeling non supervisionato

In seguito al pre-processing, viene effettuata la fase di identificazione del numero ottimale di topics, definito con K.
Questo parametro definisce il numero ideale di argomenti che sono in grado di descrivere il dataset preso in esame, ed è un valore assolutamente critico nell'analisi della Voice of Customer, in quanto è direttamente coinvolto nella qualità e nell'interpretabilità dei risultati.

In generale, si può affermare che un numero eccessivamente piccolo di topics può creare argomenti troppo ampi ed eterogenei, mentre esattamente al contrario man mano che cresce il valore del parametro K si producono topics troppo specifici (Sbalchiero S. e Maciej E., 2020).
Per queste ragioni, è importante identificare un adeguato numero di topics in modo tale da ridurre il rischio di ottenere argomenti non adatti che non possiedono informazioni rilevanti per lo svolgimento delle ricerche analitiche oppure che vi siano topics troppo puntuali che li rendono poco interpretabili.

Ulteriori problemi possono derivare dal fatto che un dataset preso in esame può essere di bassa qualità o avere dimensioni non adeguate allo studio.

A tal proposito, vista la difficoltà per gli individui di identificare un K corretto, esiste una soluzione molto apprezzata a questo problema che prende il nome di "held-out likelihood", la quale è stata selezionata come misura di performance del topic model.

Questo indicatore valuta quanto bene un trained model (ovvero un insieme di documenti, tendenzialmente il 90% del totale, utilizzato come training set) riesce a spiegare l'held-out data, costituito dal restante 10% di dati non utilizzati per sviluppare il topic model (Barravecchia F. *et al.*, 2020).

Nello specifico l'held-out likelihood (L) viene formalmente definito come il logaritmo della probabilità (p) degli held-out data ($W_{held\text{-}out}$) forniti al trained model ($M_{trained}$):

$$L = \log p(W_{held-out}|M_{trained})$$

In pratica viene calcolata la verosomiglianza tra i risultati ottenuti dal trained model e dall'held-out data, ottenendo quindi una metrica sulla bontà delle performance con un certo numero di argomenti.

A livello teorico bisognerebbe prendere in considerazione il numero ottimale di topics che massimizza il valore di held-out.

Questo indicatore permette anche di prevenire l'overfitting, ovvero un fenomeno per il quale un algoritmo di topic modeling migliorerebbe le sue performance solamente attraverso il crescere del numero di topics, e di conseguenza una situazione limite in cui spiega ciascun documento con sé stesso.

Esistono anche altri indicatori che aiutano ad identificare il numero ottimo di topics, come ad esempio i residuals oppure la semantic coherence, ma non verranno presi in considerazione in questa tesi.

Nel codice è stato scelto di valutare da 5 a 50 topics e la funzione apposita utilizzata in R per trovare il numero di topics è "searchK()".

Plottando l'output che si ottiene dall'algoritmo, si riesce a capire quale valore prendere come riferimento per il numero ottimale di topics, ovvero quelli in corrispondenza del massimo locale di Held-Out Likelihood.

Una volta terminata la fase di identificazione del numero di topics, si può passare all'applicazione dello Structural Topic Modeling (STM) sulle recensioni pre-processate utilizzando il numero ottimale di topics (K) trovato in precedenza.
Questo processo, infatti restituisce i due insiemi di valori di probabilità (per le definizioni si rimanda alla sezione 2.5):

3. Topical content (φ);
4. Topical prevalence (ϑ).

Una successiva fase è quella di labelling, ovvero si effettua una identificazione delle etichette semantiche per ciascun topic (in questo caso dovremo svolgerlo quattordici volte, una per ogni K individuato).
Essa viene svolta manualmente, in quanto non esistono ancora tecniche che permettono un'etichettatura automatica (Barravecchia F. *et al.*, 2020), ma comunque la si realizza tramite l'aiuto dell'algoritmo di topic modeling, il quale permette di ottenere varie informazioni:

1. Highest Prob: parole che hanno la maggior probabilità di appartenere a ciascun topic. Esse derivano dalla matrice del topical content in ogni topic;
2. FREX: indicatore che prende in considerazione la frequenza assoluta di ogni parola e quanto sono esclusive rispetto al topic. Viene calcolato per ciascuna parola rispetto ad ogni argomento.

3.5 Topic Modeling semi-supervisionato

L'esecuzione di questo modello e del suo corrispondente algoritmo viene fatto partire dal database utilizzato per il topic modeling non supervisionato, e questo ci permette quindi di iniziare da un punto in comune e notare poi eventuali differenze.

L'algoritmo utilizzato, come già precedentemente citato nel paragrafo 2.6, fa parte dello studio "Seeded-LDA for Topic Modeling" del 14 Ottobre 2022, utilizzando il quanteda package e la libreria GibbsLDA++, al quale verranno però effettuate alcune modifiche per adattarlo al caso specifico preso in esame nella presente tesi. Per la spiegazione specifica dei passaggi e comandi di questo algoritmo si rimanda all'appendice: "guida all'utilizzo di Seeded-LDA".

In generale, il Seeded-LDA permette agli utenti di pre-definire i topics attraverso parole chiave (seed) utili per eseguire analisi basate sull'analisi dei dati testuali nelle scienze sociali e umanistiche (Watanabe K. e Zhou Y., 2020).
Infatti dopo aver applicato alcuni comandi per la pulizia del database, si può partire con l'attività chiave che caratterizza questa tipologia di processo, ovvero trovare i seed dei topics (cioè tutte quelle parole che ci fanno capire di che cosa parla il topic) del servizio/prodotto preso in esame, in modo tale da poterli inserire nell'algoritmo ed ottenere un output più coerente.

In questa tesi verranno considerati gli argomenti K trovati con il precedente modello non supervisionato e le parole che li caratterizzano maggiormente.

Allo stesso tempo si effettua anche un perfezionamento dei topics elaborati, ovvero si procede con una loro analisi e, nel caso venga ritenuto che un argomento non sia particolarmente utile ai fini della definizione delle determinanti di qualità del prodotto o servizio, allora si potrà scartarlo.
Ad esempio, se un argomento è molto generico come potrebbe essere uno che parla della soddisfazione generale, allora dovrà essere rimosso.

Non solo può avvenire l'eliminazione di alcuni topics, ma nel caso in cui esistono due o più di essi che vengono ritenuti trattare argomenti simili ma con delle sfumature leggermente diverse, allora si possono accorpare in un unico topic.
Ad esempio, se un argomento tratta delle funzionalità di un servizio e un altro delle caratteristiche del servizio offerto, allora potranno essere uniti in un unico topic.

Di seguito, in figura 6, viene riportato un flow chart in cui viene schematizzato il ragionamento che sta dietro al fatto di mantenere o eliminare un topic oppure se unire i simili.

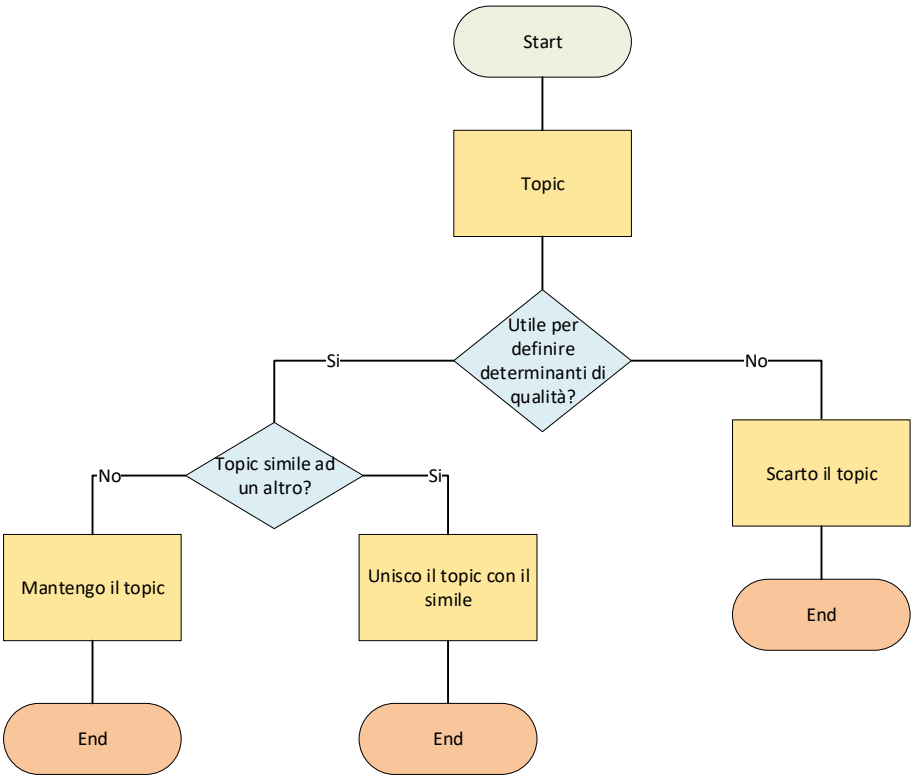


Figura 6 - Diagramma di flusso sulla scelta di mantenere, eliminare o accorpare i topics

Successivamente, per ogni topic trovato vengono inseriti i seed all'interno dell'algoritmo e infine attraverso un ulteriore comando vengono restituiti i valori di topical prevalence (ϑ), utili ai fini del lavoro per il confronto con l'altro modello non supervisionato.

Sui dati estratti verranno poi eseguiti i calcoli necessari per effettuare la validazione e trovare quindi i vari indici di performance che saranno poi confrontati con quelli ottenuti dall'algoritmo di topic modeling non supervisionato.

Si noterà nel prossimo paragrafo che l'output ottenuto dall'algoritmo possiede un ulteriore topic rispetto a quelli definiti in precedenza chiamato "other", perché vengono processati dei topics che hanno anche del rumore al loro interno e quindi ci può essere un argomento in cui vengono inseriti tutti quei concetti non legati agli altri topics dati in input.
Esso verrà considerato nella successiva fase di validazione e analisi dei risultati.

3.6 Validazione dei risultati

L'ultima attività richiesta per poter poi procedere con l'analisi dei risultati è la validazione dei risultati, il cui obiettivo è quello di verificare se effettivamente l'algoritmo ha associato correttamente i topics ai documenti, quindi questa fase permette di valutarne la sua efficacia.
In pratica permette di stabilire l'attendibilità del risultato ottenuto, quindi deve essere sempre connessa al processo di topic modeling.
Infatti non basta concentrarsi solamente sulle tecniche di estrapolazione dei topic, ma è anche necessario effettuare una validazione dei risultati ottenuti (Barravecchia *et al.*, 2022).

Inoltre, questa fase avrà un'importanza considerevole in questa tesi perché permetterà di quantificare il confronto tra i metodi di topic modeling e verificare se si ottengono risultati più consistenti inserendo dei seed di ancoraggio per l'algoritmo (topic modeling semi-supervisionato), e quindi se alle aziende conviene applicare questa tipologia oppure al contrario non vengono restituiti benefici tali da motivare l'utilizzo di una parte aggiuntiva di codice e quindi conviene rimanere con una versione non supervisionata del processo.

Il problema che nasce durante la validazione è che non esiste un procedimento standardizzato per valutare l'output.
Per risolvere questa questione, nella presente tesi viene utilizzato il metodo proposto da Barravecchia *et al.* (2022), che consiste nel calcolare una serie di indicatori che mettono in confronto l'assegnazione umana dei topics (riferito ad un campione casuale più piccolo dell'intero dataset preso in considerazione) con

quella automatica svolta dall'algoritmo STM sul dataset completo (ovvero la topical prevalence ϑ).

In pratica viene eseguita svolgendo i seguenti punti:

1. Estrazione di un campione casuale del dataset di recensioni con le rispettive topical prevalence (in questo caso di studio vengono considerate 100 recensioni casuali);
2. Assegnazione umana ad uno o più topics di tutte le recensioni casuali in base a quanto un valutatore pensa siano correlati tra loro;
3. Calcolo di una soglia dinamica per ogni recensione derivante dalla matrice di topical prevalence;
4. Confronto tra la probabilità di ciascuna recensione di appartenere ad ogni topic e la soglia dinamica precedentemente calcolata: se la probabilità è maggiore o uguale alla soglia, il topic è considerato rilevante per quanto riguarda l'assegnazione automatica;
5. Creazione di una matrice di confusione per ogni recensione, formatasi in conseguenza di un confronto tra i topics rilevanti stabiliti dal valutatore umano e quelli identificati dall'algoritmo STM.
   In tabella 1 viene riportata la matrice usata come riferimento:

		Assegnazione umana del topic	
		Assegnazione al topic T_i	Non-assegnazione al topic T_i
Assegnazione automatica del topic	Assegnazione al topic T_i	**Vero positivo (true positive tp)** Inferenza corretta Accordo tra l'assegnazione umana e automatica. Entrambe le procedure riconoscono la presenza di un topic in una recensione.	**Falso positivo (false positive fp)** Errore di I specie Disaccordo tra l'assegnazione umana e automatica. La seconda riconosce la presenza di un topic in una recensione, mentre la prima no.
	Non-assegnazione al topic T_i	**Falso negativo (false negative fn)** Errore di II specie Disaccordo tra l'assegnazione umana e automatica. La prima riconosce la presenza di un topic in una recensione mentre la seconda no.	**Vero negativo (true negative tn)** Inferenza corretta Accordo tra l'assegnazione umana e automatica. Entrambe le procedure non riconoscono la presenza di un topic in una recensione.

Tabella 1 – Matrice di confusione

6. Calcolo degli indicatori necessari per effettuare la valutazione prendendo come riferimento i valori aggregati della matrice di confusione.
   L'aggregazione consiste nelle seguenti quattro sommatorie, che riguardano ogni dimensione D della matrice:
   - Totale dei veri positivi: $TP = \sum_{i=1}^{D} tp_i$;
   - Totale dei falsi positivi: $FP = \sum_{i=1}^{D} fp_i$;
   - Totale dei veri negativi: $TN = \sum_{i=1}^{D} tn_i$;
   - Totale dei falsi negativi: $FN = \sum_{i=1}^{D} fn_i$;

   Gli indicatori che si calcoleranno sono elencati in tabella 2.

Indicatore	Descrizione	Formula	Valore target
Accuracy	L'accuratezza valuta l'efficacia dell'algoritmo in base alla sua percentuale di previsioni corrette.	$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$	>0,95
Precision	La precisione è una stima della probabilità che una previsione positiva sia corretta.	$Precision = \frac{TP}{TP + FP}$	>0,70
Recall	Il recall è la frazione della quantità totale di istanze rilevanti che sono state effettivamente recuperate.	$Recall = \frac{TP}{TP + FN}$	>0,70
F_1 score	Il F_1 score è una misura dell'accuratezza del test e dipende dai valori di precisione e recall.	$F1\ score = 2 * \frac{precision * recall}{precision + recall}$	>0,70
Fall-out	Il fall-out è la probabilità condizionata di rilevare un argomento che in realtà non è presente.	$Fall - out = \frac{FP}{FP + TN}$	<0,05
Miss rate	Il miss rate è la probabilità che l'algoritmo di topic modeling non riesca ad identificare un argomento quando in realtà è presente.	$Miss\ rate = \frac{FN}{TP + FN}$	<0,20
Specifity	La specificità misura la probabilità di trovare dei veri negativi.	$Specificity = \frac{TN}{FP + TN}$	>0,90
Negative predective value	Il negative predective value è la probabilità che l'algoritmo di topic modeling non rilevi un argomento quando non è effettivamente presente.	$NPV = \frac{TN}{TN + FN}$	>0,90
False omission rate	Il false omission rate è la proporzione di argomenti non rilevanti quando l'argomento era invece presente.	$FOR = \frac{FN}{FN + TN}$	<0,05

False discovery rate	Il false discovery rate è la proporzione di argomenti identificati erroneamente rispetto a tutti gli argomenti identificati.	$FDR = \frac{FP}{TP + FP}$	<0,05

Tabella 2 – Indicatori di performance

In figura 7 vengono schematizzate le principali fasi di validazione delle procedura proposta qui sopra.

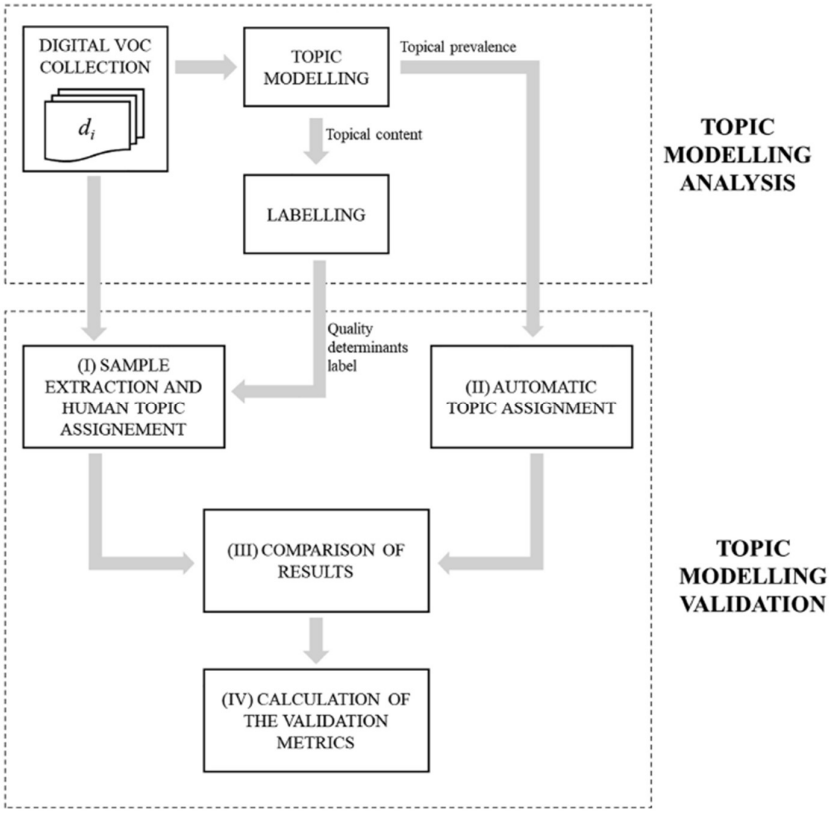


Figura 7 – Fasi della validazione dei risultati (Barravecchia et al., 2022)

3.7 Analisi e confronto dei risultati

Una volta calcolati gli indicatori, si effettua un primo confronto tra i valori ottenuti e quelli target per ogni tipologia di topic modeling, in modo tale da verificare immediatamente se l'output ottenuto può essere considerato soddisfacente.

Stabilito quindi che i risultati di entrambi i modelli garantiscono delle buone performance, si procede con il loro confronto, necessario per stabilire quale tra i due è migliore.
Inoltre, per completezza dell'analisi, verrà eseguito anche un confronto sulle matrici di confusione ottenute, in modo tale da poter spiegare cosa non va negli indicatori calcolati e perché un modello risulta meno performante dell'altro.

4 APPLICAZIONI SPERIMENTALI

In questo capitolo vengono dapprima introdotti i due casi di studio scelti per effettuare il lavoro, successivamente esposti i risultati ottenuti per ciascuna tipologia di topic modeling seguendo la metodologia spiegata nel precedente paragrafo, per poi alla fine stabilire per ciascun caso quale tra l'algoritmo non supervisionato e quello semi-supervisionato ha garantito performance migliori e quindi più conveniente da utilizzare per le aziende che si avvalgono di questi approcci innovativi.

4.1 Primo caso di studio – Uber

4.1.1 Descrizione del servizio

Il servizio sul quale si focalizzerà la prima parte dello studio è Uber, cioè una piattaforma il cui business principale è quello di offrire un servizio di trasporto passeggeri (come una sorta di taxi), ma non solo, propone ad esempio anche un servizio di food delivery.
Tuttavia l'analisi, che verrà effettuata in questo elaborato sulle recensioni realizzate dai clienti, verterà esclusivamente sul core business di Uber.

Qui di seguito verrà spiegato come funziona il processo per poter comprendere meglio i futuri topics che verranno estratti attraverso le due tipologie di topic modeling.
In figura 8 sono elencate le macro-fasi che caratterizzano il customer journey di Uber, che sono:

Figura 8 – Customer Journey di Uber

Il processo che deve compiere un cliente per utilizzare il servizio consiste nell'accedere all'applicazione sul proprio smartphone e selezionare l'opzione "Viaggia".

Dopodiché inserisce tutti i dati necessari a completare la prenotazione di un driver, come ad esempio l'orario in cui dovrà avvenire la corsa, l'indirizzo di partenza, la destinazione, la tipologia di auto e così via.

Il sistema informativo elabora i dati inseriti ed identifica un percorso possibile, individuando le auto conformi alle esigenze del cliente.

Il driver più vicino viene notificato dal sistema informativo e sceglie se confermare la presa in carico della corsa oppure se rifiutarla; in quest'ultimo caso la richiesta viene passata all'autista successivo in ordine di vicinanza al cliente.

Dopo aver accettato la corsa, il driver si dirige verso il punto di partenza; nel frattempo l'utente vede aggiornato lo stato della sua richiesta e, da questo momento fino alla conclusione del servizio, può osservare in tempo reale la posizione e il tempo stimato di percorso.

Il servizio di Uber rende disponibili all'autista e al cliente i reciproci numeri telefonici, così da permettere fra loro una comunicazione semplice e diretta in caso di eventuali difficoltà nell'incontrarsi.

L'app informa l'utente dell'arrivo del driver e, prima di convalidare l'inizio della corsa, l'autista si accerta dell'identità del passeggero.

Raggiunta la destinazione finale, il driver conferma la conclusione della corsa; successivamente all'utente viene richiesto un feedback facoltativo e l'importo, anche nullo, da destinare come mancia all'autista.

Il processo si conclude con l'invio automatico al cliente della fattura di pagamento, tramite e-mail.

Tutte queste attività vengono schematizzate in figura 9 tramite un'apposita scheda di processo:

	ATTIVITA'	Touch Point	S.I.	Inizio/Fine
1	Accede all'app e seleziona "viaggia"	V		INPUT
2	Seleziona orario di partenza	V	V	
3	Inserisce indirizzo di destinazione	V	V	
4	Calcola percorso e individua driver disponibili		V	
5	Mostra tipologie di auto disponibili		V	
6	Sceglie l'auto	V	V	
7	Sceglie la modalità di pagamento	V	V	
8	Conferma il punto di partenza	V	V	
9	Invia notifica di richiesta corsa a driver disponibile		V	
10	Visualizza notifica e sceglie se accettare la corsa			
11	Rifiuta la corsa		V	
12	Conferma la corsa e si dirige verso il punto di partenza della corsa		V	
13	Notifica l'utente che la corsa è confermata		V	
14	Aggiorna e mostra all'utente in tempo reale la posizione del driver e il tempo di arrivo stimato	V	V	
15	Raggiunge il punto di partenza			
16	Notifica l'arrivo del driver all'utente		V	
17	Entra in auto			
18	Verifica identità dell'utente e conferma l'inizio della corsa		V	
19	Aggiorna e mostra all'utente in tempo reale la posizione del driver e il tempo di arrivo stimato		V	
20	Raggiunge la destinazione finale			
21	Conferma la conclusione della corsa		V	
22	Chiede all'utente un rilascio feedback e possibile mancia al driver		V	
23	Sceglie se valutare la corsa	V		
24	Rilascia feedback	V	V	
25	Imposta l'importo da rilasciare al driver come mancia	V	V	
26	Invia la fattura del pagamento all'utente via e-mail	V	V	OUTPUT

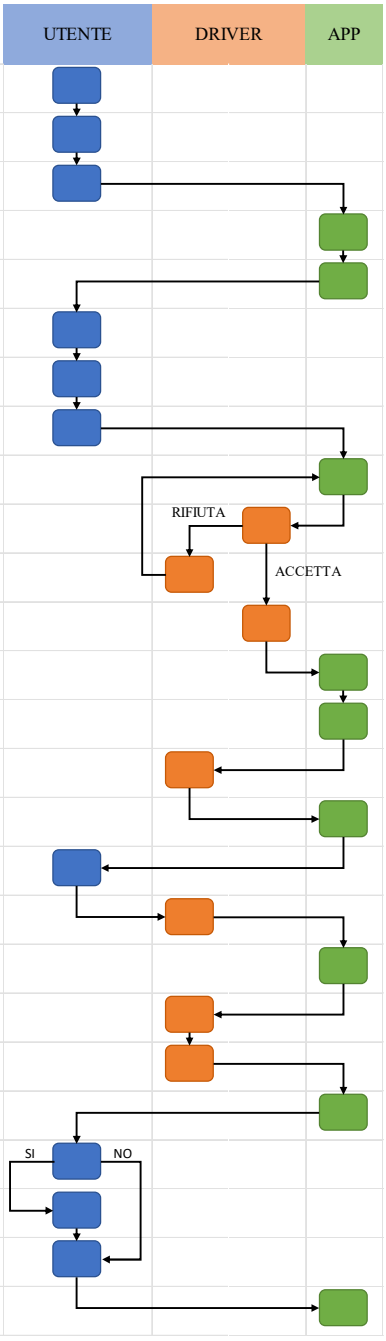


Figura 9 – Scheda di processo del servizio

4.1.2 Web Scraping

La prima fase su cui si opera è quella di web scraping, la quale permette di estrarre da internet le informazioni utili per lo sviluppo del lavoro.

Come già citato nella sezione di metodologia, sono state ricercate e poi scaricate recensioni inerenti al servizio di Uber di consumatori che parlano la lingua inglese, per avere una miglior elaborazione da parte dei software utilizzati (Octoparse e ParseHub) e una più ampia fornitura di reviews rispetto a quelle che si potrebbero ottenere in italiano.

Le fonti utilizzate per estrarre la Voice of Customer di Uber sono:

1. Google Play Store
2. ProductReview.com.au
3. Consumeraffairs.com
4. Yelp – Los Angeles
5. Hellopeter.com

I metadati estratti sono la data di inserimento della recensione, il testo della recensione e il rispettivo rating.
Inoltre, è stato aggiunto un codice identificativo ID e la rispettiva fonte di provenienza per ciascuna review.
Nella tabella 3 è possibile osservare un esempio delle informazioni estratte e salvate in un file di formato .csv:

Data	ID	Recensione	Rating	Fonte
09/03/2015	ID00001	Good car transportation service helps you get around places quicker cheap and very safe to use	5	Yelp
28/06/2014	ID00002	Easy to use app, background checks and	4	Yelp

		there's almost always a driver en route!		
27/02/2018	ID00003	Private Taxi service that is safe and reliable, even through the traffic Uber is on time	4	HelloPeter
08/09/2021	ID00004	My Trip was cancelled twice and they took my money and till today have not come back to me or refund . This is terrible	1	HelloPeter
14/02/2022	ID00005	Why have this service if its going to be cancelled multiple times for a short trip,this has happened to me over the last few weeks,so back to taxi for me	1	ProductReview

Tabella 3 – Esempio recensioni e metadati estratti

Le recensioni estratte si riferiscono al periodo compreso tra il 2012 e il 2022 e sono pari a 12399.

Ad esse è stata applicata una prima di pulitura, ovvero sono state eliminate quelle scritte nel 2012 e 2013, in quanto rappresentavano una percentuale poco significativa rispetto all'intero database (1,09% del totale) e facevano riferimento agli inizi dell'entrata nel mercato di Uber, quindi i clienti avevano meno pretese sulla qualità del servizio.

Inoltre sono state eliminate anche quelle recensioni che contenevano vocaboli attinenti al servizio Uber Eats, come ad esempio Uber Eats, food, drink, eat, delivery, meal, lunch, dinner, facendo attenzione a ricercare tutte le loro varianti.

Ulteriori step per depurare il database di Uber sono quelli elencati nel paragrafo 3.2 della sezione metodologia, ovvero sono state eliminate anche le reviews che non rispettano un certo numero di caratteri:

- < 35 caratteri;
- > 3500 caratteri;
- > del valore della soglia dinamica calcolata.

Il database finale così ottenuto contiene 10197 recensioni e sarà utilizzato per effettuare il topic modeling semi-supervisionato e non, per poi poter confrontare la bontà degli output ottenuti.

Le principali caratteristiche del campione ottenuto sono esposte nelle seguenti figura 10, figura 11 e figura 12:

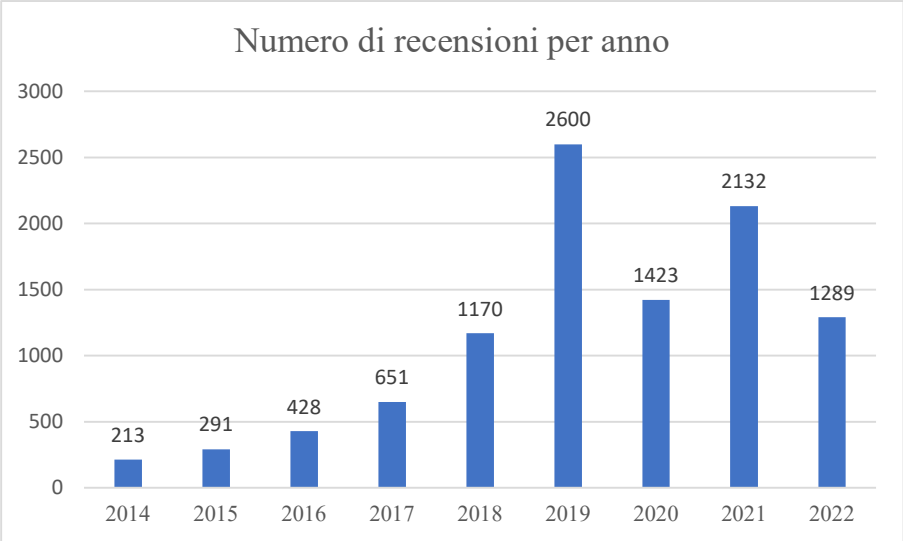


Figura 10 – Numero di recensioni per anno

Da questa figura si nota come ci sia stata una crescita del numero di recensioni man mano che il servizio si diffondeva, per poi fermarsi nel 2020 a causa della situazione pandemica globale.
L'anno 2022 possiede meno recensioni perché non è stato considerato nella sua interezza.

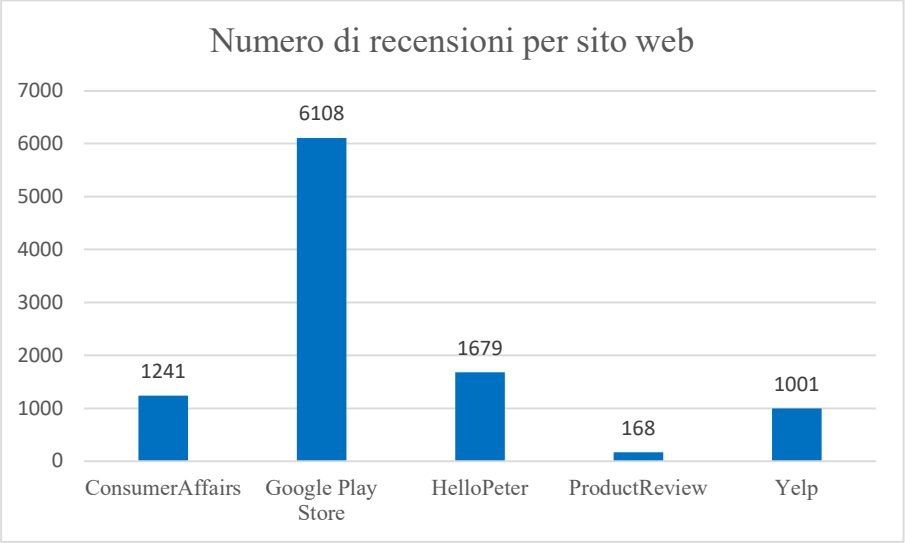


Figura 11 – Numero di recensioni per sito web

Dal grafico soprastante si vede immediatamente che il principale sito di aggregazione di recensioni per la creazione del database è stato il Google Play Store.

Infatti, essendo che Uber è un servizio utilizzato dallo smartphone, è più facile che un utente scriva la sua opinione direttamente sul sito da cui scarica l'applicazione, senza dover andare su pagine web specifiche.

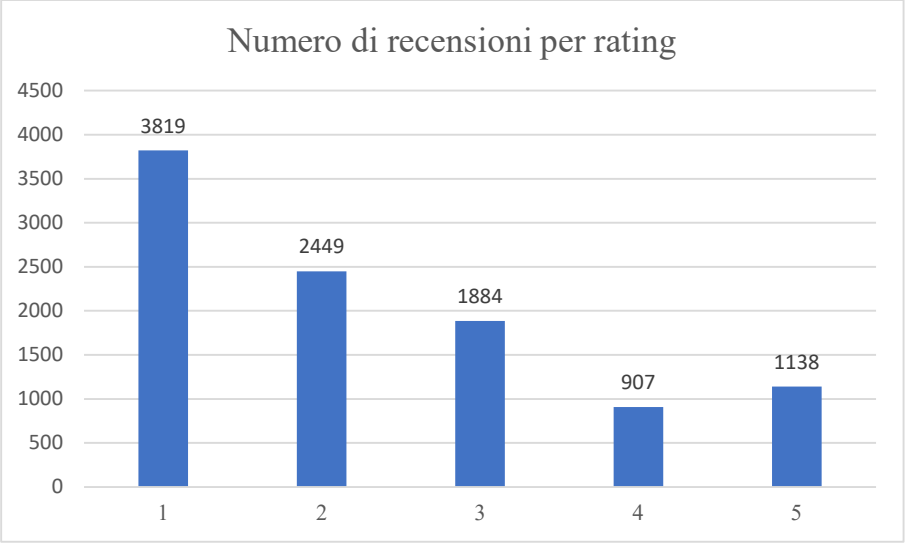


Figura 12 – Numero di recensioni per rating

Il grafico mostrato in figura 12 dà un'idea generale sulla valutazione che i clienti forniscono al servizio.
Per quanto riguarda Uber, si ha una tendenza discendente man mano che il ranking cresce, indice che le persone non gradiscono molto il servizio in oggetto.
Infatti, la maggioranza dei voti è pari a 1, ovvero la valutazione più bassa che possano inserire.

4.1.3 Topic Modeling non supervisionato

Una volta ottenuto il database si può iniziare la fase di topic modeling (semi-supervisionato e non) per trovare le determinanti latenti di qualità del servizio Uber.

Nel caso dell'algoritmo non supervisionato in questa tesi viene usata la tecnica STM, citata nei precedenti paragrafi, che verrà implementata attraverso il software RStudio.

Innanzitutto si effettua la fase di pre-processing, la cui compilazione della rispettiva riga di codice restituisce il database di input ripulito di 1 documento che non conteneva parole e di 12000 su 13551 termini a causa della loro frequenza.
L'output depurato ottenuto quindi è comprensivo di 10196 documenti, 1551 termini e 232715 tokens.
La seguente tabella 4 mostra in sintesi i risultati ottenuti sul dataset in seguito al pre-processing:

Indicatore	**Uber**
Dataset originale	12399
Dataset post soglia dinamica e eliminazione recensioni <35 caratteri	10197
Dataset post pre-processing	10196
Riduzione volume dati	≈18%

Tabella 4 – Output pre-processing Uber

Dopodiché si identifica il numero ottimale di topics (K) attraverso l'indicatore "held-out likelihood" (per approfondimenti si rimanda al paragrafo 3.4) calcolato dall'algoritmo.

L'output che si ottiene è stato plottato anche insieme ad altri indicatori restituiti dall'algoritmo (figura 13) e graficamente si nota come siano possibili due risultati:

- un K pari a 14 (possiede un valore di held-out likelihood di -6,126);
- un K pari a 23 (con un valore di held-out likelihood di -6,116).

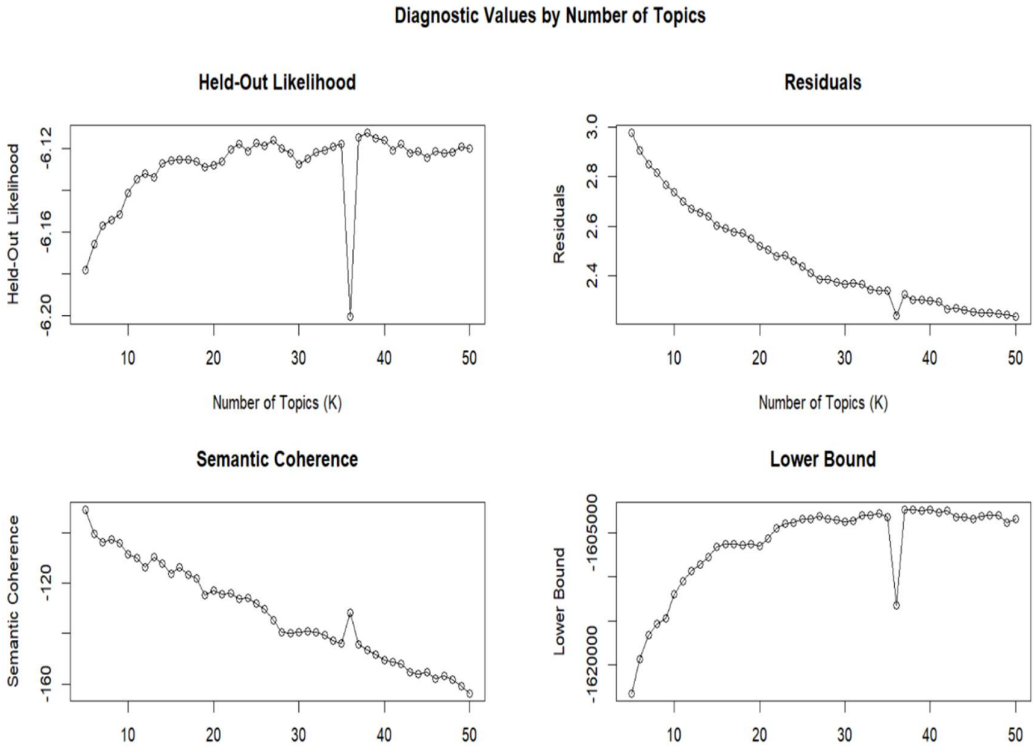


Figura 13 – Output numero ottimale di topics

Come si vedrà nei successi paragrafi il valore che è stato ritenuto ottimale è quello con K=14, ma in una fase della presente tesi a valle dell'applicazione del topic modeling non supervisionato, per effettuare un successivo controllo sugli indicatori di qualità degli algoritmi di topic modeling semi-supervisionato, verrà valutato se modificare il presente valore di K ottimale.

Finita questa fase si è passati all'applicazione dello Structural Topic Modeling (STM) che permette di effettuare l'attività di labelling, ovvero l'etichettatura dei topics, e restituisce i valori di topical content e topical prevalence, dove i secondi saranno necessari per ottenere gli indicatori richiesti dalla fase di validazione.

In tabella 5 vengono inserite le parole chiave ottenute automaticamente dall'algoritmo che permettono di stabilire le etichette, le vere e proprie etichette (definite label) e la descrizione di cosa tratta ciascun topic, entrambi ottenuti manualmente in seguito ad un'analisi dell'output.

N°	Keywords	Label
1	Highest Prob: driver, cancel, trip, charg, wait, fee, accept, min, minut, request FREX: cancel, fee, accept, min, trip, wait, anoth, request, driver, move	**CANCELLAZIONE CORSA**
	Descrizione: riguarda le richieste di annullamento della prenotazione della corsa da parte degli utenti.	
2	Highest Prob: car, driver, vehicl, drive, passeng, dollar, safe, nice, clean, safeti FREX: clean, car, safe, luggag, seat, insur, smell, vehicl, polit, damag 8	**ESPERIENZA DI VIAGGIO**
	Descrizione: tiene conto di diversi aspetti come qualità del viaggio, aspetto dell'auto nonché della professionalità del conducente.	
3	Highest Prob: use, driver, servic, rate, realli, taxi, experi, star, peopl, alway FREX: rate, star, recommend, conveni, easi, taxi, better, year, bit, overal	**SODDISFAZIONE GENERALE**
	Descrizione: riguarda il livello di soddisfazione generale del passeggero e tiene conto dei diversi aspetti che caratterizzano il servizio complessivo.	
4	Highest Prob: account, money, email, refund, trip, charg, still, receiv, request, bank FREX: account, refund, deduct, bank, log, sign, password, email, receiv, week	**GESTIONE ACCOUNT**
	Descrizione: si riferisce a lamentele dell'utente circa problemi di gestione dell'account, quali addebiti non autorizzati e difficoltà di accesso al proprio profilo personale.	

5	Highest Prob: app, locat, updat, problem, issu, fix, pleas, work, map, add FREX: updat, fix, locat, uninstal, app, select, add, screen, map, enter	**FUNZIONAMENTO APPLICAZIONE**
	Descrizione: riguarda problemi con le funzionalità offerte dall'applicazione e problemi di funzionamento in generale.	
6	Highest Prob: time, minut, wait, driver, everi, mani, arriv, late, alway, sometim FREX: time, wast, everi, mani, arriv, minut, late, journey, three, long	**ATTESA STIMATA**
	Descrizione: si riferisce ad errori nella stima del tempo di attesa per l'arrivo del driver.	
7	Highest Prob: driver, way, rout, around, home, walk, wrong, area, direct, lyft FREX: walk, rout, road, around, turn, gps, street, area, traffic, direct	**PROBLEMI DI LOCALIZZAZIONE**
	Descrizione: riguarda il mancato aggiornamento dei percorsi che genera l'allungamento del tragitto o errori nell'indirizzo di partenza e destinazione.	
8	Highest Prob: charg, card, price, credit, cost, use, surg, lyft, quot, delet FREX: quot, surg, card, credit, rip, delet, price, gift, cost, fraud	**ADDEBITI ERRATI**
	Descrizione: si riferisce a problematiche di addebito: costi maggiori rispetto a quanto preventivato o addebiti non autorizzati.	
9	Highest Prob: ride, schedul, work, tip, tri, find, morn, confirm, app, get FREX: schedul, tip, ride, advanc, earli, flight, morn, share, strand, appoint	**PRBLEMI GESTIONE CORSA**
	Descrizione: riguarda problematiche inerenti alla schedulazione di una corsa, suddivisione di quest'ultima con un altro passeggero, condivisione delle informazioni della stessa con altri.	
10	Highest Prob: driver, contact, number, tri, help, call, lost, report, noth, messag FREX: number, lost, contact, assist, deliveri, son, forgot, heard, return, websit	**OGGETTI SMARRITI**
	Descrizione: fa riferimento a oggetti persi dai passeggeri, ai quali Uber non ha fornito alcun tipo di supporto.	

11	Highest Prob: servic, provid, travel, transport, becom, driver, pathet, distanc, destin, rider FREX: transport, travel, provid, pathet, becom, altern, disgust, public, special, per	**SERVIZIO DI TRASPORTO**
	Descrizione: riguarda aspetti del servizio di trasporto in generale.	
12	Highest Prob: custom, servic, issu, support, care, complaint, compani, respons, resolv, even FREX: custom, care, support, complaint, resolv, question, communic, repli, issu, read	**CUSTOMER SERVICE**
	Descrizione: si riferisce a problemi con il supporto fornito ai clienti da Uber.	
13	Highest Prob: pay, payment, fare, cash, amount, paid, money, driver, chang, method FREX: cash, payment, pay, paid, method, fare, toll, amount, paytm, full	**METODI DI PAGAMENTO**
	Descrizione: riguarda aspetti e problematiche inerenti alle modalità di pagamento.	
14	Highest Prob: book, cab, reach, auto, driver, experi, show, tri, destin, station FREX: book, auto, cab, reach, station, train, earlier, mom, come, face	**PRENOTAZIONE CORSA**
	Descrizione: fa riferimento ad aspetti e problematiche circa la prenotazione della corsa.	

Tabella 5 – Label, keywords e descrizione dei topic di Uber

Tutte le voci ottenute corrispondono alle determinanti latenti di qualità di Uber che sono state captate dalle numerose recensioni online.

Una volta che si possiede questo output si può procedere con i calcoli della validazione, seguendo i passaggi elencati nel precedente paragrafo, i cui risultati verranno esposti nelle seguenti tabelle 6 e 7:

		Assegnazione umana del topic	
		Assegnazione al topic T_i	Non-assegnazione al topic T_i
Assegnazione automatica del topic	Assegnazione al topic T_i	**True positive** TP=106	**False positive** FP=43
	Non-assegnazione al topic T_i	**False negative** FN=49	**True negative** TN=1202

Tabella 6 – Matrice di confusione di Uber (topic modeling non supervisionato)

Indicatore	Risultato ottenuto	Valore target
Accuracy	0,9343	>0,95
Precision	0,7114	>0,70
Recall	0,6839	>0,70
F_1 score	0,6974	>0,70
Fall-out	0,0345	<0,05
Miss rate	0,3161	<0,20
Specifity	0,9655	>0,90
Negative predective value	0,9608	>0,90
False omission rate	0,0392	<0,05
False discovery rate	0,2886	<0,05

Tabella 7 – Indicatori di performance di Uber (topic modeling non supervisionato)

In linea generale gli indicatori calcolati mostrano delle performance soddisfacenti dell'algoritmo di topic modeling non supervisionato, in riferimento a questo numero di topics.

Gli unici due che si discostano molto dai valori target sono il miss rate e il false discovery rate.
Il primo fa pensare che l'algoritmo non riesca ad individuare dei topics effettivamente presenti nel dataset, e quindi restituisca tanti falsi negativi; mentre il secondo indice dice che esistono dei topics che non rappresentano correttamente la loro idea centrale, rispetto a tutti gli altri argomenti, quindi esistono un numero più o meno rilevante di falsi positivi.

4.1.4 Topic modeling semi-supervisionato

Partendo dallo stesso database utilizzato per il topic modeling non supervisionato, per avere un punto in comune tra le due metodologie, e averlo ripulito nuovamente, si procede con la definizione dei topics, ovvero se gli argomenti trovati in precedenza vadano bene oppure vadano eliminati o accorpati con altri.

Nel caso specifico del presente elaborato è stato ritenuto che per il servizio di Uber diano informazioni troppo generiche e che quindi possano essere esclusi dall'analisi, i seguenti topics:

- Soddisfazione generale;
- Funzionamento applicazione;
- Servizio di trasporto.

In particolare il topic "Funzionamento applicazione" è stato ritenuto che al suo interno comprendesse gli argomenti "Problemi di localizzazione", "Addebiti errati", "Metodi di pagamento" e "Prenotazione corsa", quindi è stata fatta la scelta di lasciare solo i casi specifici che dettagliano il servizio.
In contemporanea è stato considerato che i topics "Addebiti errati" e "Metodi di pagamento", si riferiscono entrambi al macro-argomento del pagamento ma con un'accezione leggermente diversa, quindi è stata fatta la scelta di accorparli in un unico topic di nome "Pagamento del servizio".

Le attività appena descritte vengono riassunte nella seguente figura 14:

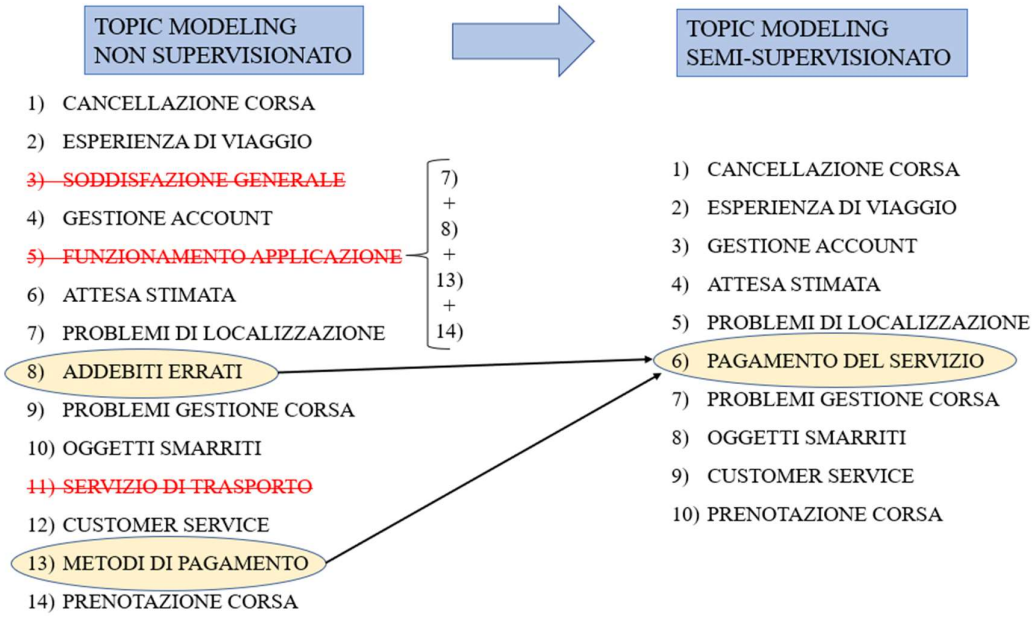


Figura 14 – Raggruppamento topic (la parentesi graffa stabilisce che l'argomento di riferimento è troppo generico e di conseguenza può essere rimosso in quanto i casi specifici sono dettagliati nei topic contenuti nella parentesi)

In seguito sono poi stati estratti i seed di ogni determinante di qualità (aiutandosi con le parole più frequenti per ogni topic stabilite dalla precedente tipologia di algoritmo).

Il risultato delle attività di definizione dei topics e dei seed per il servizio Uber è riportato nella seguente tabella 8:

TOPIC	**SEED**
CANCELLAZIONE CORSA	Cancel, wait, minut, request, penal, penalty, ban.
ESPERIENZA DI VIAGGIO	Car, driver, vehicl, drive, safe, nice, clean, safety, seat, smell, damage, insur, smoke, dirti.

GESTIONE ACCOUNT	Account, email, request, sign, password, code, login, log.
ATTESA STIMATA	Time, minut, wait, arriv, late, wast, journey, beforehand, fluctuat, shift, readi.
PROBLEMI DI LOCALIZZAZIONE	Way, rout, around, home, wrong, area, direct, road, gps, street, freeway, longest, neighborhood, shortest, entranc, corner, shorter, quicker.
PAGAMENTO DEL SERVIZIO	Charg, card, price, credit, cost, quot, fraud, inflat, surcharg, pay, payment, fare, cash, amount, paid, money, toll, paytm, tax, balanc.
PROBLEMI GESTIONE CORSA	Ride, schedul, find, confirm, app, get, earli, morn, advance, unlock.
OGGETTI SMARRITI	Contact, number, help, call, lost, report, noth, messag, assist, forgot, return, package, cellphon, ring, cell, key.
CUSTOMER SERVICE	Custom, issu, support, care, complaint, respons, resolv, question, communic, repli, resolute, servic.
PRENOTAZIONE CORSA	Book, auto, driver, destin, cab.

Tabella 8 – Topics e seed di Uber (topic modeling semi-supervisionato)

Una volta ottenuti i valori di topical prevalence (ϑ), si può procedere con la fase di validazione (per i dettagli si riveda il paragrafo 3.6) in comune con quella riferita al topic modeling non supervisionato.

Utilizzare lo stesso processo permette anche di confrontare i risultati ottenuti tra le due diverse metodologie di topic modeling, consentendo quindi di poter stabilire

attraverso una procedura quantitativa se l'utilizzo di una versione semi-supervisionata apporta dei miglioramenti, e di conseguenza se alle aziende conviene applicare quest'ultimo.

Per la fase di validazione del modello semi-supervisionato sono stati utilizzati i topics descritti nella precedente tabella 8.

Dal confronto tra i valori elaborati automaticamente dall'algoritmo e quelli ottenuti attraverso una comprensione umana delle recensioni che si riferiscono a determinati argomenti, sono state ottenute le seguenti somme di true positive, false positive, false negative e true negative riportate nella prossima tabella 9:

		Assegnazione umana del topic	
		Assegnazione al topic T_i	Non-assegnazione al topic T_i
Assegnazione automatica del topic	Assegnazione al topic T_i	**True positive** TP=58	**False positive** FP=36
	Non-assegnazione al topic T_i	**False negative** FN=44	**True negative** TN=962

Tabella 9 – Output matrice di confusione Uber (topic modeling semi-supervisionato)

Partendo da questi valori vengono calcolati gli indicatori sui quali si eseguirà il confronto tra le due tipologie di algoritmi di topic modeling.

Essi vengono riportati nella seguente tabella 10:

Indicatore	Risultato ottenuto	Valore target
Accuracy	0,9273	>0,95
Precision	0,6170	>0,70
Recall	0,5686	>0,70
F_1 score	0,5918	>0,70
Fall-out	0,0361	<0,05
Miss rate	0,4314	<0,20
Specifity	0,9639	>0,90
Negative predective value	0,9563	>0,90
False omission rate	0,0437	<0,05
False discovery rate	0,3830	<0,05

Tabella 10 – Output indicatori di performance Uber (topic modeling semi-supervisionato)

Dai risultati ottenuti dagli indicatori si nota come le performance dell'output dell'algoritmo di topic modeling semi-supervisionato non siano molto soddisfacenti.

Infatti ben cinque di essi risultano discostarsi di molto dai valori target, ovvero il 50% del totale.

Gli indicatori che non rispettano l'obiettivo sono:

1. Precision: questo porta a pensare che spesso quando l'algoritmo e l'uomo sono d'accordo sul riferimento di una certa recensione su un determinato topic, in verità non è corretto, nonostante comunque la differenza tra il valore calcolato e quello target non sia enorme;
2. Recall: vuol dire che non sono state recuperate molte istanze rilevanti sul totale, cioè che il numero di true positive non sia abbastanza elevato rispetto ai false negative, e questo porta ad avere una maggior quota di recensioni che secondo l'umano riguardano un determinato argomento mentre per

l'algoritmo no, e questo non va bene in quanto in un mondo ideale i due soggetti dovrebbero essere d'accordo;

3. F1 Score: il mancato rispetto del valore target è una conseguenza di precision e recall, in quanto il presente indicatore è dipendente da entrambi. Questo spiega una scarsa accuratezza del test;
4. Miss rate: porta a pensare che l'algoritmo è poco propenso a riuscire ad identificare un topic effettivamente presente;
5. False discovery rate: vuol dire che sono stati identificati molti topics errati, quindi il numero di falsi positivi è rilevante.

Per quanto riguarda l'accuracy, è stato ritenuto che nonostante non rispetti a pieno il valore target, è comunque vicino ad esso e quindi è stata considerata una buona performance a quel livello.

4.1.5 Confronto indicatori topic modeling non supervisionato e semi-supervisionato

Una volta ottenuti i valori di tutti gli indici necessari, si può procedere con l'obiettivo vero e proprio della presente tesi, ovvero quello di confrontare quale modello convenga applicare alle aziende per poter capire effettivamente quali sono le determinanti di qualità del proprio prodotto o servizio.

Innanzitutto si possono notare le differenze tra il totale di true positive, false positive, false negative e true negative in tabella 11, ovvero quei valori da cui dipendono i vari indicatori di performance dei modelli.

	Non supervisionato	**Semi-supervisionato**
TP	106	58
FP	43	36
FN	49	44
TN	1202	962

Tabella 11 – Quantità di true positive, false positive, false negative e true negative per tipologia di topic modeling

Essendo che il totale di tutti i valori è differente in quanto per il modello non supervisionato sono stati presi come riferimento 14 topics e 100 recensioni casuali, di conseguenza si avrà un totale di 1400 valori, mentre nel caso semi-supervisionato sono stati considerati 11 argomenti sempre per 100 recensioni casuali, quindi in totale si avranno 1100 valori, si procederà ad effettuare un calcolo percentuale della loro presenza, riportato in tabella 12:

	Non supervisionato	**Semi-supervisionato**
TP	7,57%	5,27%
FP	3,07%	3,27%
FN	3,50%	4,00%
TN	85,86%	87,45%

Tabella 12 - Percentuale di true positive, false positive, false negative e true negative per tipologia di topic modeling

Si nota come la quantità di false positive e di false negative siano simili in entrambi i modelli.

La vera differenza, quindi, si trova nei true positive e nei true negative.

I primi diminuiscono del 2,3% nel modello semi-supervisionato, questo potrebbe voler significare che questa tipologia di modello permette di essere più attenti ad individuare il topic a cui si sta riferendo una recensione, e quindi di ottenere una maggior precisione sia a livello automatico dell'algoritmo sia a livello umano, di conseguenza ottenendo un minor numero di argomenti che si riferiscono ad una recensione.

All'opposto potrebbe anche voler dire che non si riescono ad individuare dei veri argomenti da etichettare per ciascuna recensione.

Per capire quale delle due versioni è corretta si utilizzano gli indicatori, i quali garantiscono informazioni più strutturate e adeguate.

I secondi (true negative), invece, aumentano dell'1,59% con l'algoritmo di topic modeling semi-supervisionato, questo potrebbe significare che, continuando il discorso precedente, nel caso fosse più facile individuare gli argomenti di ciascuna

recensione, allora sarebbe anche più probabile che per gli altri topics non vi sia un collegamento, quindi verranno trovati più veri negativi.

Nella seguente tabella 13, verranno messi adiacenti i valori degli indicatori ottenuti con i due modelli in modo tale da avere una miglior percezione delle differenze tra i due.

Indicatore	Non supervisionato	Semi-supervisionato
Accuracy	0,9343	0,9273
Precision	0,7114	0,6170
Recall	0,6839	0,5686
F_1 score	0,6974	0,5918
Fall-out	0,0345	0,0361
Miss rate	0,3161	0,4314
Specifity	0,9655	0,9639
Negative predective value	0,9608	0,9563
False omission rate	0,0392	0,0437
False discovery rate	0,2886	0,3830

Tabella 13 – Valori degli indicatori per tipologia di topic modeling

Dalla presente tabella si nota come tutti gli indicatori calcolati per la versione non supervisionata di topic modeling siano migliori, e quindi garantiscano performance decisamente superiori rispetto alla controparte semi-supervisionata.
Infatti, per quegli indicatori la cui bontà viene espressa da un output maggiore, il modello non supervisionato è sempre risultato essere superiore rispetto all'altro approccio utilizzato, e questi indicatori sono:

- Accuracy;
- Precision;

- Recall;
- F_1 score;
- Specificity;
- Negative predective value.

Allo stesso tempo, anche per gli indicatori il cui obiettivo è quello di essere più piccoli possibile, il modello di topic modeling non supervisionato risulta ancora una volta essere migliore, e questi indicatori sono:

- Fall-out;
- Miss rate;
- False omission rate;
- False discovery rate.

Quindi concludendo il caso specifico del servizio Uber, si può affermare che applicare una metodologia più grezza, come quella utilizzata nella prima parte della tesi ovvero il topic modeling non supervisionato con tecnica STM, risulta garantire performance migliori.
La conseguenza di questa affermazione porta a dire che Uber, o più in generale qualsiasi azienda che produca beni o fornisca servizi al pubblico, possa continuare ad utilizzare questa tecnica, senza dover aggiungere parti di codice che non garantiscono un valore aggiunto alla loro ricerca, anzi portano ad una perdita di valore del tutto.

Ovviamente applicato ad un solo caso specifico, la conclusione non può avere un valore scientifico apprezzabile, quindi nei prossimi paragrafi verrà analizzato un ulteriore prodotto/servizio, con l'obiettivo di scoprire i risultati di quello studio e vedere se si ha un cambiamento nell'output finale e quindi con performance di topic modeling semi-supervisionato migliori dell'altro modello, oppure viene confermato quanto appena detto.

4.2 Secondo caso di studio – Smartphone

4.2.1 Descrizione del prodotto

La seconda analisi sulla quale verterà la presente tesi, prenderà come riferimento non più un semplice servizio, ma una tipologia di prodotto che fornisce una moltitudine di servizi come gli smartphone e la rispettiva digital Voice of Customer di tre specifici modelli.

Essi possono essere considerati facenti parte della macrocategoria dei product-service system (PSS), ovvero un insieme commercializzabile di prodotti e servizi la cui unione permette di soddisfare le esigenze di un cliente.
Non per forza un PSS deve essere fornito da un'unica azienda, ma possono essere realizzate delle partnership tra diverse imprese: nel caso specifico degli smartphone infatti c'è la società produttrice degli smartphone che realizza e mette a disposizione l'hardware sul quale poi potranno essere eseguiti servizi di terze parti tramite il download di applicazioni che permettono allo strumento di soddisfare le più diverse richieste di ciascun consumatore.

La sinergia tra prodotti e servizi ha permesso alle aziende manifatturiere tradizionali di evolversi, non solo per rimanere al passo con i cambiamenti del mercato, ma anche per aumentare i propri profitti grazie alla caratteristica intrinseca dei servizi di garantire un maggior valore aggiunto e di conseguenza una redditività migliore rispetto ai prodotti.
Infatti, di fronte alla contrazione dei mercati e all'aumento della mercificazione dei propri prodotti, le aziende manifatturiere hanno notato come la fornitura di servizi possa essere vista come un nuovo percorso verso profitti e crescita (Sutanto A. *et al.*, 2015).
I PSS hanno permesso alle aziende di non immettere nel mercato dei semplici prodotti, ma di creare delle vere e proprie "esperienze", che di conseguenza incuriosiscono maggiormente i possibili clienti e li rendono più propensi al consumo.

In generale possono essere identificate tre diverse classi di PSS (Tukker A. e Tischner U., 2006):

1. Product-oriented services: in questo caso la proprietà del prodotto è in mano al consumatore, e ad esso vengono aggiunti dei semplici servizi, come ad esempio un contratto di manutenzione;
2. Use oriented services: il prodotto rimane di proprietà del fornitore di servizi, il quale vende le sue funzioni tramite ottiche di condivisione, pooling o leasing;
3. Result-oriented services: il fornitore del PSS sostituisce i prodotti con dei servizi, ad esempio le segreterie telefoniche vengono sostituite dal voice-mail.

Tornando più nello specifico al settore degli smartphone è noto come nel corso degli anni ci sia stata un'ampia evoluzione e ad un perfezionamento di questi prodotti, grazie al miglioramento della tecnologia e alla crescente competizione che spinge i produttore a migliorare sempre di più per non perdere quote di mercato.

Questi fattori hanno anche portato, di anno in anno, ad avere smartphone con un ciclo di vita più breve e soprattutto a una loro omologazione sia a livello hardware sia software, portando quindi la concorrenza a sfidarsi sul prezzo del prodotto e sull'efficienza del software.

Tutto questo provoca una minore comprensione dei trend di mercato per le aziende produttrici, ma grazie alla digital Voice of Customer e quindi in particolare alle recensioni online realizzate dai consumatori, esse possono comprendere cosa viene richiesto dal pubblico, i loro interessi, le loro preoccupazioni, e quindi riescono a personalizzare il proprio PSS in modo tale da soddisfare il maggior numero di clienti e non rimanere indietro sulle esigenze in continuo mutamento.

4.2.2 Web Scraping

L'estrazione dei dati è stata nuovamente realizzata attraverso pagine web in lingua inglese per la quantità di recensioni disponibili grazie ai numerosi paesi anglofoni e anche perché i software di scraping lavorano meglio in questa lingua.

Sono stati utilizzati ancora una volta i software ParseHub e Octoparse, confermando una preferenza per il secondo.

Le recensioni estratte riguardano tre modelli diversi di smartphone: "IPhone 12" della Apple Inc., "Samsung Galaxy S20 FE" della Samsung e "Xiaomi Mi 10" della Xiaomi Inc..
La ragione della loro scelta è stata che soddisfano le caratteristiche degli attuali smartphone, pur non essendo usciti molto recentemente, inoltre nel momento del loro ingresso nel mercato appartenevano tutti a un'alta fascia di prezzo.

Le fonti utilizzate per la raccolta delle recensioni sono state:

1. Amazon;
2. Gadget360;
3. Ebay;
4. Idealo;
5. CustomerReview.

Le informazioni ritenute nuovamente rilevanti per lo svolgimento del lavoro sono la data in cui la recensione è stata inserita, il corpo della recensione, il rating e la fonte di provenienza.

Nella tabella 14 che segue è possibile osservare un esempio di metadati estratti e salvati in un file di formato .csv:

Data	Recensione	Rating	Fonte
24/01/2022	So far so good after over a month of use. Camera is very good and clear. Battery could be better Two charge a day for normal use. Finger print works flow less	4	Amazon
18/02/2021	Best features in this price range....awesome for gaming and photography .Camera gives high quality sharp images and screen refresh is also good	5	Gadget360
01/09/2021	Great phone. I upgraded from an phone 6 to a 12. I'm glad I made the purchase. I'm completely satisfied with this product.	4	Customer review
11/10/2021	Ordered phone twice and received defective product both time! Once it was some WiFi issues and later it was manufacturing defect!	1	Amazon
02/09/2021	Got a battery problem and an alert from phone itself that it cannot verify if the battery is genuine. Did not expect this within 2 weeks of buying.	1	Idealo

Tabella 14 – Esempio recensioni e metadati estratti

Il numero di recensioni di smartphone estratte è pari a 11369 distribuite lungo un arco temporale che parte da Ottobre 2020 ad Aprile 2022.
A questo dataset è poi applicata la soglia dinamica per ripulire il file da recensioni non accurate, ottenendo quindi un database contenente 10726 reviews, preso come riferimento per effettuare il topic modeling semi-supervisionato e non, e infine confrontare i due output.

Le principali caratteristiche del campione sono schematizzate nelle figure che seguiranno:

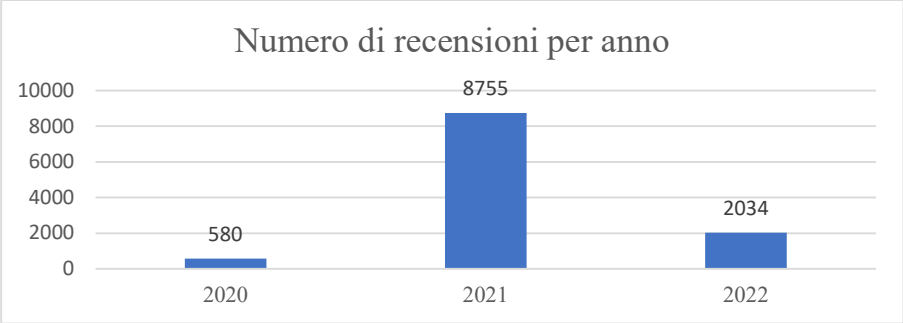


Figura 15 – Numero di recensioni per anno

Gli anni presi come riferimento sono pochi perché la vita degli smartphone è di qualche anno; inoltre considerando che due tipologie su tre (Iphone 12 e Samsung Galaxy S20 FE) sono uscite a fine 2020, si può affermare che le recensioni di quel periodo sono poche, e si ha correttamente un picco nel 2021 in quanto è l'anno di spicco nelle vendite di questi telefoni, mentre nel 2022 iniziano ad uscire le nuove versioni e quindi si ha un ulteriore calo di reviews.

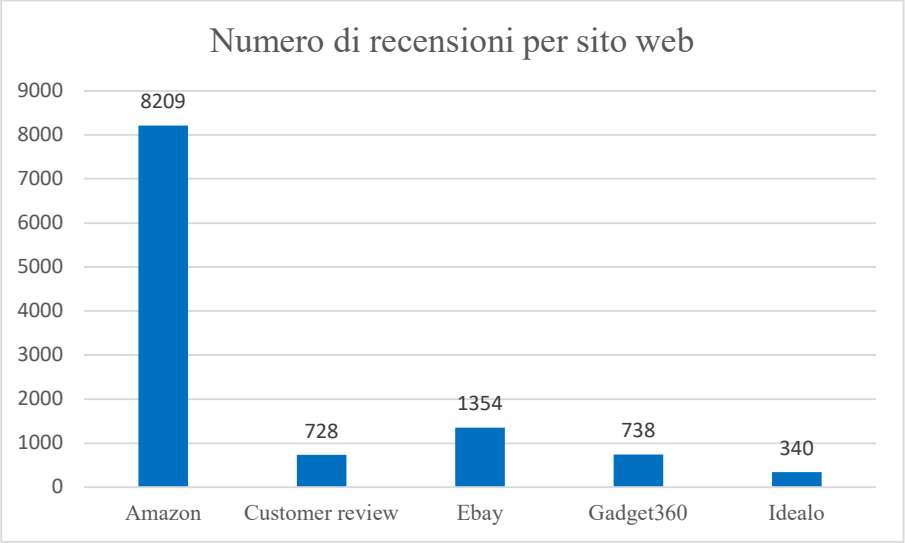


Figura 16 – Numero di recensioni per sito web

Il grafico soprastante mostra come la quasi totalità delle recensioni estratte provenga da Amazon, ovvero direttamente dall'e-commerce.

Probabilmente in quel sito si ha la possibilità di avere un maggior numero di recensioni perché una volta effettuato l'acquisto e ricevuto, è il sito stesso che ti invita a rilasciare la tua opinione, in modo tale da permettere anche ad altri utenti di capire com'è il prodotto desiderato e nel caso effettuare l'acquisto.

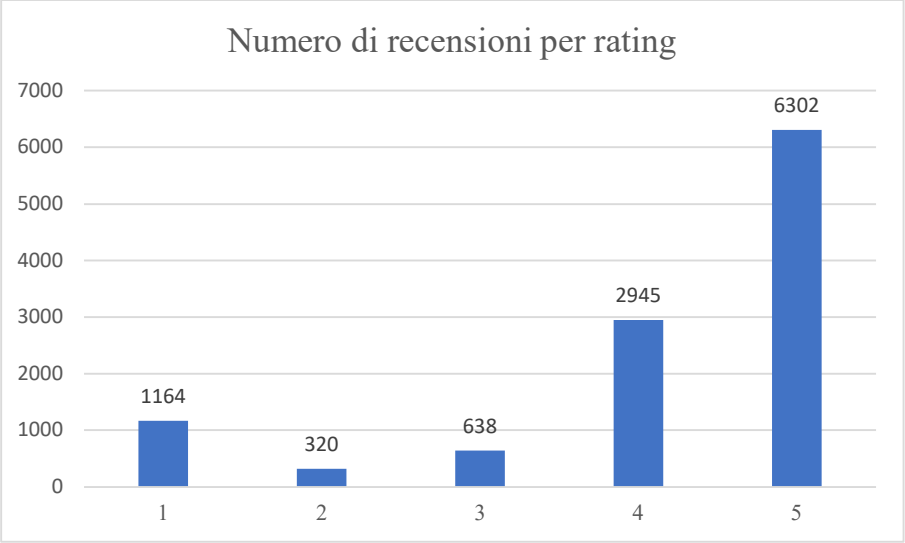


Figura 17 – Numero di recensioni per rating

In figura 17, si vede come i clienti abbiano apprezzato le tre tipologie di smartphone prese come riferimento nel presente studio.
Infatti si ha una tendenza crescente man mano che il ranking aumenta, raggiungendo il picco a 5, ovvero il voto più alto che un consumatore può inserire.

4.2.3 Topic Modeling non supervisionato

In seguito all'ottenimento del database, è possibile iniziare la fase di topic modeling necessario per trovare le determinanti di qualità utili alle aziende che producono le tipologie di smartphone prese come riferimento.

Tutte le fasi di topic modeling non supervisionato seguono la tecnica prevista dallo Structural Topic Modeling (STM) descritto nei precedenti paragrafi, in particolare nella sezione di metodologia.

Innanzitutto si procede con la fase di pre-processing, che sinteticamente comprende (per maggiori dettagli si rimanda al paragrafo 3.3):

1. La conversione del testo delle recensioni in minuscolo;
2. L'eliminazione della punteggiatura;
3. La rimozione delle stopwords;
4. L'eliminazione delle parole poco frequenti;
5. Il processo di stemming;
6. La rimozione di parole non collegate al contenuto del topic;
7. La sostituzione degli n-grammi.

Una volta ripulito il database, si può procedere con la fase di identificazione del numero ottimale di topics, ovvero K.
Data la difficoltà ad individuare questo valore, si ricorre di nuovo al metodo definito "held-out likelihood", utilizzando nell'algoritmo un parametro c (ovvero il range entro cui valutare il numero di topics ottimale) che varia da 5 a 50.
Nella seguente figura 18 viene mostrato l'output ottenuto applicando l'algoritmo.

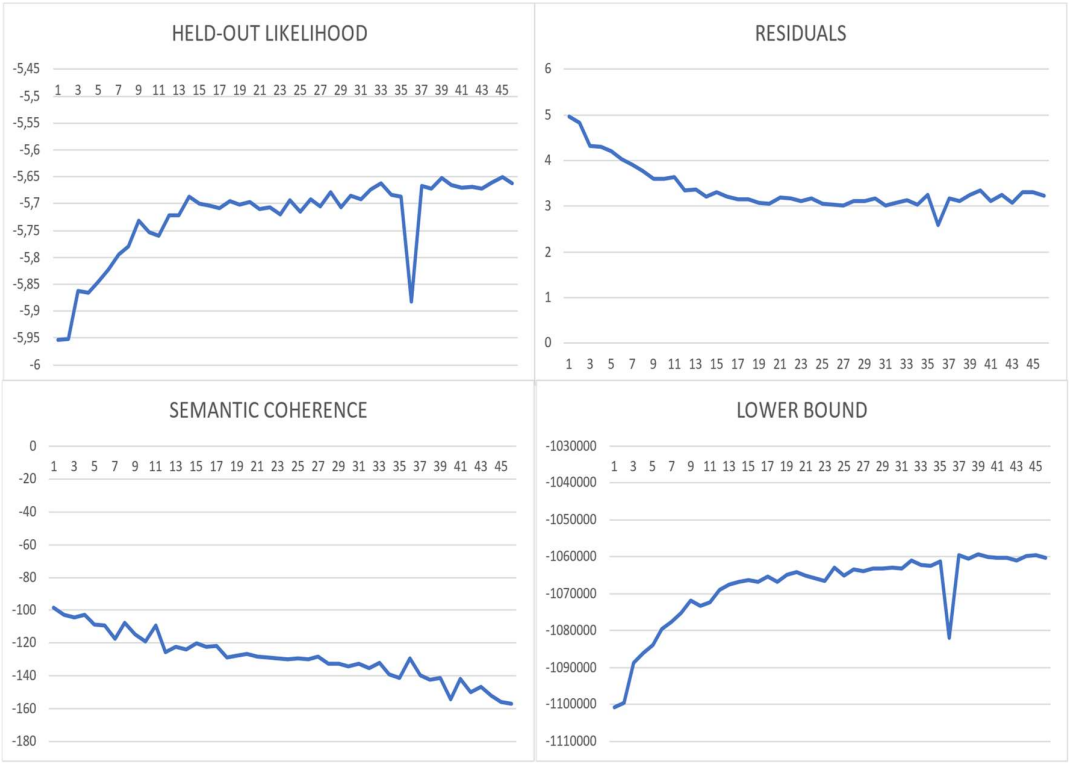


Figura 18 – Output numero ottimale di topics

Dai grafici si evince che il numero ottimale è K=18, pertanto poi lo script dell'algoritmo estrarrà 18 topics.
Anche in questo caso verrà poi valutato se questo numero andrà bene anche per il topic modeling semi-supervisionato oppure se si dovrà intervenire e modificarlo.

Ottenuto questo valore, si può iniziare la vera e propria applicazione dello STM sulle recensioni pre-processate utilizzando il K appena individuato.
L'output che viene restituito è composto dal topical content (φ) e dalla topical prevalence (ϑ), quest'ultima ci servirà nell'importante fase di validazione dei risultati, necessaria per ottenere valori che servono per realizzare la valutazione del target di questa tesi.

Prima di procedere con quell'ultimo passaggio, però, è necessario attuare la fase di labelling, ovvero bisogna identificare i topics ed applicargli manualmente delle etichette semantiche, una per ogni K trovato.

In questa fase è stato problematico effettuare l'associazione tra i concetti individuati dall'algoritmo e l'etichetta, in quanto il rischio è quello di imbattersi in topics confusi e non particolarmente riconoscibili.
Questo problema è stato risolto assegnando dei labels più generici (questo fatto fa pensare che molto probabilmente quando verrà applicato il topic modeling semi-supervisionato verrà rimosso qualche topics, perché essendo che attraverso i seeds si aiuta l'algoritmo, gli argomenti più generici dovrebbero scomparire o comunque essere accorpati con altri).

Infine, si può procedere con la validazione dei risultati seguendo i passaggi descritti più dettagliatamente nel paragrafo 3.6 e qui sotto riassunti:

1. Estrazione di un campione casuale dal database ripulito con le rispettive topical prevalence;
2. Assegnazione umana delle recensioni casuali a uno o più topics;
3. Calcolo della soglia dinamica;
4. Confronto tra la probabilità di ciascuna recensione di appartenere ad ogni topic e la soglia dinamica;
5. Creazione della matrice di confusione per ogni recensione (true positive, false positive, false negative e true negative);
6. Calcolo degli indicatori utilizzando i valori aggregati della matrice di confusione.

Nella seguente tabella 15 vengono riportati gli output conseguenti all'applicazione delle fasi citate precedentemente, ovvero le keywords dall'algoritmo STM, le etichette (label) dal labelling e l'aggiunta di una descrizione sull'argomento trattato da ogni topic ottenuta manualmente dopo aver analizzato lo stesso.

N°	Keywords	Label
1	Highest Prob: awesom, purchas, last, devic, day, expect, satisfi, deliveri, like, deliv FREX: deliveri, awesom, satisfi, deliv, like, arriv, purchas, packag, seal, pack	**Delivery**
	Descrizione: il topic si riferisce all'ultimo step relativo all'acquisto, ovvero la procedura di trasporto e consegna. In aggiunta, racchiude anche i riscontri degli utenti relativi proprio a quest'ultimo passaggio.	
2	Highest Prob: excel, take, photo, superb, load, sound, option, usual, top, especi FREX: excel, load, superb, photo, mark, take, usual, smart, handl, clariti	**Photo Quality**
	Descrizione: il topic rimanda alla qualità delle fotografie memorizzate e visualizzate sul dispositivo. In particolare, parla della nitidezza e dei colori delle immagini.	
3	Highest Prob: video, pictur, light, definit, front, weight, mode, clear, watch, imag FREX: video, imag, watch, youtub, light, pic, pictur, photographi, captur, edit	**Graphic Quality**
	Descrizione: il topic parla di tutto ciò che può riguardare video, foto, luminosità e colori dello schermo.	
4	Highest Prob: price, money, worth, buy, rang, total, wast, expens, india, lot FREX: money, worth, price, wast, total, afford, job, worthi, expens, penni	**Price**
	Descrizione: il topic è relativo al prezzo d'acquisto del device ed alle offerte connesse.	
5	Highest Prob: app, call, issu, sim, screen, card, tri, cant, updat, time FREX: sim, app, card, call, notif, hear, slot, text, wifi, messag	**Connectivity**
	Descrizione: il topic si riferisce a tutto ciò che può riguardare il mondo della connettività, dall'utilizzo della SIMcard e dell'operatore telefonico fino alla qualità delle chiamate.	
6	Highest Prob: mobil, compar, bit, smartphon, around, point, flagship, cost, low, that FREX: mobil, smartphon, bit, compar, budget, class, point, white, around, spec	**Comparison**

	Descrizione: il topic si concentra su dei possibili confronti con altri devices simili, sotto vari aspetti. Di conseguenza, si riferisce anche a vantaggi e svantaggi d'acquisto.	
7	Highest Prob: now, month, problem, switch, sinc, love, use, anoth, hope, till FREX: month, love, switch, till, hang, now, sinc, hope, gift, daughter	**Buy Reason**
	Descrizione: il topic fornisce informazioni sul perché si è deciso di acquistare lo smartphone in oggetto. Da questo deriva anche una forte componente relativa ai periodi durante l'anno (festività, periodi lavorativi/scolastici e life-cicle del dispositivo sostituito).	
8	Highest Prob: batteri, qualiti, life, fast, perform, display, happi, super, overal, smooth FREX: life, batteri, design, happi, build, qualiti, super, beauti, fast, perform	**Performance**
	Descrizione: il topic rimanda a qualsiasi aspetto relativo alle performance, alla qualità ed all'esperienza di utilizzo del dispositivo.	
9	Highest Prob: camera, use, buy, time, perfect, year, come, thank, brand, need FREX: camera, thank, use, wow, fantast, beast, wonder, brand, perfect, own	**Camera**
	Descrizione: il topic rimanda alla qualità della camera presente all'interno del dispositivo. In particolare, parla della nitidezza e della risoluzione delle fotografie.	
10	Highest Prob: screen, fingerprint, display, sensor, reader, finger, slow, print, rate, pretti FREX: fingerprint, print, reader, sensor, finger, notch, bright, lcd, scanner, amol	**Display**
	Descrizione: il topic parla di tutto ciò che concerne il display, il touch-screen ed i sensori digitali.	
11	Highest Prob: far, face, charg, charger, case, box, quick, adapt, recognit, glass FREX: far, cabl, recognit, addit, adapt, headphon, usb, protector, plug, yet	**Case**
	Descrizione: il topic analizza tutto ciò che concerne la parte hardware del dispositivo, dalla custodia agli ingressi USB.	
12	Highest Prob: valu, deal, offer, galaxi, pay, version, seri, wish, phone, choic FREX: valu, deal, version, trade, choic, pay, line, die, hai, okay	**Version**

	Descrizione: il topic analizza la specifica versione acquistata all'interno della gamma (ad esempio le versioni "Pro" oppure "Galaxy"), suggerendone confronti e paragoni.	
13	Highest Prob: better, upgrad, old, differ, hand, plus, big, year, feel, way FREX: upgrad, better, hand, max, differ, faster, glad, smaller, old, notic	**Upgrade**
	Descrizione: il topic richiama a tutto ciò che concerne gli aggiornamenti del dispositivo, suggerendone confronti con le versioni precedenti (aggiornamenti software).	
14	Highest Prob: phone, new, featur, look, size, simpli, learn, bought, complaint, enjoy FREX: phone, simpli, new, learn, negat, featur, look, navig, enjoy, world	**Usability**
	Descrizione: il topic si riferisce all'esperienza d'utilizzo, intesa come facilità di apprendimento, semplicità d'utilizzo ed ergonomia.	
15	Highest Prob: work, littl, color, get, size, fine, small, must, colour, fit FREX: must, blue, fine, color, button, work, fit, small, littl, que	**Design**
	Descrizione: il topic analizza il colore, la dimensione e tutto ciò che può riguardare la forma dello smartphone.	
16	Highest Prob: charg, heat, game, charger, oneplus, usag, hour, issu, con, support FREX: con, usag, pros, game, watt, hrs, oneplus, secur, heat, band	**Battery**
	Descrizione: il topic si concentra sulla capacità di ricarica del dispositivo, parlando di aspetti tecnici (come Ampere e Watt), calore generato e tempistiche.	
17	Highest Prob: buy, servic, phone, replac, custom, help, store, issu, receiv, bought FREX: servic, replac, defect, custom, order, verizon, repair, transfer, warranti, center	**Assistance**
	Descrizione: il topic si riferisce all'attività d'assistenza. In particolare, parla di esperienze di guasto, cause di riconsegna, sostituzione e garanzia.	
18	Highest Prob: experi, recommend, disappoint, long, bought, sure, find, enough, friend, high FREX: high, recommend, disappoint, friend, everyon, experi, pixel, trust, long, flaw	**Feedback**

	Descrizione: il topic è dedicato ai feedback generati, alle raccomandazioni ed ai consigli che gli utenti darebbero ad altri users (come parenti o amici).

Tabella 15 - Label, keywords e descrizione dei topic degli smartphone

I topics riportati nella precedente tabella corrispondono alle determinanti di qualità riferite alle tipologie di smartphone scelte, le quali sono state estratte dalle recensioni online realizzate dai consumatori.

In seguito si può procedere con la fase di validazione, di cui verranno esposti i risultati ottenuti nelle seguenti tabelle 16 e 17:

		Assegnazione umana del topic	
		Assegnazione al topic T_i	Non-assegnazione al topic T_i
Assegnazione automatica del topic	Assegnazione al topic T_i	**True positive** TP=141	**False positive** FP=40
	Non-assegnazione al topic T_i	**False negative** FN=83	**True negative** TN=1536

Tabella 16 – Output matrice di confusione smartphone (topic modeling non supervisionato)

Indicatore	Risultato ottenuto	Valore target
Accuracy	0,9317	>0,95
Precision	0,7790	>0,70
Recall	0,6295	>0,70
F_1 score	0,6963	>0,70
Fall-out	0,0254	<0,05
Miss rate	0,3705	<0,20
Specifity	0,9746	>0,90
Negative predective value	0,9487	>0,90
False omission rate	0,0513	<0,05
False discovery rate	0,2210	<0,05

Tabella 17 – Output indicatori di performance smartphone (topic modeling non supervisionato)

La tabella degli indicatori mostra che le performance dell'output dell'algoritmo di topic modeling non supervisionato soddisfano abbastanza i valori target ai quali si puntava.

Gli indicatori precision, fall-out, specifity e negative predective value rispettano in pieno l'obiettivo da raggiungere, mentre sono stati considerati buoni anche accuracy, F1 score e false omission rate perché si avvicinano molto al target, e considerato che quest'ultimo è semplicemente un valore di riferimento non per forza i risultati lo devono rispettare, ma è sufficiente che si avvicinino.

Al contrario, ci sono tre indicatori che non rientrano nell'obiettivo e sono:

1. Recall: significa che il numero di true positive non è abbastanza ampio in confronto ai false negative, con una conseguente elevata quota di recensioni che secondo l'umano riguardano un determinato topic mentre per l'algoritmo no. Questo non va bene perché ovviamente dovrebbero essere d'accordo le due varianti;

2. Miss rate: vuol dire che l'algoritmo non è molto propenso a riuscire ad identificare un topic effettivamente presente;
3. False discovery rate: significa che il numero di falsi positivi è rilevante, pertanto sono stati identificati molti topics non corretti, e quindi un numero di true positive troppo basso.

4.2.4 Topic modeling semi-supervisionato

Una volta ottenuti i valori benchmark dal modello non supervisionato si può procedere con l'obiettivo della tesi, ovvero quello di realizzare un modello di topic modeling semi-supervisionato, in questo caso applicato al database precedentemente realizzato riferito agli smartphone.
Anche in questo caso viene riapplicato l'algoritmo presente nello studio "Seeded-LDA for Topic Modeling", con le modifiche già utilizzate in precedenza.
I dettagli dei passaggi dell'applicazione del metodo possono essere trovati nel paragrafo 3.5.

Una fase fondamentale però è quella che riguarda i seed dei topics necessari per il perfezionamento dell'algoritmo semi-supervisionato rispetto a quello non supervisionato, ed anche in questo caso verranno determinati manualmente tramite le parole che caratterizzano maggiormente i K argomenti trovati con il precedente modello non supervisionato.
Inoltre, verrà ripetuta l'analisi sui K topics con la quale si verificherà se sono tutti necessari oppure se ne esistono alcuni che non sono particolarmente utili al caso di studio oppure se ne sono presenti altri che trattano argomenti simili o con sfumature leggermente diverse che verranno accorpati in un unico topic.

Nel caso specifico degli smartphone è stato ritenuto utile eliminare i seguenti topics in quanto forniscono informazioni troppo generiche e non utili allo studio:

- Version;
- Feedback.

In aggiunta sono stati rimossi due ulteriori argomenti perché contengono informazioni generiche che possono essere ricavate da altri:

- Comparison;
- Buy reason.

Infatti i contenuti di entrambi è stato ritenuto che possano essere individuati dai topics “Photo quality”, “Graphic quality”, “Price”, “Performance”, “Camera”, “Display”, “Case”, “Usability”, “Design” e “Battery”.
Allo stesso tempo è stato considerato che alcuni topics possano essere accorpati in macroargomenti in quanto sottoinsieme di topics simili, ed essi sono:

- Photo quality e Camera: uniti per rappresentare il macroargomento di nome “Camera quality”;
- Graphic quality e Display: accorpati per formare “Display quality”.

Infine il topic “Case” è stato incluso in “Design”, perché il primo può essere considerato una caratteristica specifica del design di uno smartphone.

Le attività appena descritte vengono riassunte nella seguente figura 19:

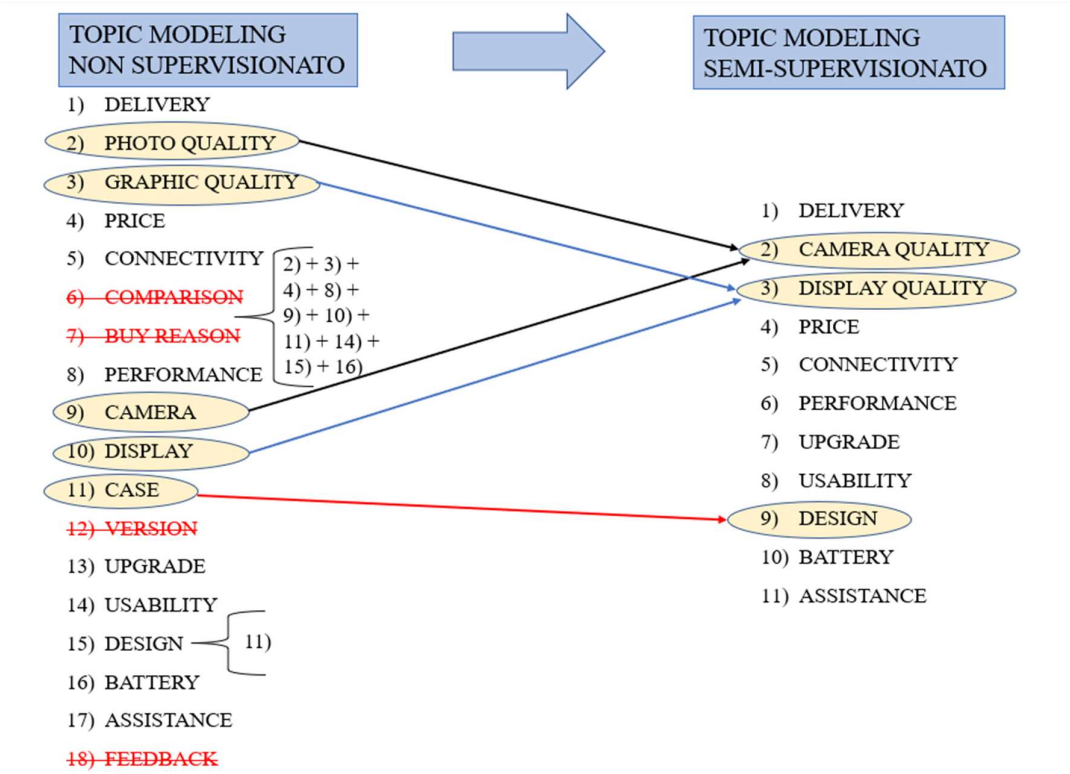


Figura 19 - Raggruppamento topic (la parentesi graffa stabilisce che l'argomento di riferimento è troppo generico e di conseguenza può essere rimosso in quanto i casi specifici sono dettagliati nei topic contenuti nella parentesi)

In conseguenza all'aver individuato i topics da utilizzare nel topic modeling semi-supervisionato, si devono individuare i seed di ciascuna determinante di qualità.

Nella seguente tabella 18 verrà esposto l'output ottenuto:

TOPIC	**SEED**
DELIVERY	Day, deliveri, deliv, arriv, packag, seal, pack.
CAMERA QUALITY	Photo, superb, sound, top, smart, clarity, camera, perfect, wow, fantast, wonder.
DISPLAY QUALITY	Video, pictur, light, definit, watch, imag, pic, photographi.
PRICE	Price, money, worth, buy, total, expens, afford, whorti.
CONNECTIVITY	Sim, card, app, call, notif, hear, text, wifi, messag.
PERFORMANCE	Batteri, qualiti, fast, perform, super.
UPGRADE	Upgrad, old, better, faster
USABILITY	Phone, size, simpli, learn, enjoy, featur.
DESIGN	Littl, color, size, fine, small, colour, fit, blue, button, face, case, glass, cabl, adapt, usb, protector.
BATTERY	Charg, heat, charger, watt, hrs.
ASSISTANCE	Servic, replac, custom, help, issu, repair, warranti, center.

Tabella 18 – Topics e seed riferiti agli smartphone (topic modeling semi-supervisionato)

I seed stabiliti vengono poi inseriti nel codice dell'algoritmo di topic modeling semi-supervisionato per ciascun topic.

Una volta che sono state date in input tutte le informazioni necessarie all'algoritmo e viene eseguito il run, vengono restituiti in output i parametri di topical prevalence (ϑ) e topical content (φ), dove per la fase di validazione vengono considerati solo i primi, i quali vengono salvati in un apposito file .csv e successivamente rielaborati in un formato adatto a lavorare sul software Excel.

Come già accaduto per il caso Uber si noterà, nel prossimo paragrafo di validazione e analisi dei risultati, che l'output ottenuto dal processo di topic modeling semi-supervisionato crea un ulteriore argomento chiamato "other", perché anche in questo caso i vari topics hanno del rumore al loro interno e quindi l'algoritmo che li processa ne crea uno nuovo in cui racchiude i concetti non legati ai topics stabiliti in precedenza.

I valori di topical prevalence (ϑ) ottenuti come output dell'algoritmo di topic modeling semi-supervisionato, permettono di procedere con l'importante fase di validazione dei risultati.
Essa viene realizzata sempre utilizzando il metodo proposto da Barravecchia F. *et al.,* in modo tale da rimanere coerente con le precedenti validazioni, e per i dettagli della procedura si rimanda al paragrafo 3.6 della presente tesi.
I topics utilizzati in questa fase sono quelli riportati in tabella 18.
Si procede dunque al confronto tra i topics individuati automaticamente dall'algoritmo e quelli trovati in seguito ad un'analisi umana, e questo permette di ricavare la matrice di confusione con i rispettivi valori totali di true positive, false positive, false negative e true negative (tabella 19).

		Assegnazione umana del topic	
		Assegnazione al topic T_i	Non-assegnazione al topic T_i
Assegnazione automatica del topic	Assegnazione al topic T_i	**True positive** TP=67	**False positive** FP=83
	Non-assegnazione al topic T_i	**False negative** FN=84	**True negative** TN=966

Tabella 19 – Output matrice di confusione Smartphone (topic modeling semi-supervisionato)

In seguito ai valori sopracitati vengono calcolati gli indicatori utilizzati poi per realizzare il confronto tra il topic modeling semi-supervisionato e non.
I risultati ottenuti vengono riportati nella seguente tabella 20:

Indicatore	**Risultato ottenuto**	**Valore target**
Accuracy	0,8608	>0,95
Precision	0,4467	>0,70
Recall	0,4437	>0,70
F_1 score	0,4452	>0,70
Fall-out	0,0791	<0,05
Miss rate	0,5563	<0,20
Specifity	0,9209	>0,90
Negative predective value	0,9200	>0,90
False omission rate	0,0800	<0,05
False discovery rate	0,5533	<0,05

Tabella 20 – Output indicatori di performance smartphone (topic modeling semi-supervisionato)

Da quest'ultima tabella si evince come il modello semi-supervisionato non garantisca delle buone performance rispetto ai valori target presi come riferimento per la bontà dei risultati.
A questo proposito si nota che ben 6 dei 10 indicatori calcolati si discostano molto dall'obiettivo, e allo stesso tempo ulteriori due si avvicinano ma non lo rispettano a pieno (nonostante questo, vengono comunque considerati buoni).

Gli indicatori che non rispettano l'obiettivo sono:

1. Accuracy: nonostante la distanza dal valore target non sia esagerata, è comunque stato considerato che la performance non sia di buon livello, e quindi che l'algoritmo non è molto efficace nelle previsioni corrette. Questo è dovuto dal fatto che il numero di true positive è basso;
2. Precision: questo porta a pensare che molto spesso non è corretta l'individuazione di un certo topic riferito ad una determinata recensione, nonostante l'algoritmo e l'uomo siano d'accordo che invece siano legati;
3. Recall: significa che non sono state recuperate molte istanze rilevanti rispetto al totale, cioè il numero di true positive non è sufficientemente elevato rispetto alla quantità di false negative, con una conseguente maggior quota di recensioni che secondo l'umano riguardano un certo topic mentre per l'algoritmo di topic modeling no. Ovviamente questo non va bene perché in teoria le due versioni dovrebbero concordare;
4. F1 Score: essendo dipendente da precision e recall, anche questo indicatore non rispetta l'obiettivo, ciò vuol dire che il test non è accurato;
5. Miss rate: ci dice che si ha una bassa propensione dell'algoritmo a identificare un topic effettivamente presente;
6. False discovery rate: il mancato rispetto del valore target vuol dire che il numero di falsi positivi è rilevante, quindi sono stati individuati molti topics errati.

Per quanto riguarda gli indicatori accuracy, precision e recall, essi non rispettano i valori target perché in generale sono stati trovati pochi true positive, ovvero che

sia l'algoritmo che l'analisi umana siano d'accordo nel sostenere che una determinata recensione faccia riferimento ad un certo topic.
Fall-out e False omission rate sono stati considerati accettabili nonostante non rispettino a pieno il target, perché la differenza tra il valore puntuale e l'obiettivo non è molto elevata, tenendo sempre presente che questi ultimi sono valori di riferimento e non vincolano al raggiungimento stretto del traguardo.

In questa fase sono state riscontrate due problematiche che probabilmente hanno portato ad avere basse performance dell'algoritmo di topic modeling semi-supervisionato.
La prima è stata che, molte volte le recensioni degli utenti, nonostante fossero lunghe, racchiudessero poche informazioni ma di tanti topics diversi, ad esempio di uno smartphone un consumatore spiega allo stesso tempo la qualità della camera, lo stato della batteria e il design del dispositivo, e questo ha portato l'algoritmo ad individuare pochi argomenti o addirittura solamente uno.
Un secondo problema è stato riscontrato con il topic "other" generato automaticamente dal modello, perché a differenza del precedente caso di studio riferito a Uber, dove non sono state notate particolari differenze tra la decisione umana e quella automatica, nel caso specifico degli smartphone è capitato diverse volte che a livello di decisione umana venivano individuati diversi argomenti riferiti a una specifica recensione, e quindi questo faceva escludere il topic other, mentre per l'algoritmo era comunque presente.

4.2.5 Confronto indicatori topic modeling non supervisionato e semi-supervisionato

Grazie al calcolo degli indicatori ottenuti dall'analisi degli output del topic modeling non supervisionato e semi-supervisionato, si può effettuare il confronto tra i due modelli e verificare quale tra i due garantiscono migliori performance, e quindi stabilire se alle aziende conviene applicare uno o l'altro per stabilire quali sono le determinanti di qualità del prodotto o servizio che offrono.

Come nel precedente caso di studio, si parte analizzando le differenze tra le quantità totali di true positive, false positive, false negative e true negative (presenti in tabella 21).

	Non supervisionato	**Semi-supervisionato**
TP	141	67
FP	40	83
FN	83	84
TN	1536	966

Tabella 21 – Quantità di true positive, false positive, false negative e true negative per tipologia di topic modeling

Anche in questo caso tra il modello non supervisionato e quello semi-supervisionato cambia il numero di topics analizzati, infatti nella prima tipologia si hanno 18 argomenti mentre nella seconda 12, e tenendo sempre come riferimento le 100 recensioni casuali estratte, questo fa cambiare la quantità dei valori analizzati, rispettivamente 1800 e 1200, con la conseguenza di non poter utilizzare nel confronto le semplici quantità ma solamente attraverso la loro presenza percentuale, che viene calcolata e riportata nella seguente tabella 22.

	Non supervisionato	**Semi-supervisionato**
TP	7,83%	5,58%
FP	2,22%	6,92%
FN	4,61%	7,00%
TN	85,33%	80,50%

Tabella 22 – Percentuale di true positive, false negative e true negative per tipologia di topic modeling

Dal confronto che si evince dai valori percentuali, si nota subito come utilizzando un algoritmo semi-supervisionato diminuisce l'accordo tra le assegnazioni automatiche e quelle umane.

Infatti i valori di accordo sia nel caso positivo che in quello negativo (ovvero true positive e true negative) diminuiscono percentualmente del 2,25% e del 4,83% rispettivamente.

Allo stesso tempo aumentano i valori di disaccordo, quindi crescono del 4,7% i false positive e del 2,39% i false negative, confermando quindi l'idea che si aveva quando sono stati calcolati gli indicatori del modello semi-supervisionato, ovvero le performance di quest'ultimo non convincono, anzi peggiorano nonostante vengano fornite delle parole d'ancoraggio all'algoritmo.

Per confermare quest'ultima affermazione, in tabella 23 vengono riportati, uno di seguito all'altro, i valori degli indicatori ottenuti con i due modelli, in modo tale da poter scorgere meglio le differenze tra essi.

Indicatore	**Non supervisionato**	**Semi-supervisionato**
Accuracy	0,9317	0,8608
Precision	0,7790	0,4467
Recall	0,6295	0,4437
F_1 score	0,6963	0,4452
Fall-out	0,0254	0,0791
Miss rate	0,3705	0,5563
Specifity	0,9746	0,9209
Negative predective value	0,9487	0,9200
False omission rate	0,0513	0,0800
False discovery rate	0,2210	0,5533

Tabella 23 – Valori degli indicatori per tipologia di topic modeling

Dalla tabella precedente si evince subito come tutti gli indicatori derivati dal topic modeling semi-supervisionato siano peggiori rispetto alla controparte non supervisionata.
Infatti, tutti quegli indicatori i cui output devono avvicinarsi all'unità per garantire performance migliori, sono più elevati nel caso del modello non supervisionato, ed essi sono:

- Accuracy;
- Precision;
- Recall;
- F_1 score;
- Specificity;
- Negative predective value.

Contemporaneamente, anche tutti gli indicatori che devono avvicinarsi a zero per delle buone prestazioni, risultano essere più bassi nel caso di topic modeling non supervisionato.
Questi indicatori risultano essere:

- Fall-out;
- Miss rate;
- False omission rate;
- False discovery rate.

In conclusione del caso di studio degli smartphone, si può affermare che il topic modeling non supervisionato garantisce un output più performante e adatto alle aziende che producono smartphone, nonostante sia una metodologia più grezza e non venga dato alcun aiuto in input all'algoritmo.

5 DISCUSSIONE E CONCLUSIONI

La mole di dati disponibile gratuitamente online su qualsiasi prodotto o servizio presente sul mercato non può rimanere inutilizzata, infatti se sfruttata adeguatamente possono trarne vantaggi sia i clienti, ottenendo informazioni utili tramite recensioni e quindi potendo prendere decisioni più consapevoli su un eventuale acquisto o utilizzo di un prodotto/servizio, e allo stesso tempo anche le aziende per migliorarsi e analizzare i competitors.

Gli algoritmi di topic modeling riescono a sfruttare questi dati per ottenere informazioni utili nei vari contesti, ma esistendo diverse tipologie, bisogna anche capire qual è la più adeguata e performante.

L'obiettivo del presente lavoro, quindi, era quello di stabilire se l'utilizzo di un algoritmo di topic modeling semi-supervisionato permettesse di raggiungere performance migliori rispetto a quelle conseguibili con una versione non supervisionata nell'ottenimento delle determinanti latenti di qualità di un prodotto o servizio.

Applicando una stessa metodologia per le due diverse versioni di topic modeling e confrontando gli output ottenuti (in particolar modo degli indicatori di performance), è stato possibile giungere al risultato che, al contrario di quanto preventivato, attraverso l'algoritmo semi-supervisionato si ottengono performance qualitativamente inferiori per entrambi i casi di studio affrontati.

In particolare, tutti gli indicatori di performance provenienti dallo studio effettuato da Barravecchia F. *et al.* (2021) risultano peggiorare (come riportato nelle tabelle 13 e 23, rispettivamente nei paragrafi 4.1.5 e 4.2.5).

Le cause del peggioramento potrebbero essere diverse, ad esempio quelle individuate sono:

1. Topic "other": in entrambi i casi di studio, il topic modeling semi-supervisionato ha creato in automatico un topic generico definito "other", che raccoglie tutti i rumori contenuti nelle recensioni.

Questo potrebbe quindi causare degli errori nei riconoscimenti automatici degli argomenti di determinate recensioni, e conseguentemente performance non accettabili;

2. Assegnazione di seed inadeguati: potrebbero essere state identificate delle parole d'ancoraggio non troppo inerenti, e questo al posto che facilitare l'algoritmo a riconoscere e ad assegnare dei topics alle recensioni estratte, lo ha messo in difficoltà.
3. Eliminazione o accorpamento di topics: nella fase di labelling del processo semi-supervisionato sono stati accorpati topics riguardanti argomenti simili ma con sfumature leggermente diverse, ed eliminati altri che invece erano troppo generici o che non sono stati ritenuti utili all'analisi delle determinanti di qualità.

   Questo fatto però, potrebbe aver portato a una minor propensione dell'algoritmo ad assegnare correttamente i topics alle recensioni;
4. Recensioni contenenti informazioni sintetiche su tanti argomenti diversi: soprattutto nel caso di studio riferito agli smartphone, è stato notato che sono presenti diverse recensioni che parlano di moltissimi topics differenti in maniera molto sintetica.

   In questi casi l'algoritmo non è riuscito ad individuarli e ad assegnarli tutti, e conseguentemente le performance peggiorano;
5. Modello di topic modeling semi-supervisionato non adeguato o da perfezionare: esistono diversi studi e metodologie che adottano questo modello in maniera differente.

   Potrebbe essere che utilizzare algoritmi facenti riferimento ad altri ricercatori e quindi che forniscono librerie diverse o addirittura usano un altro linguaggio di programmazione garantiscano output migliori e più performanti (in questo elaborato è stato preso come riferimento R, mentre durante la fase di ricerca di papers e applicazioni sono stati trovati degli esempi in Python).

Per queste ragioni potrebbero essere importanti lavori futuri che confermino o confutino gli output di questo elaborato.
Essi potrebbero approfondire quanto già svolto oppure realizzare nuovi studi, ad esempio:

1. Confrontare gli output di due algoritmi di topic modeling semi-supervisionato appartenenti a studi diversi utilizzando lo stesso linguaggio di programmazione;
2. Confrontare gli output di due algoritmi di topic modeling semi-supervisionato utilizzando due diversi linguaggi di programmazione;
3. Cambiare il numero ottimale di topics e di conseguenza tutti gli argomenti e i seed del presente studio, in modo tale da confermare che un modello non supervisionato garantisca performance migliori.

In generale, questo studio potrebbe essere preso e analizzato da aziende che desiderano implementare processi innovativi di analisi dei dati, ma che non conoscono le differenze tra il topic modeling non supervisionato e quello semi-supervisionato.
Infatti troverebbero le spiegazioni di entrambi i modelli e, soprattutto, basandosi sui due casi di studio realizzati e i loro rispettivi output, potrebbero convincersi che non vale la pena investire tempo e risorse nell'utilizzo di versioni che sulla carta dovrebbero essere maggiormente performanti e poi nella realtà dei fatti non lo sono (come dimostrato nella presente tesi per quanto riguarda l'algoritmo semi-supervisionato di topic modeling).

In conclusione dell'analisi, dai risultati ottenuti per ciascun caso di studio, si denota come l'applicazione dell'algoritmo di topic modeling non supervisionato, e quindi più grezzo, garantisce una miglior individuazione delle determinanti di qualità del proprio prodotto e servizio, di conseguenza non è necessario l'utilizzo di metodologie più sofisticate ma meno performanti.
Questa ricerca è importante perché il risultato che si ottiene può avere un impatto più o meno grande nell'ambito della qualità e dell'analisi dei dati, dal momento

che, dal punto di vista della ricerca, i ricercatori potrebbero concentrarsi maggiormente sul continuo miglioramento degli algoritmi non supervisionati già realizzati visto che la controparte non fornisce risultati adeguati.

Questo studio è vantaggioso anche dal punto di vista delle aziende, le quali non devono per forza imparare nuovi processi (che ovviamente comporterebbero un dispendio di risorse che potrebbero essere reindirizzate su altre attività più redditizie), ma utilizzare quelli già in essere in quanto il topic modeling non supervisionato è già molto sfruttato dalle imprese, per ottenere performance elevate utili per un continuo miglioramento del proprio prodotto o servizio e mantenere un vantaggio competitivo sulla concorrenza.

BIBLIOGRAFIA:

Barravecchia Federico, Mastrogiacomo Luca, Franceschini Fiorenzo. "Categorizing Quality Determinants in Mining User-Generated Contents*". MDPI Sustainability Vol. 12, 9944 (2020).*

Barravecchia Federico, Mastrogiacomo Luca, Franceschini Fiorenzo. "Digital voice-of-customer processing by topic modeling algorithms: insights to validate empirical results". *International Journal of Quality & Reliability Management Vol. 39, No. 6 (2022): pp. 1453-1470.*

Benoit Kenneth, Watanabe Kohei, Wang Haiyan, Nulty Paul, Obeng Adam, Müller Stefan, Matsuo Akitaka, Lowe William, Müller Christian. "Quantitative Analysis of Textual Data - Package 'quanteda'". *CRAN (13 Ottobre 2022, version 3.2.3).*

Blei David M., Ng Andrew Y., Jordan Michael I.. "Latent Dirichlet Allocation". *Journal of Machine Learning Research Vol. 3 (2003): pp. 993-1022.*

Bo Zhao. "Web Scraping". *Springer International Publishing AG (outside the USA). L.A. Schintler, C.L. McNeely (eds.), Encyclopedia of Big Data (2017).*

Brown Patrick G.. "QFD: echoing the voice of the customer". *AT&T technical journal (March/April 1991).*

Diouf Rabiyatou, Birregah Babiga, Sarr Edouard Ngor, Bousso Mamadou, Sall Ousmane, Mbaye Sény Ndiaye. "Web Scraping: State-of-the-Art and Areas of Application", *2019 IEEE International Conference on Big Data (2019).*

Franceschini Fiorenzo. "Dai prodotti ai servizi – Le nuove frontiere per la misura della qualità", *UTET (2001, 1° edizione).*

Glez-Peña Daniel, Lourenço Anália, López-Fernández Hugo, Reboiro-Jato Miguel, Fdez-Riverola Florentino. "Web scraping technologies in an API world". *Briefings in bioinformatics Vol. 15, No. 5 (2013): pp. 788-797.*

Griffin Abbie, Hauser John. “The Voice of the Customer”. *Wiley International Encyclopedia of Marketing, John Wiley & Sons Ltd (1991).*

Jelodar Hamed, Wang Yongli, Yuan Chi, Feng Xia, Jiang Xiahui, Li Yanchao, Zhao Liang. “Latent Dirichlet Allocation (LDA) and topic modeling: models, applications, a survey”. *Multimedia Tools and Applications Vol. 78, No. 11 (2019): pp. 15169-15211.*

Mazur Glenn. “Voice of the customer (define): QFD to define value”. *Annual Quality Congress Proceedings, Milwaukee Vol. 57 (2003): pp. 151-157.*

Roberts Margaret E., Stewart Brandon M., Tingley Dustin. “STM: An R Package for Structural Topic Models”. *Journal of Statistical Software Vol. 91, No. 2 (2019): pp. 1-40.*

Roberts Margaret E., Stewart Brandon M., Tingley Dustin, Lucas Christopher, Leder-Luis Jetson, Gadarian Shana Kushner, Albertson Bethany, Rand David G.. “Structural Topic Models for Open-Ended Survey Responses”. *American Journal of Political Science, Vol. 58, No. 4 (2014): pp. 1064-1082.*

Sbalchiero Stefano, Eder Maciej. “Topic modeling, long texts and the best number of topics. Some Problems and solutions”. *Springer Nature B.V. 2020, Quality & Quantity. Vol. 54 (2020): pp. 1095–1108.*

Sutanto Agus, Yuliandra Berry, Tjahjono Benny, Hadiguna Rika Aampuh. “Product-service system design concept development based on product and service integration”. *J. Design Research, Vol. 13, No. 1 (2015).*

Tukker Arnold, Tischner Ursula. “Product-services as research field: past, present and future. Reflections from a decade of research”. *Journal of Cleaner Production Vol. 14 (2006): pp. 1552-1556.*

Watanabe Kohei, Zhou Yuan. “Theory-Driven Analysis of Large Corpora: Semisupervised Topic Classification of the UN Speeches”. *Social Science Computer Review 2022 Vol.40, No.2: pp. 346-366.*

SITOGRAFIA:

https://www.agnesevardanega.eu/wiki/r/gestione_dei_dati/importare_i_dati

https://cran.r-project.org/web/packages/seededlda/seededlda.pdf

https://www.galganogroup.com/wp-content/uploads/2017/03/14ArfLippiMethodo-maggio.pdf (2014)

https://github.com/dondealban/learning-stm

https://www.html.it/guide/guida-r/

https://medium.com/pew-research-center-decoded/overcoming-the-limitations-of-topic-models-with-a-semi-supervised-approach-b947374e0455

https://quanteda.io/reference/dfm_select.html

https://www.rdocumentation.org/packages/utils/versions/3.6.2/topics/head

https://www.r-project.org/about.html

https://www.tiobe.com/tiobe-index/

APPENDICE: guida all'utilizzo di Seeded-LDA

All'interno della presente appendice verranno spiegati tutti i vari comandi e i procedimenti che sono stati utilizzati nell'applicazione del topic modeling semi-supervisionato, che sono:

1. "require(seededlda)" e "require(quanteda)": sono pacchetti delle librerie che devono essere caricati prima di poter implementare qualsiasi comando. In particolare, il comando è "require()" mentre "seededlda" e "quanteda" sono le librerie;
2. "read.csv2()": successivamente è necessario immettere il database all'interno del programma RStudio, in modo tale da poter eseguire le successive operazioni di pulitura e di topic modeling. Per effettuare ciò in questa tesi viene fatta una modifica rispetto al paper utilizzato come riferimento, infatti viene impiegato il comando "read.csv2()" perché il database contenente tutte le recensioni era stato preventivamente salvato in formato .csv, e questa linea di codice permette proprio di fornire in input il file desiderato se è nel formato numerico italiano, ovvero legge i file con ";" come separatore di campo e "," come separatore decimale;
3. "head()": questo comando restituisce la prima o l'ultima parte di un vettore, una matrice, una tabella, un data frame o una funzione.
   È utile per creare un corpo del dataset sul quale verranno applicate le funzioni presenti nei prossimi punti che permettono una pulizia di quest'ultimo;
4. "tokens()": è un comando che prende come riferimento le recensioni del corpo creato precedentemente ed elimina tutte le informazioni testuali non necessarie alla comprensione del testo, come ad esempio la punteggiatura, i vari simboli che potrebbero essere presenti, i numeri, gli n-grammi, un po' come fatto per la parte di topic modeling non supervisionato, ma attraverso una diversa linea di codice;

5. “dfm()”: comando che viene attivato sul testo ripulito dalla precedente funzione, il quale permette di creare una matrice sparsa di caratteristiche del documento da un carattere, un corpus, un tokens o anche da un altro oggetto dfm (Kenneth Benoit *et al.*, Quantitative Analysis of Textual Data, Package ‘quanteda’, versione 3.2.3, 13 Ottobre 2022), e comprende due sottocomandi:
    a. “dfm_remove(stopwords(‘en’), min_nchar=2)”: permette di rimuovere delle features al dfm appena estratto (in questo caso specifico le stopwords di lingua inglese e tutte le parole che contengono meno di due caratteri, ma in generale può anche eliminare termini da un dictionary che colui che sta effettuando l’analisi ha creato per elencare le varie parole non utili ai fini dello studio);
    b. “dfm_trim(min_termfreq=0.90,termfreq_type=”quantile”, max_docfreq=0.1, docfreq_type=”prop”)”: permette di tagliare il dfm utilizzando come parametro una soglia di frequenza sia delle parole sia dei documenti e restituisce quindi una matrice di dimensioni ridotte (come riportato dal Benoit K. *et al.*, nella guida sul pacchetto quanteda citata in precedenza).
6. “dictionary()”: è una funzione che permette di creare un dizionario di parole, ovvero un elenco di caratteri denominato con classi speciali, che può derivare da una lista di parole (come in questi casi di studio) oppure può essere importato da una fonte esterna che può avere diversi formati (WordStat, LIWC, Lexicoder v2 e v3, Yoshikoder) (Kenneth Benoit *et al.*, Quantitative Analysis of Textual Data, Package ‘quanteda’, versione 3.2.3, 13 Ottobre 2022).
7. “textmodel_seededlda()”: attraverso questo comando vengono richiamati la matrice dfm e il dizionario creati in precedenza.

In generale, le funzioni di tipo textmodels facenti parte del pacchetto quanteda permettono di avere modelli per il ridimensionamento e la classificazione dei dati testuali.

Nello specifico, "textmodel_seededlda()" implementa il semi-supervised Latent Dirichlet (seeded-LDA) attraverso un codice adottato dalla libreria GibbsLDA++ (Xuan-Hieu Phan, 2007) e permette ai ricercatori di specificare i topics utilizzando il vocabolario dei seed.

Esso restituisce una lista dei parametri del modello, ovvero il theta (ϑ) che rappresenta la distribuzione dei topics nei documenti, phi (φ) cioè la distribuzione delle parole nei topics, e altri parametri poco interessanti nel presente studio (alpha e beta) (Kohei Watanabe *et al.*, Seeded-LDA for Topic Modeling, Package 'seededlda', versione 0.8.2, 14 Ottobre 2022).

8. "write.csv (unlist(slda$theta, file=' ')": questo comando permette di esportare e salvare i valori di topical prevalence (ϑ) in un file in formato .csv.