

POLITECNICO DI TORINO

Collegio di Ingegneria Gestionale

Corso di Laurea Magistrale in Ingegneria Gestionale



Politecnico di Torino

Tesi di Laurea Magistrale

Metodi e strumenti per l'analisi della digital VoC (Voice of Customer)

Relatore: Fiorenzo Franceschini

Candidato: Stefano Coccato

Co-relatore: Federico Barravecchia

Anno accademico 2022-2023

Indice

1	Lista acronimi.....	2
2	Introduzione	3
3	Descrizione del problema e analisi della letteratura	6
3.1	Descrizione del problema	6
3.2	Il text mining	7
3.2.1	Steps del text mining	8
3.2.2	Applicazioni del text mining.....	9
3.3	Scelta del prodotto	11
4	Strumenti utilizzati.....	12
4.1	Structural Topic Modeling.....	12
4.2	Output dello Structural Topic Modeling	12
4.3	Funzionamento dello Structural Topic Modeling.....	13
4.3.1	Raccolta dei dati	14
4.3.2	Pre- processing	14
4.3.3	Identificazione del numero ottimale di topics	16
4.3.4	Applicazione del topic modeling.....	17
4.3.5	Labeling.....	19
4.4	OCTOPARSE-WEB SCRAPING.....	20
4.4.1	Flusso di lavoro di Octoparse.....	21
4.5	RSTUDIO.....	22
4.6	ChatGPT	23
4.6.1	Alla base di ChatGPT.....	24
5	Applicazione degli strumenti utilizzati al caso studio	26
5.1	Raccolta dei dati e pre-processing.....	26
5.2	Indentificazione del numero ottimale di topics e applicazione STM.....	27
6	Metodo proposto per l'estrazione dei requisiti	29
7	Confronto dei risultati ottenuti tramite estrazione manuale, estrazione tramite il metodo delle parole e estrazione tramite ChatGPT.....	41
7.1	Analisi dei risultati: estrazione manuale e metodo delle parole	47
7.2	Estrazione dei requisiti con ChatGPT e confronto con estrazione manuale.....	48
7.2.1	Analisi dei risultati: metodo manuale e ChatGPT.....	50
7.3	Confronto tra ChatGPT e metodo delle parole.....	50
8	Conclusione	54
9	Bibliografia e sitografia	57

1 Lista acronimi

ACRONIMO	DESCRIZIONE
UGC	User Generated Contents
QFD	Quality Function Deployment
NLP	Natural Processing Language
AI	Artificial Intellingence
LDA	Latent Dirichlet Allocation
STM	Structural Topic Modeling
FREX	Frequent Exclusive
GPT	Generative Pre-trained Transformer
VPT	Veri Positivi Totali
FPT	Falsi Positivi Totali
VNT	Veri Negativi Totali
FNT	Falsi Positivi Totali
NPV	Negative Predicted Value
FOR	False Omission Rate
FDR	False Discovery Rate
MVP	Minimum Viable Product

2 Introduzione

La comprensione dei requisiti di un prodotto/servizio è di fondamentale importanza per poter realizzare prodotti/servizi che soddisfano il più possibile i bisogni dei clienti e la soddisfazione dei clienti è un determinante del successo all'interno di un mercato competitivo (Barravecchia, Mastrogiacomo, e Franceschini 2022). Infatti, una maggiore soddisfazione dei clienti può portare ad un aumento della probabilità che questi acquistino, ricomprino e addirittura consiglino il prodotto/servizio ad altre persone. La comprensione dei requisiti di un prodotto/servizio, oltre a servire nello sviluppo di un nuovo prodotto, consente di identificare errori all'interno delle caratteristiche di un prodotto/servizio già sul mercato, migliorare la customer experience e consente anche di effettuare un'analisi della concorrenza rispetto al mercato.

Attualmente, per ottenere i requisiti relativi ad un prodotto/servizio si utilizzano ancora i metodi tradizionali, quali i classici questionari e interviste, nelle loro varie sfaccettature. Tuttavia, al giorno d'oggi, a far concorrenza ai metodi tradizionali c'è la possibilità di sfruttare il materiale disponibile online contenente informazioni sui prodotti/servizi, di cui la maggior parte generato dagli utenti. I dati generati dagli utenti, UGC, sono definiti come *“creative works that are published on publicly accessible websites and are created without a direct link to monetary profit or commercial interest”* (OECD 2001 citato in Mastrogiacomo, Barravecchia, Franceschini, e Marimon 2021). Proprio per il fatto che queste informazioni online sono frutto di uno sforzo individuale ed indipendente, e quindi non condizionato dagli interessi espliciti di un'organizzazione, da loro un valore potenzialmente molto elevato. Tuttavia, proprio perché provenienti dalla mera creatività dell'utente, gli UGC si presentano principalmente in forma testuale e destrutturata e di conseguenza difficili da trattare accuratamente ed in maniera esaustiva. Alcuni esempi di UGC sono le recensioni, la discussione relativa ad un certo argomento presente all'interno di un forum di discussione online oppure le interazioni che avvengono su un social network tra gli utenti. I fenomeni appena descritti, ovvero quelli di esprimere la propria opinione relativamente ad un prodotto/servizio online, rappresentano il passa parola digitale, il quale è stato definito come Voice-of-Customers (VoC) (Ozdogoglu, Kapucugil, e Celik 2018, citato in Barravecchia, Mastrogiacomo, e Franceschini 2022).

Ci sono diversi siti online (come Amazon, Facebook, TripAdvisor, etc.) che incoraggiano gli utenti, anche attraverso promozioni o regali, a contribuire attivamente e spontaneamente alla generazione di contenuti relativi alla loro esperienza con il prodotto/servizio e ad esprimere un feedback (Guo,

Barnes, e Jia 2017 citato in Mastrogiacomo, Barravecchia, Franceschini, e Marimon 2021). L'incentivazione alla creazione di UCG deriva dal fatto che, come diversi autori hanno sottolineato (Tirunillaie Tellis 2012 citato in Mastrogiacomo, Barravecchia, Franceschini, e Marimon 2021), questi influenzano la domanda di un determinato prodotto/servizio e, a catena, le vendite e le performance di quel prodotto/servizio. In particolare, la domanda di un prodotto/servizio è influenzata dal numero di recensioni, dalla qualità delle recensioni, dalla credibilità della piattaforma che aggrega le recensioni e dalla percezione che colui interessato all'acquisto ha relativamente alla conoscenza del prodotto/servizio di coloro che hanno espresso un feedback. Questi elementi influenzano il processo di acquisto del compratore, ma non solamente il processo d'acquisto di coloro che acquistano online, anche quello di coloro che acquisteranno offline. L'informazione online serve a questo secondo gruppo per acquisire informazioni e consapevolezza sul prodotto/servizio a cui sono interessati.

Di conseguenza, per un'organizzazione, avere prodotti/servizi che soddisfano in maniera opportuna i bisogni dei clienti è cruciale per ottenere testimonianze positive sulla proprio offerta e poter beneficiare dell'effetto che queste hanno su altri soggetti interessati a quel genere di prodotto/servizio.

In questo lavoro di tesi, a valle delle considerazioni appena fatte, sarà proposta una metodologia che comprenda l'analisi di UGC al fine dell'estrazione dei requisiti di prodotto. Il metodo proposto si pone l'obiettivo di estrarre un campione di recensioni rappresentativo dell'insieme di UCG da cui poter estrarre i requisiti manualmente. In pratica, questa metodologia può essere considerata come un filtro che consente al soggetto interessato all'estrazione dei requisiti di analizzare un campione di UGC anziché la totalità dei documenti con l'obiettivo di ottenere gli stessi risultati. Successivamente, sarà valutata la capacità di estrazione dei requisiti dalle recensioni di un modello che fa uso di tecniche avanzate di machine learning, ChatGPT, e di un modello più semplice che fa uso di insiemi di parole significative. Questa analisi sarà effettuata al fine di valutare la capacità di estrazione dei requisiti nel caso in cui il processo fosse completamente automatizzato rispetto all'estrazione manuale. I risultati di ogni metodo saranno confrontati con i risultati ottenuti tramite l'estrazione manuale, la quale sarà considerata come corretta inferenza, al fine di poter capire quanto l'estrazione dei requisiti automatizzata attraverso l'utilizzo tecniche di machine learning avanzate o puramente statistiche sia distante o vicino all'idealità. Per quanto riguarda la metodologia che fa leva sulle parole significative per l'estrazione dei requisiti dalle recensioni, è ovvio semplici parole

difficilmente riescono a dichiarare esplicitamente un requisito infatti sarà valutata la loro capacità di suggerirne l'esistenza. A valle del confronto, saranno analizzati e commentati i risultati ottenuti e tratte le conclusioni.

Dopo un breve introduzione relativa allo scopo del presente lavoro, l'elaborato presenta cinque sezioni ed una conclusione in merito all'analisi eseguita.

La prima sezione (Capitolo 3) riporta una descrizione dei metodi tradizionali utilizzati per l'estrazione dei requisiti, in particolare saranno illustrati i loro punti deboli. In seguito, sarà effettuata un'analisi della letteratura relativa agli strumenti utilizzati per l'analisi di documenti che si presentano in forma testuale. Infine, sarà esposto il prodotto pilota del lavoro e le ragioni per cui è stato scelto.

Nella seconda sezione (Capitolo 4) saranno esposti gli strumenti utilizzati. Sarà posta particolare enfasi alla spiegazione dello Structural Topic Modeling, un algoritmo di text mining che ha avuto un ruolo fondamentale nella costruzione della procedura proposta per l'estrazione dei requisiti.

Nella terza sezione (Capitolo 5) sarà spiegato come e in quali occasioni sono stati utilizzati gli strumenti descritti nel capitolo 4 all'interno del lavoro.

Nella quarta sezione (Capitolo 6) sarà illustrata la procedura proposta che porta alla costruzione del campione di UCG da cui estrarre i requisiti manualmente.

Nella quinta sezione (Capitolo 7) sarà valutata la capacità della procedura proposta nel capitolo 6 e di ChatGPT di individuare correttamente i requisiti all'interno delle recensioni rispetto all'estrazione manuale. A valle del primo confronto saranno messe a confronto le performance dei due metodi.

Infine, la conclusione propone alcune considerazioni sui risultati ottenuti all'interno del capitolo 7, miglioramenti che si potrebbero apportare alla procedura proposta e ambiti in cui la procedura proposta può essere utilizzata.

3 Descrizione del problema e analisi della letteratura

3.1 Descrizione del problema

Il problema in analisi è quello di cercare di valorizzare le informazioni presenti sul web relative a prodotti/servizi al fine della comprensione dei requisiti.

Attualmente, la comprensione dei fattori determinanti la qualità di un prodotto/servizio è ancora effettuata tramite l'analisi quantitativa delle risposte ottenute a questionari e interviste (De Vellis 2016 citato in Mastrogiacomo, Barravecchia, Franceschini, e Marimon 2021). La progettazione di questa tipologia di approccio segue 4 step fondamentali:

- Specificazione del dominio del questionario/intervista
- Creazione di un insieme iniziale di domande da porre all'interno del questionario/intervista basandosi sulla letteratura del prodotto/servizio in questione e su relative interviste e sondaggi
- Perfezionamento dell'insieme iniziale di domande
- Ottimizzazione della scala di misura delle risposte

L'utilizzo di questi metodi è ampiamente diffuso tra le diverse organizzazioni, tuttavia presenta numerosi difetti. In primo luogo, utilizzano un gran numero di risorse, sia in termini di tempo che in termini di persone coinvolte nella progettazione e successiva analisi delle risposte.

In secondo luogo, la qualità dei risultati derivanti dall'utilizzo di queste tecniche è profondamente legata all'impegno che i partecipanti impiegano nel rispondere alle domande ed alla complessità delle stesse (Groves 2006 citato in Mastrogiacomo, Barravecchia, Franceschini, e Marimon 2021).

Inoltre, i metodi tradizionali presentano ulteriori difetti relativi a:

- limitata dimensione del campione di clienti a cui il questionario è sottoposto;
- la creazione dell'insieme iniziale di domande da sottoporre per misurare la qualità del prodotto/servizio è condizionata dalla soggettività degli esperti del settore poiché alcuni fattori determinanti la qualità potrebbero essere omessi o sottostimati;
- la difficoltà con cui le risposte contenenti potenziali errori sono identificate.

Per sopperire a queste limitazioni, l'utilizzo delle informazioni ottenute direttamente dal web al fine di ottenere insights relativi alla qualità percepita di un prodotto/servizio è una possibile strada alternativa ai metodi tradizionali. In particolare, gli UGC rappresentano una fonte di informazioni

low-cost, non condizionata dalla soggettività degli esperti del settore e non vincolata dalla struttura definita in fase di progettazione (Mastrogiacomo, Barravecchia, Franceschini, e Marimon 2021).

Attualmente l'analisi degli UGC, e in particolare delle recensioni, è svolta attraverso l'utilizzo di tecniche di text mining. Queste tecniche sono capaci di ottenere informazioni tramite l'analisi di documenti scritti in linguaggio naturale.

3.2 Il text mining

Il text mining è una branca del Natural Language Processing che si occupa di estrarre in maniera automatica informazioni da dati che si presentano in forma testuale e destrutturata. Nonostante lo studio di tecniche per processare documenti scritti in linguaggio naturale risalga alla metà del secolo scorso, il text mining, per come lo conosciamo oggi, ha iniziato a svilupparsi principalmente grazie all'avvento di algoritmi di machine learning, il cui sviluppo è stato possibile grazie alla crescita delle capacità computazionali dei computer. Tuttavia, i primi algoritmi di machine learning dedicati all'analisi di documenti scritti utilizzavano regole rigide composte da costrutti if-else. Col passare del tempo, quest'ultimo approccio è stato superato grazie all'incorporazione di modelli statistici. Questi algoritmi, dopo essere stati ulteriormente adattati alle necessità odierne grazie all'arrivo di tecniche di machine learning avanzate come il deep learning e le reti neurali, sono ancora oggi utilizzati. Inoltre, i campi di applicazioni del text mining si sono ampliati e, soprattutto, è possibile ottenere risultati migliori in quelli già esistenti proprio perché l'utilizzo di queste tecniche avanzate consente di comprendere e processare il linguaggio umano in maniera molto più accurata.

Oltre ad esserci state le condizioni adeguate da un punto di vista tecnologico per lo sviluppo del text mining, il bisogno di sviluppare queste tecniche è dovuto alla prevalente quantità di dati in forma testuale rispetto ai dati numerici. La maggior parte delle informazioni di una società, infatti, proviene da chat, e-mail, blog e forum, ovvero da fonti in cui l'informazione si presenta in forma scritta. Inoltre, al giorno d'oggi, oltre a chat, e-mail, blog e forum è presente altro materiale di pubblico accesso presente sul web che si presenta sotto forma di testo, come le recensioni presenti su una piattaforma.

Il text mining è un campo multidisciplinare, all'interno del quale si presentano le seguenti discipline:

- information retrieval, che si occupa della rappresentazione, memorizzazione e organizzazione dell'informazione testuale.

- data mining, un processo che si occupa di analizzare grandi quantità di dati numerici al fine di estrarne informazioni di valore. Benché data mining e text mining sono processi che sono collegati, essi presentano qualche differenza. In primo luogo, il data mining cerca di estrarre informazioni di valore precedentemente sconosciute dall'analisi di dati numerici. Queste informazioni sono implicite nei dati ed è difficile ottenerle senza che i dati siano analizzati con opportune tecniche. Al contrario, il text mining, parte da input di tipo testuale in cui l'informazione contenuta è esplicita in quanto è scritta. Il suo compito è quello di cercare di automatizzare l'estrazione di queste informazioni. Inoltre, come conseguenza al tipo di dati che trattano, ci sono delle differenze rispetto a come sono organizzati i dati ricevuti in input. In particolare, nel caso del data mining i dati sono organizzati in maniera strutturata invece, i dati ricevuti in input nel processo di text mining sono destrutturati poiché scritti in linguaggio naturale. La differenza tra la tipologia di dati trattati, porta anche a diversità in termini del processo con cui dal dato in input si arriva all'output. Infatti, nel text mining i dati in forma testuale prima di essere analizzati sono sottoposti ad una fase di preelaborazione che ha lo scopo di organizzare i dati in maniera strutturata e secondo una codifica che consenta l'utilizzo di machine learning. Una volta che le informazioni sono strutturate a queste è applicata la tecnica di analisi che è necessaria per l'obiettivo.
- machine learning, al fine di rendere il processo di reperimento di informazione dal testo automatizzato, efficiente e che produca risultati il più simili possibili a quelli che sarebbero ottenuti se l'analisi testuale fosse eseguita da una persona.
- statistica, utilizzata all'interno delle tecniche di elaborazione dei dati strutturati per ottenere il risultato desiderato.
- NLP, un'evoluzione della linguistica computazionale che utilizza metodi di varie discipline, come informatica, AI, linguistica e data science, per consentire ai computer di comprendere il linguaggio umano nelle forme scritte e verbali. Questa disciplina è utilizzata nella fase che porta ad avere le informazioni organizzate in maniera strutturata.

3.2.1 Steps del text mining

Come detto in precedenza, il processo di text mining richiede qualche step aggiuntivo rispetto ad un processo di semplice data mining. In particolare, come è mostrato in figura 1, le fasi del processo di text mining possono essere generalizzate in 5 passi principali: raccolta dei dati da diverse fonti, pre-processing, creazione del database strutturato, mining e analisi.

Durante la prima fase sono raccolti i dati da emails, siti internet, recensioni, a seconda della necessità. Una volta che si hanno i dati a disposizione, questi dati sono “preprocessati” al fine di rimuovere informazioni non volute e organizzare il testo in una maniera che il computer possa eseguirci delle operazioni. A valle della fase di preelaborazione, i dati saranno organizzati in maniera strutturata così che possano essere manipolati nella fase successiva, quella di mining. Durante questa fase, i dati sono analizzati utilizzando le tecnica più opportuna per lo scopo che si vuole raggiungere. Infine, durante la fase di analisi, colui interessato ai risultati verifica ed interpreta l’output prodotto nella fase di mining.

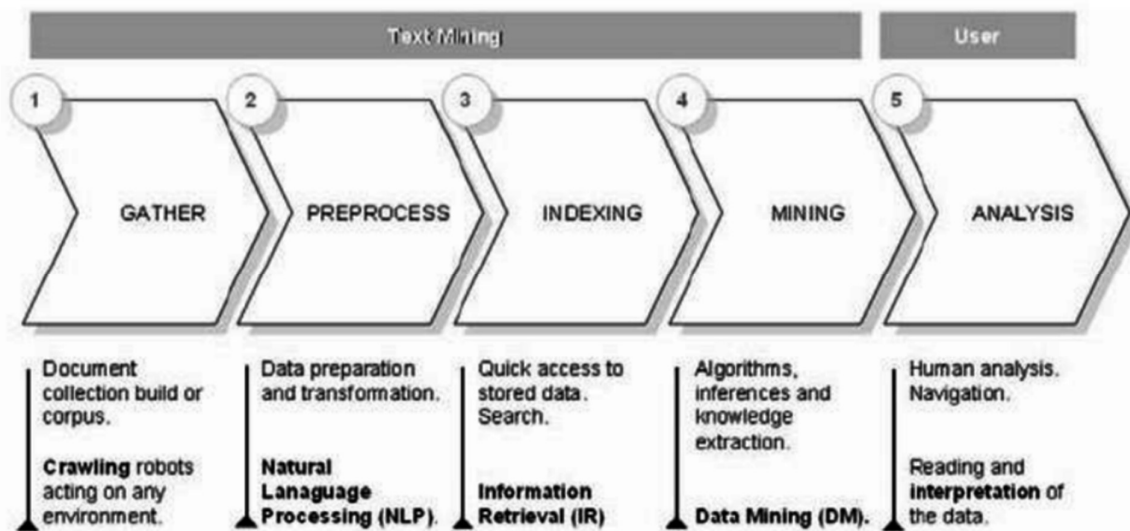


Figura 1- Fasi del processo di text mining.

3.2.2 Applicazioni del text mining

Il text mining può essere utilizzato in un’ampia varietà di contesti, per applicazioni differenti e per svolgere compiti complessi.

Tra le diverse applicazioni del text mining quella diffusa quella più nota è la sentiment analysis, la quale cerca di comprendere il sentimento presente in un feedback nei confronti di un’identità. Un’identità può essere un prodotto, un servizio, un’azienda, un’esperienza o qualsiasi altra cosa su cui è possibile esprimere un giudizio.

Oltre alla sentiment analysis esistono altre applicazioni del il text mining e una di queste è il topic modeling, utilizzato all'interno del presente lavoro. Gli algoritmi di topic modeling utilizzano tecniche statistiche avanzate al fine di identificare gli argomenti trattati all'interno di un documento che può essere, ad esempio, un insieme di recensioni o un articolo di giornale. Il compito di leggere e interpretare grandi quantità di dati è molto difficile da svolgere per l'essere umano e, inoltre, questi algoritmi risultano essere molto più efficaci ed efficienti dell'uomo.

La ragione dell'applicazione di un algoritmo di topic modeling all'analisi degli UGC è dovuta al fatto che la conoscenza degli argomenti trattati all'interno delle recensioni siano un indicatore degli aspetti rilevanti per la qualità del prodotto/servizio recensito. Tuttavia, come sarà spiegato nei capitoli successivi, conoscere quali sono gli aspetti determinanti la qualità di un prodotto/servizio non implica conoscerne esattamente quali sono i requisiti del prodotto/servizio.

Inizialmente, il metodo di topic modeling più utilizzato per identificare gli aspetti determinanti la qualità di un prodotto/servizio a partire dagli UGC è stato un modello probabilistico del testo chiamato Latent Dirichlet Allocation. Tuttavia, questo metodo non tiene in considerazione alcuni importanti metadati associati alla recensione (Mastrogiacomo, Barravecchia, Franceschini, e Marimon 2021), quali ad esempio il rating, la posizione geografica del recensore, la data ed eventualmente ulteriori metadati caratteristici della recensione. Questi metadati possono offrire ulteriori informazioni di valore. Ad esempio, considerando il rating, oltre agli argomenti trattati all'interno delle recensioni, si può capire quali aspetti del prodotto/servizio non sono apprezzati dai clienti (recensioni con rating basso) e quali invece sono apprezzati (rating alto).

Al fine di tenere in considerazione anche i metadati legati alla recensione si è passati ad utilizzare un altro algoritmo di topic modeling chiamato Structural Topic Modeling, variante del metodo LDA. Lo Structural Topic Modeling, oltre ad inferire i vari argomenti trattati all'interno di un insieme di documenti, considera anche i relativi metadati permettendo di effettuare un'inferenza potenzialmente più esaustiva (Mastrogiacomo, Barravecchia, Franceschini, e Marimon 2021). Inoltre, oltre alla considerazione dei metadati legati alle recensioni, l'STM presenta altri benefici rispetto al metodo LDA:

- non richiede l'impostazione di nessun parametro di input a differenza di LDA che oltre ad avere la necessità di stimare i parametri in input del modello, è altamente sensibile a questi;
- ha una metodologia ben definita che indica i vari steps necessari da seguire, dall'estrazione dei dati all'analisi dei risultati.

Per queste ragioni, all'intero di questo lavoro di tesi sarà utilizzato lo Structural Topic Modeling.

3.3 Scelta del prodotto

Per conseguire lo scopo del lavoro è stato scelto di analizzare le recensioni relative al prodotto Sony ZX Series Wired On-Ear Headphones, Black MDR-ZX110.

La scelta del seguente prodotto è avvenuta principalmente per due ragioni. In primo luogo, questo prodotto era l'unico a cui facevano riferimento un numero di recensioni di gran lunga maggiore rispetto al numero di recensioni relative a ciascuna delle altre opzioni. In particolare, mettendo insieme le recensioni presenti su due siti relative al prodotto scelto è stato possibile avere a disposizione oltre 22000 recensioni da utilizzare come input per il lavoro. In secondo luogo, pensando al processo di estrazione dei requisiti, è stato pensato che la familiarità con questa tipologia di prodotti potesse aiutare a valutare in maniera più semplice e, allo stesso tempo, più critica i risultati che si sarebbero ottenuti.



Figura 2-Prodotto pilota, Sony ZX Series Wired On-Ear Headphones, Black MDR-ZX110

4 Strumenti utilizzati

L'obiettivo del seguente capitolo è quello di illustrare gli strumenti utilizzati per conseguire il lavoro. Il primo paragrafo presenterà il funzionamento del Structural Topic Modeling, in particolare saranno illustrati l'output che si ottiene e i vari steps necessari per ottenerlo. All'interno dei paragrafi successivi saranno presentati Octoparse, il software utilizzato per effettuare la raccolta dati, Rstudio, l'ambiente di sviluppo del Structural Topic Modeling che è stato utilizzato anche per implementare il metodo proposto, e ChatGPT.

4.1 Structural Topic Modeling

Lo Structural Topic Modeling rientra all'interno degli strumenti di machine learning che utilizzano tecniche di text mining per l'estrapolazione delle informazioni presenti all'interno di un insieme di documenti che si presentano in forma testuale. In particolare, lo Structural Topic Modeling si occupa di identificare gli argomenti (topics) trattati all'interno dell'insieme di documenti in input, associare ad ogni topic un insieme di parole e le proporzioni in cui queste lo descrivono e, infine, definire per ogni documento quali topics sono trattati al suo interno e in quali proporzioni.

4.2 Output dello Structural Topic Modeling

Prima di spiegare i vari step necessari per ottenere l'output dello Structural Topic Modeling sarà descritto proprio quest'ultimo.

L'output del Structural Topic Modeling è composto da due elementi:

- il *Topical Content*, ovvero una distribuzione di probabilità associata ad ogni topic che indica la proporzione con cui ciascuna parola presente all'interno del vocabolario contribuisce a definire il topic. Il vocabolario è un insieme di parole standardizzate che è stato definito nei primi passi dell'algoritmo ed è caratteristico dell'insieme di documenti che si sta analizzando. Ogni parola può contribuire a definire più topic, in proporzioni differenti;
- la *Topical Prevalence*, ovvero una distribuzione multinomiale di probabilità associata ad ogni documento. La distribuzione multinomiale di probabilità indica la probabilità con cui ciascun topic è trattato all'interno di ciascun documento.

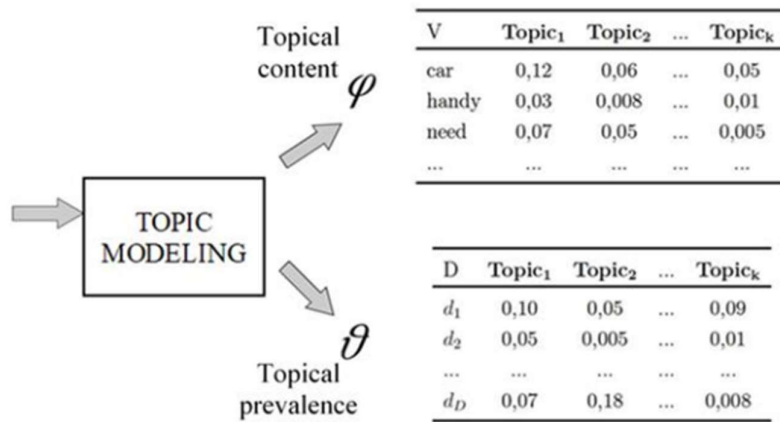


Figura 3- Output dello Structural Topic Modeling (Barravecchia, Mastrogiacomo e Franceschini 2021)

4.3 Funzionamento dello Structural Topic Modeling

Come è mostrato in figura 4, l'algoritmo di topic modeling segue 5 steps fondamentali:

- 1) Raccolta dei dati
- 2) Pre processing (preelaborazione)
- 3) Identificazione del numero ottimale di topics
- 4) Applicazione del topic modeling
- 5) Labeling

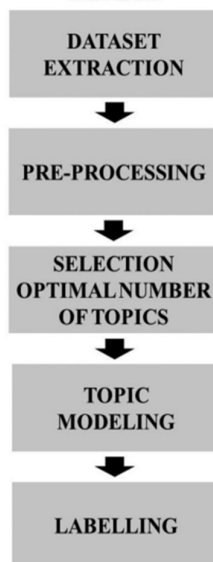


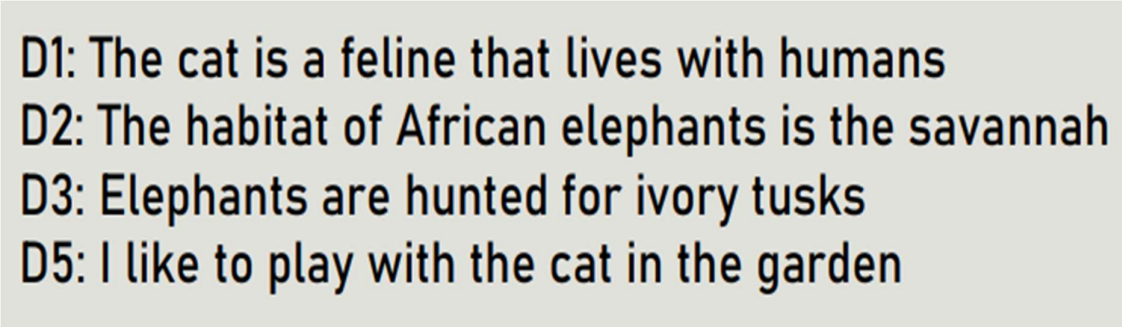
Figura 4- Steps dell'STM

4.3.1 Raccolta dei dati

La prima fase dell'algoritmo consiste nel raccogliere gli UGC dai siti online, dai social media o da qualunque piattaforma online in cui sono presenti recensioni relative ad un prodotto/servizio.

Questa attività, detta web scraping, si può fare utilizzando alcune librerie già incluse all'interno dei programmi di text mining (Zaki e McColl-Kennedy 2020 citato in Mastrogiacomo, Barravecchia, Franceschini, e Marimon 2021) oppure utilizzando un software dedito proprio a questa attività, come è stato fatto per il presente lavoro.

La scelta di utilizzare un software per effettuare web scraping è stata fatta poiché questi programmi hanno un'interfaccia utente molto intuitiva e consentono all'utente di avere un'anteprima di quali dati saranno scaricati e di come saranno strutturati, consentendo eventualmente di effettuare le opportune modifiche ed ottenere il risultato desiderato. Inoltre, oltre alla raccolta delle recensioni, i software di web scraping consentono di ottenere i metadati rilevanti associati alle recensioni in maniera molto semplice. Tuttavia, ai fini dell'obiettivo di questo lavoro i metadati non hanno giocato un ruolo fondamentale. Un esempio esplicativo di come si presentano i dati raccolti tramite il web scraping e processati dall'algoritmo è mostrato in figura 5.



D1: The cat is a feline that lives with humans
D2: The habitat of African elephants is the savannah
D3: Elephants are hunted for ivory tusks
D5: I like to play with the cat in the garden

Figura 5-Esempio di un insieme di documenti ottenuti tramite web-scraping.

4.3.2 Pre- processing

Mentre la raccolta dati non è un vero e proprio passo dell'algoritmo ma è necessario, il pre-processing è la prima vera fase del Structural topic Modeling.

Una volta che si ha a disposizione il database contenente i documenti da analizzare, questi devono essere standardizzati e unificati al fine di migliorare l'efficienza dell'algoritmo di topic modeling.

Le operazioni eseguite sui documenti in forma testuale sono (Mastrogiacomo, Barravecchia, Franceschini, e Marimon 2020; Mastrogiacomo, Barravecchia, Franceschini, e Marimon 2021):

- convertire tutto il testo in carattere minuscolo al fine di eliminare l'ambiguità che si può creare con le stesse parole in maiuscolo;
- rimozione della punteggiatura e dei numeri poiché non contribuiscono molto a generare il contenuto dei topics (topical content) e di conseguenza sono di scarsa rilevanza per i fini dell'algoritmo;
- rimozione delle *stopwords* inglesi, quali ad esempio "the", "and", "when", "is", "at", "which", "on", etc.;
- rimozione di parole composte da meno di due caratteri o da più di quindici caratteri;
- rimozione delle parole che si presentano con una bassa frequenza all'interno della totalità dei documenti. In particolare, parole con meno di 15 occorrenze sono rimosse in quanto la loro inclusione potrebbe distorcere il risultato dell'algoritmo o non sarebbe rappresentativa di nessun argomento specifico (topic);
- normalizzazione del testo utilizzando il "Porter stemmer" per aggregare parole simili in un unico termine. In questa fase, le desinenze delle parole in inglese sono rimosse ed aggregate in un'unica parola in base alla radice. Ad esempio, alle parole "likes", "liked", "likely" e "linking" sarà rimossa la desinenza e saranno sostituite tutte con la stessa parola "like";
- rimozione delle parole solitamente non correlate a nessun topic come, ad esempio, "another", "mean", "problem", "review", "made", "did", "done", etc.;
- sostituzione degli n-gram più comuni con una singola parola. Un n-grams è una sequenza ripetitiva di parole utilizzate per definire un concetto. Ad esempio, la parola "customer service", la quale definisce un concetto proprio, sarà unita al fine di formare una singola parola.

A valle di tutti questi passaggi il contenuto delle recensioni sarà semplificato e pronto per essere analizzato. Inoltre, le parole sopravvissute al processo di stemming andranno a costituire il vocabolario, cioè l'insieme delle parole utilizzate all'interno dell'insieme di documenti "stemmati". La figura 6 mostra un esempio esplicativo del processo di stemming applicato ai documenti in figura 5.

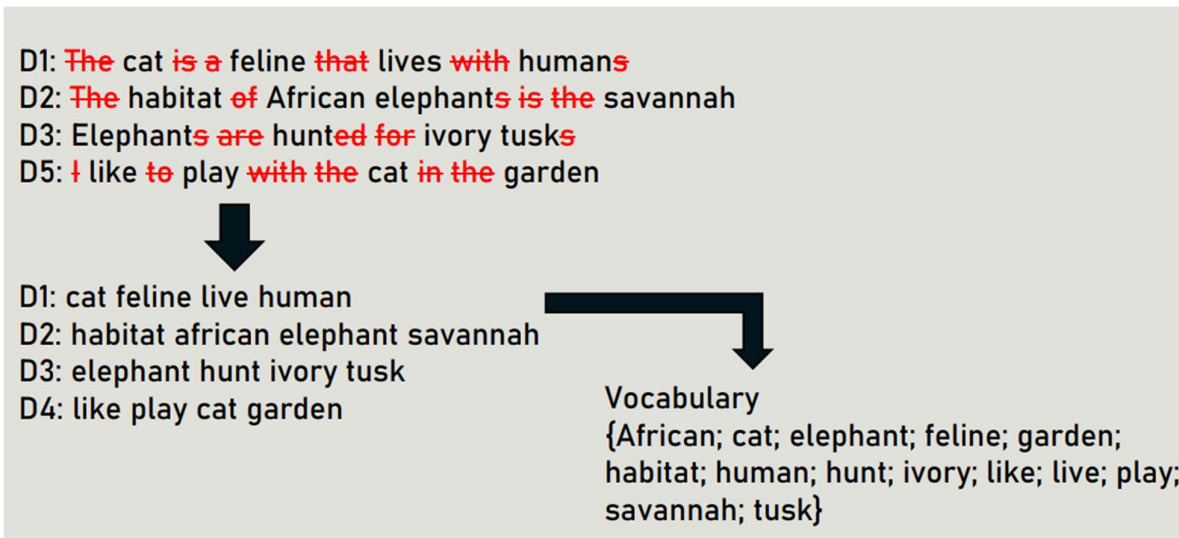


Figura 6- Esempio del processo di stemming eseguito su un insieme di 4 documenti. A valle di questa fase i documenti sono semplificati ed è creato il vocabolario.

4.3.3 Identificazione del numero ottimale di topics

Un parametro fondamentale dell'algoritmo è il numero ottimale di topics in grado di descrivere l'insieme di documenti. Per trovare questo parametro, l'algoritmo è ripetuto iterativamente facendo variare il numero di topics e misurandone le performance ad ogni iterazione. Una volta calcolate le performance in funzione del numero di topics sarà scelto il numero di topics che produce le performance migliori.

Per misurare le performance dell'algoritmo sono state proposte diverse metriche. In questo lavoro sono state utilizzate metriche predittive poiché sono le più diffuse. Quando si adotta questa tipologia di metriche solitamente si utilizza una parte dei documenti (90%), detta training set, a cui si applica l'algoritmo mentre la parte rimanente dei documenti, detta test set, si utilizza per misurare le performance dell'algoritmo impostato con il training set.

La metrica predittiva più comune utilizzata per misurare la performance dell'algoritmo è l'held-out likelihood. Questa metrica misura quanto l'algoritmo impostato con il training set di documenti è in grado di descrivere un insieme di documenti mai visti, il test set. Inoltre, può anche essere vista come una misura di quanto l'algoritmo di topic modeling è in grado di descrivere la variabilità dell'insieme totale di documenti. L'insieme di valori che può assumere l'held-out likelihood varia da $[-\infty; 0]$. A valori maggiori corrispondono performance migliori in quanto il modello sviluppato è considerato più solido statisticamente. Dal momento in cui esistono più modi per calcolare questa metrica ne

sarà data una definizione generale. L'held-out likelihood è definita come il logaritmo della probabilità di descrivere il test set dato il modello impostato con il training set (Mastrogiacomo, Barravecchia, Franceschini, e Marimon 2020):

$$L = \log (\text{prob}(data_{test-set} | M_{trained})) \quad (4.1)$$

Una volta calcolate le performance in funzione del numero di topics si otterrà un grafico simile a quello nella figura 7. Come si può osservare, l'asse delle ordinate rappresenta l'held-out likelihood e l'asse delle ascisse il numero di topics: si sceglierà il numero di topic con held-hout likelihood maggiore.

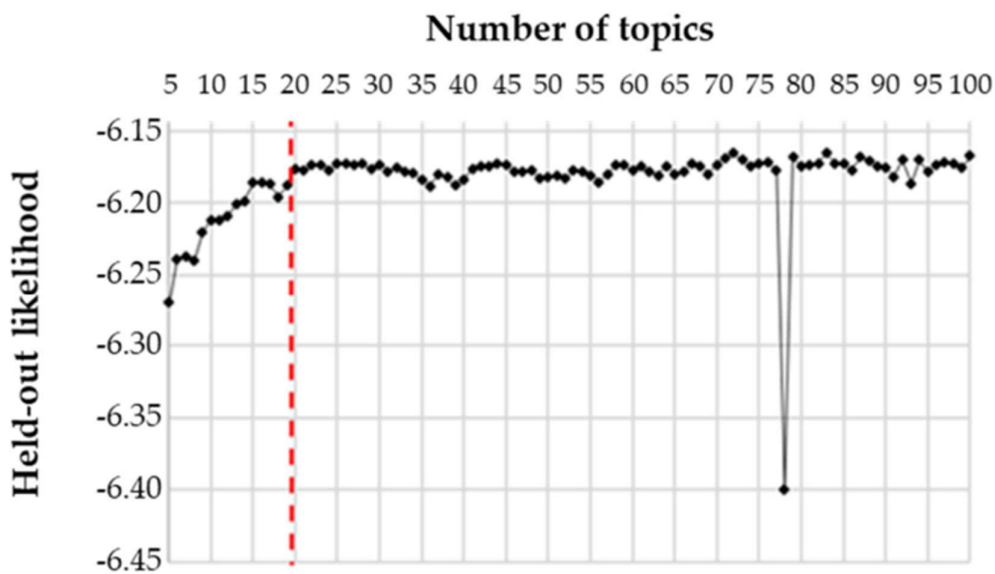


Figura 7- Grafico rappresentante l'held-out likelihood di una serie generica di documenti (Mastrogiacomo, Barravecchia, e Franceschini 2021). In questo caso è stato scelto un numero di topic pari a 20 in quanto ad un incremento del numero di topics non si ottengono performance migliori.

4.3.4 Applicazione del topic modeling

Una volta trovato il numero di topics ottimale, l'algoritmo è pronto per essere applicato all'intero insieme di documenti.

In input riceverà:

- il numero di topics, K, definito al passo precedente;

- una matrice contenente i documenti processati nei passi precedenti sulle righe e le parole del vocabolario sulle colonne. Come mostrato in figura 8, per ogni documento sarà presente un 1 se la parola del vocabolario è presente al suo interno e 0 altrimenti.

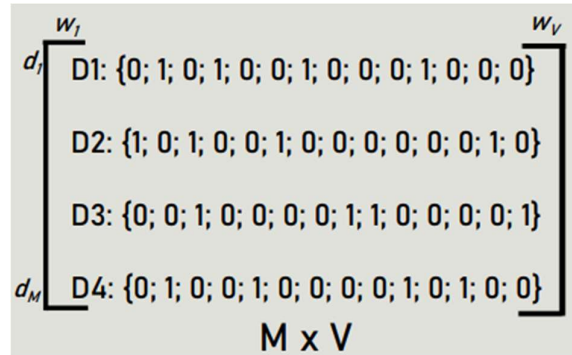


Figura 8: matrice di incidenza documenti-vocabolario.

Questa matrice in input sarà trattata dall'algorithmo di topic modeling come il prodotto di due matrici incognite. Nonostante siano incognite, conoscendo la matrice risultante si può dire con certezza che la prima matrice avrà sulle righe i documenti e la seconda matrice avrà sulle colonne le parole del vocabolario. Inoltre, considerando l'output del Structural Topic Modeling e le regole del prodotto tra matrici risulta che sulle colonne della prima matrice e sulle righe della seconda ci saranno sicuramente i vari topics. Le matrici ottenute saranno rispettivamente la matrice contenente la topical prevalence e il topical content. In figura 9 è mostrato un esempio generico della matrice in input (matrice $M \times V$) e della sua scomposizione nella matrice di topical prevalence e di topical content

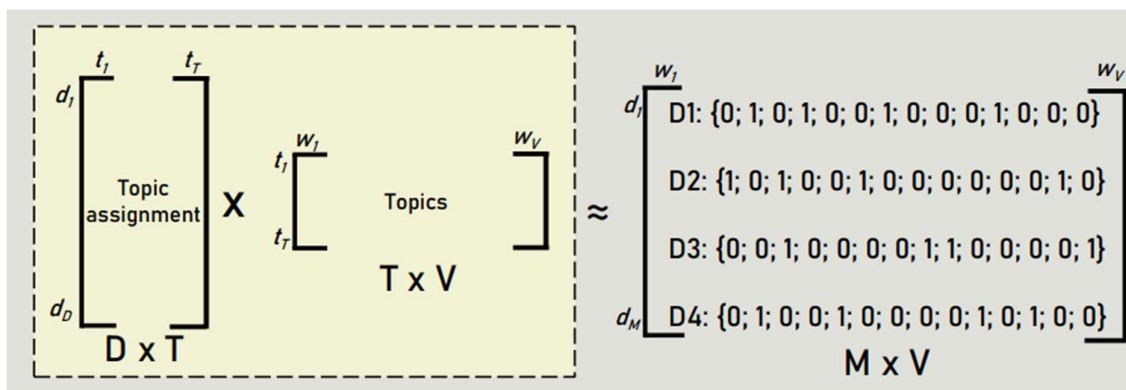


Figura 9- Da sinistra a destra: matrice topical prevalence, matrice topical content e la matrice in input

Per ottenere la matrice di input $M \times V$, dove M e V indicano rispettivamente il numero di documenti e il numero di parole all'interno del vocabolario, sarà necessario avere la prima matrice $M \times T$ e la seconda $T \times V$. Il parametro T non è incognito, ma è esattamente il numero di topic ottimale, ovvero $T=K$.

L'inferenza a ritroso fatta dal Structural Topic Modeling è un problema di fattorizzazione matriciale, cioè di decomposizione di una matrice nel prodotto di più matrici.

Riassumendo, l'algoritmo restituirà la prima matrice contenente la topical prevalence θ_k e la seconda matrice contenente il topical content β_k . Le figure 10 e 11 mostrano rispettivamente esempi di righe della matrice topical prevalence e di righe della matrice topical content.

$$\begin{array}{ll} \theta_{d1} = \{0; 0,75; 0,25\} & \theta_{d2} = \{0,65; 0; 0,35\} \\ \theta_{d3} = \{0,95; 0; 0,05\} & \theta_{d4} = \{0; 0,75; 0,25\} \end{array}$$

Figura 10- Righe di una matrice generica di topical prevalence composta complessivamente da tre topics. Ad esempio, il documento θ_{d1} è composto per il 0% dal topic 1, per il 75% dal topic 2 e per 25% dal topic 3.

$$\begin{array}{l} \beta_1 = \{0,12; 0; 0,25; 0; 0; 0,08; 0; 0,15; 0,15; 0; 0,05; 0; 0,1; 0,1\} \\ \quad \rightarrow \text{TOPIC 1 = Elephant} \\ \beta_2 = \{0; 0,45; 0; 0,25; 0,05; 0; 0,05; 0; 0; 0,5; 0; 0,15; 0; 0\} \\ \quad \rightarrow \text{TOPIC 2 = Cat} \\ \beta_3 = \{0,15; 0; 0; 0; 0,25; 0,20; 0,10; 0; 0; 0,15; 0; 0; 0,15; 0\} \\ \quad \rightarrow \text{TOPIC 3 = Places} \end{array}$$

Figura 11- Righe di una matrice generica di topical content composta complessivamente da un vocabolario di 14 parole. Ad esempio, la prima parola del vocabolario contribuisce a definire il topic 1 per il 12%, non contribuisce a definire il topic 2 e contribuisce a definire il topic 3 per il 15%.

4.3.5 Labeling

Per completezza, nonostante non sia stata rilevanti al fine del presente lavoro, è descritta un'ultima fase dell'STM. Dopo aver prodotto il proprio output, lo Structural Topic Modeling contiene un'ultima fase, Labeling, in cui con l'intervento umano i vari topics vengono etichettati, cioè ad ogni argomento è assegnato un titolo indicativo del contenuto del topic. Questa fase serve per aiutare l'umano a

trovare l'etichetta più opportuna per ogni topic. In particolare, è utile a ridurre il carico cognitivo durante il processo di assegnazione del titolo, a non lasciare spazio alla soggettività, a riassumere le informazioni su un topic con un titolo e comunicare i risultati più agevolmente. L'STM, a questo proposito, identifica le parole più rilevanti per ogni topic. Queste parole sono identificate attraverso diversi criteri, solitamente proposti insieme. Alcuni tra i vari criteri utilizzati sono il criterio della probabilità più alta (Highest probability) e il criterio delle parole frequenti ed esclusive (FREX) (Roberts, Stewart, e Tingley 2019 citati in Mastrogiacono, Barravecchia, Franceschini, e Marimon 2021). Il primo criterio propone le N parole con la probabilità più alta, dove N è un parametro scelto da colui che utilizza l'STM e che di default è impostato uguale a 8. Il secondo, invece, propone parole che sono allo stesso tempo frequenti ed esclusive per altri topics.

Oltre ai due metodi appena citati ne esistono altri che seguono regole diverse. Lo Structural Topic Modeling solitamente propone i risultati ottenuti con diversi criteri al fine di offrire un visione più ampia possibile per coloro che intollerano i vari topics.

4.4 OCTOPARSE-WEB SCRAPING

Il web scraping è un metodo automatico che estrae grandi quantità di dati dai siti web. Questa attività è estremamente conveniente quando si ha a che fare con la raccolta di dati su larga scala. L'estrazione di grandi quantità di dati dal web è un processo che necessiterebbe di anni ed un intero team dedicato se fosse eseguito manualmente mentre, grazie all'impiego di metodi di automazione intelligenti è possibile estrarli velocemente e con semplicità.

Il software scelto per eseguire il web scraping è stato Octoparse. Questo strumento consente di estrarre i dati dei prodotti dai principali siti e-commerce come Amazon, Ebay, Walmart e simili. Al login si presenta una schermata in cui è possibile inserire il link della pagina da cui si vogliono scaricare i dati e il software successivamente mostrerà una simulazione della pagina come se si fosse effettivamente presenti con l'opzione di poter selezionare i dati da scaricare. In figura 12 è mostrata la pagina iniziale di Octoparse.

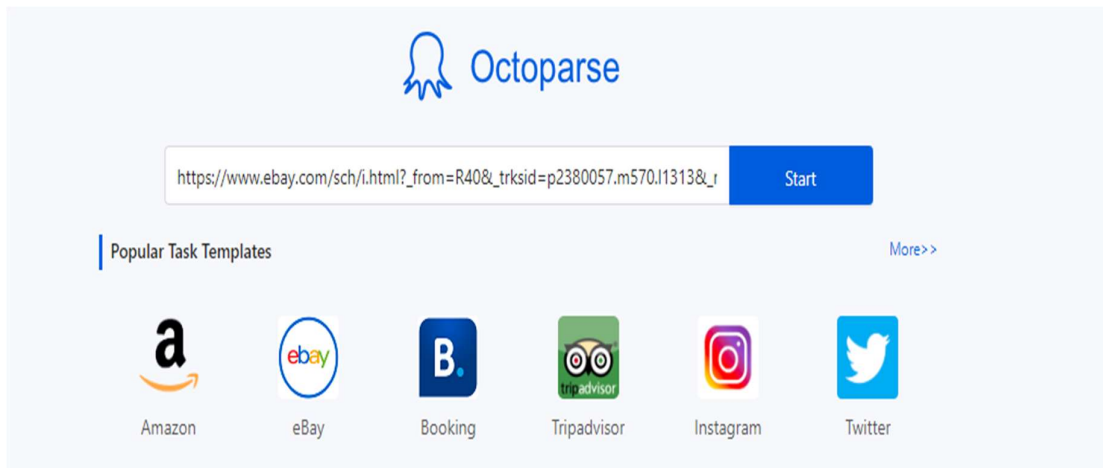


Figura 12- Schermata iniziale di octoparse.

All'interno della simulazione della pagina web, è presente un riquadro che mostra un'anteprima di come saranno organizzati i dati scaricati permettendo di effettuare modifiche prima di avviare lo scraping. Una volta che lo scraping è stato eseguito, Octoparse consente di esportare il file in diversi formati.

La figura 13 mostra un esempio di come Octoparse mostra l'anteprima dell'organizzazione dei dati all'utente. Si può notare che i dati sono organizzati in forma tabellare, si possono aggiungere o rimuovere dati o spostare le colonne. Ad esempio, nel caso in cui per qualche ragione possa risultare più conveniente avere la colonna "Title 1" tra le colonne "Title" e "Title_URL" è possibile spostarla.

Data preview: 63 line(s) of data & 16 fields captured							
#	Title	Title_URL	Image	Title1	Title2	Info	View_URL
1	DELL LAPTOP LA...	https://www.ebay.co...	https://i.ebayim...	WINDOWS 10 - ...	Refurbished · De...	Refurbished	
2	HP ProBook 655...	https://www.ebay.co...	https://i.ebayim...	Complete Teste...	Refurbished · HP...	Refurbished	https://www.ebay.c...
3	Dell Chromeboo...	https://www.ebay.co...	https://i.ebayim...	Refurbished · De...		Refurbished	

Figura 13- Anteprima fornita da Octoparse su come saranno organizzati i dati.

4.4.1 Flusso di lavoro di Octoparse

Il flusso di lavoro consente di simulare i comportamenti che gli utenti hanno all'interno delle pagine web, come aprire la pagina web, cliccare sugli elementi presenti all'interno della pagina, cliccare il pulsante per cambiare pagina o scorrere la pagina verso l'alto o verso il basso. In pratica, Octoparse

è in grado di trasferire operazioni ripetitive di estrazione manuale nel processo di estrazione automatizzato e recuperare i dati di cui si ha bisogno in maniera strutturata.

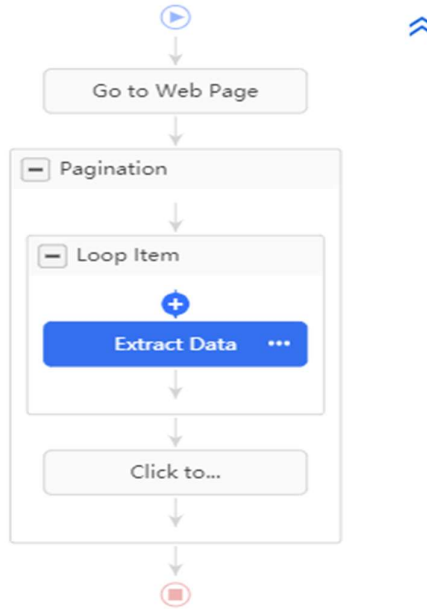


Figura 14- Esempio di un flusso di lavoro di Octoparse.

Facendo riferimento alla figura 14, saranno illustrati i vari step del flusso di lavoro:

-Step 1: Go to Web Page, apre la pagina web indicata definita dal link

-Step 2: Pagination, automatizza il processo di cambio pagina

Step 3: Loop Item, automatizza il processo di cambio item della pagina

-Step 4: Extract Data, estrae i dati selezionati dagli item

-Step 5: Click to Paginate, clicca sul pulsante di cambio pagina.

Una volta estratti tutti i dati da una pagina, ripete il processo di estrazione dati dalla nuova pagina; una volta terminati i dati da estrarre all'interno di quella pagina, cambia nuovamente pagina e ripete il processo di estrazione e cambio pagina finché Octoparse non arriva all'ultima pagina.

-Step 7: Non essendo presente alcun pulsante per cambiare pagina il work flow termina

4.5 RSTUDIO

RStudio è un ambiente di sviluppo integrato in cui è possibile elaborare dati, eseguire dai calcoli più semplici ai più complessi, ottenere rappresentazioni grafiche attraverso script scritti in linguaggio R.

Come mostrato in figura 15, RStudio è composto di quattro diverse aree:

1. Scripting area, in cui è possibile aprire, creare e scrivere lo script;
2. Console area, in cui vengono eseguiti i comandi;
3. Workspace/History area, in cui è possibile trovare una lista di tutti gli oggetti creati nello spazio di lavoro in cui si sta lavorando;
4. Visualization area, in cui è possibile caricare facilmente i pacchetti, aprire file di aiuto e visualizzare i grafici. Inoltre, all'interno di quest'area è presente un comando di supporto, Help, in cui sono spiegati i parametri in input e l'output delle varie funzioni su richiesta dell'utente.

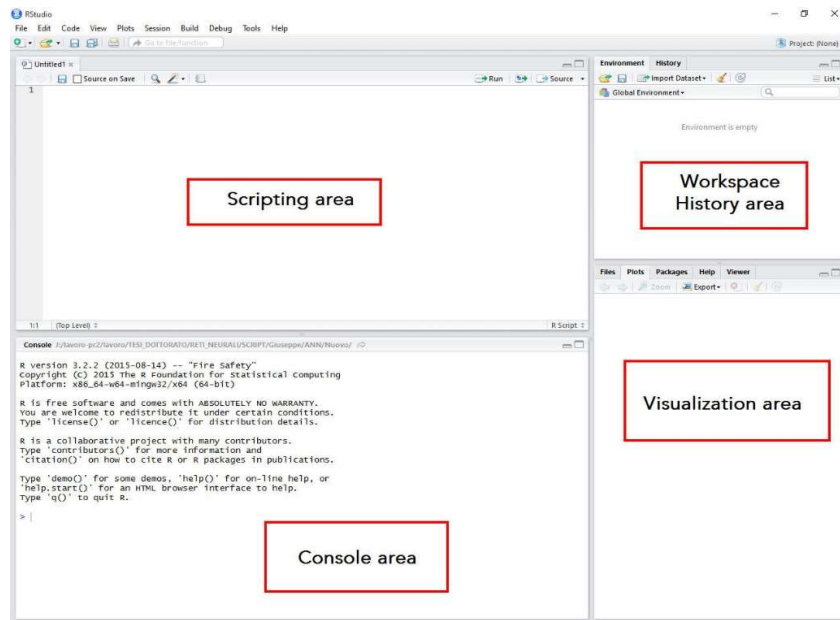


Figura 15- Layout di RStudio

4.6 ChatGPT

ChatGPT è un modello avanzato di intelligenza artificiale sviluppato da OpenAI. Si tratta di un sistema di generazione di linguaggio naturale progettato per comprendere e generare testo in linguaggio umano in modo coerente e contestualmente agli input forniti dall'utente. Il nome "ChatGPT" deriva dalla sua principale applicazione, ossia la capacità di sostenere conversazioni virtuali con gli utenti. Questo modello di IA è addestrato su un vasto corpus di testi provenienti da internet che coprono una vasta gamma di argomenti e stili di scrittura. Grazie a questo addestramento, ChatGPT è in grado di rispondere a una vasta gamma di domande, fornire

spiegazioni, generare testo creativo, assistere nell'elaborazione del linguaggio naturale e molto altro. Inoltre, la sua capacità di apprendimento automatico gli consente di migliorare costantemente la sua performance grazie all'interazione con gli utenti e ai feedback ricevuti.

L'obiettivo principale di ChatGPT è quello di fornire un'interazione umano-computer più naturale ed efficace. Può essere utilizzato in una varietà di contesti, tra cui l'assistenza virtuale, il supporto all'utente, la creazione automatica di contenuti, la traduzione automatica, la generazione di codice e altro ancora.

Tuttavia, è importante notare che ChatGPT non è privo di limitazioni. Può produrre risposte errate o insensate, può essere influenzato da dati di addestramento di bassa qualità e può generare contenuti inappropriati.

4.6.1 Alla base di ChatGPT

ChatGPT è un modello di linguaggio che utilizza una rete neurale profonda per generare testo in modo coerente e contestualmente agli input forniti. La sua capacità di generare testo di alta qualità si basa su due principali componenti:

1. **Pre-training**, fase durante la quale il modello viene addestrato su una vasta quantità di testo proveniente da Internet. Durante questo processo, ChatGPT impara le strutture linguistiche, la grammatica, il contesto e lo stile di scrittura presenti nei testi di addestramento. Questa fase consente a ChatGPT di acquisire una conoscenza generale del linguaggio umano.
2. **Fine-tuning**. Dopo la fase di pre-training il modello viene ulteriormente addestrato su dati specifici e annotati da esperti. Questa fase di fine-tuning permette di adattare il modello a compiti specifici e di rendere il suo comportamento più controllato. Ad esempio, per creare ChatGPT, il modello è stato addestrato su conversazioni umane al fine di comprenderne il contesto e rispondere in modo significativo.

All'interno di questo lavoro, ChatGPT è stato utilizzato per estrarre i requisiti dalle recensioni e confrontare i risultati ottenuti con i risultati ottenuti tramite l'estrazione manuale.

La figura 16 rappresenta l'interfaccia utente di Chat GPT appena dopo il log-in. Come si può osservare è possibile creare diverse "Chat" e il basso al centro c'è la casella che consente a Chat GPT l'input testuale.

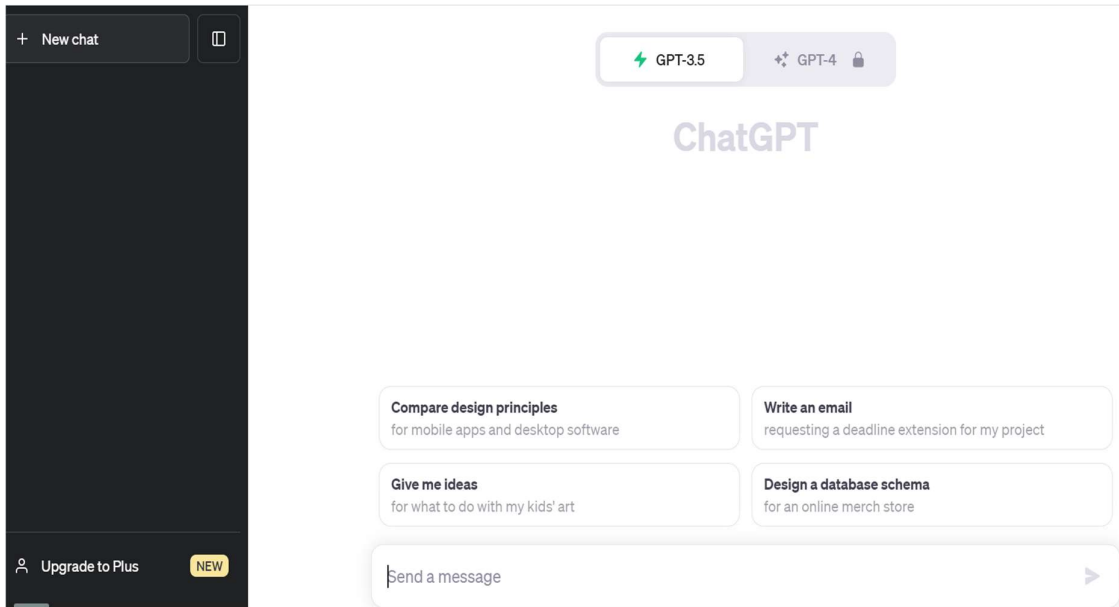


Figura 16- Interfaccia utente ChatGPT

5 Applicazione degli strumenti utilizzati al caso studio

5.1 Raccolta dei dati e pre-processing

La raccolta dei dati in input all'algoritmo di topic modeling è stata eseguita utilizzando Octoparse su due diversi siti in cui il prodotto Sony ZX Series Wired On-Ear Headphones, Black MDR-ZX110 è stato recensito: Amazon e Bestbuy. La scelta di effettuare il web scraping esclusivamente su due siti su cinque in cui erano presenti recensioni del prodotto è stata fatta per la seguente ragione: i tre siti da cui non sono state scaricate recensioni contribuivano, insieme, al 3% delle recensioni totali e di conseguenza non sono stati considerati. Raccogliendo le recensioni esclusivamente da Amazon.com e Bestbuy.com sono state scaricate complessivamente 22656 recensioni insieme ai relativi metadati.

I metadati non sono rilevanti allo scopo del presente lavoro ma sono stati utilizzati per identificare le recensioni in lingua inglese e scartare le altre poiché per l'applicazione dello Structural Topic Modeling è necessario che le recensioni siano in lingua inglese visto che l'insieme delle stopwords è stato definito esclusivamente in questa lingua. Il criterio utilizzato per discriminare una recensione scritta in inglese da una recensione scritta in un'altra lingua è stato osservare il paese di provenienza della recensione. In particolare, erano presenti recensioni scritte in portoghese, in francese, in spagnolo, in giapponese e in altre lingue diverse dall'inglese. Queste recensioni sono state rimosse sulla base del paese da cui proveniva la recensione; ad esempio, se nell'apposita colonna c'era scritto "Reviewed in Brazil" oppure "Reviewed in Germany", è stato presupposto che le recensioni fossero tutte scritte in una lingua diversa dall'inglese, come infatti lo era la maggior parte. Questo ragionamento è stato fatto poiché un controllo puntuale per verificare che ogni recensione fosse scritta in inglese sarebbe stato troppo complesso da eseguire manualmente. A valle di questo procedimento, sono state tenute le recensioni provenienti dagli Stati Uniti, dal Regno Unito, dal Canada e dall'India in quanto si presentavano quasi totalmente in lingua inglese, ottenendo così un totale di 21868 recensioni come input dell'STM.

Per identificare le recensioni scritte in inglese seguendo il criterio appena descritto è stato necessario manipolare i dati scaricati attraverso l'utilizzo di Rstudio. Inizialmente, i dati scaricati dai due siti sono stati salvati in due file differenti in quanto alcuni metadati, essendo Amazon e Bestbuy organizzati diversamente, hanno portato ad avere due database non omogenei nel contenuto di alcune colonne. In particolare, da Amazon.com la data e il luogo della recensione erano informazioni scaricabili separatamente mentre su Bestbuy.com erano un unico dato. Di conseguenza, i due database sono stati resi omogenei separando data e luogo di provenienza dalle recensioni di

Bestbuy e creando due colonne apposte per ogni dato. A valle di questo passaggio, i due database risultavano essere omogenei quindi sono stati uniti ed è stata effettuata l'eliminazione delle recensioni considerate scritte non in inglese secondo il criterio sopra descritto.

Una volta che il database in input è stato preparato, è stata effettuata la fase di pre-processing, a valle della quale le recensioni totali ed effettivamente analizzate dallo Structural Topic Modeling sono state 21776.

5.2 Identificazione del numero ottimale di topics e applicazione STM

Al fine di identificare il numero ottimale di topics, l'algoritmo è stato fatto girare variando il numero di topics. Nel caso specifico, il numero di topics è stato fatto variare da un minimo di 5 topics ad un massimo di 50 topics ed è stata calcolata l'held-out likelihood ad ogni iterazione. Sono stati ottenuti i valori rappresentati nella figura 17.

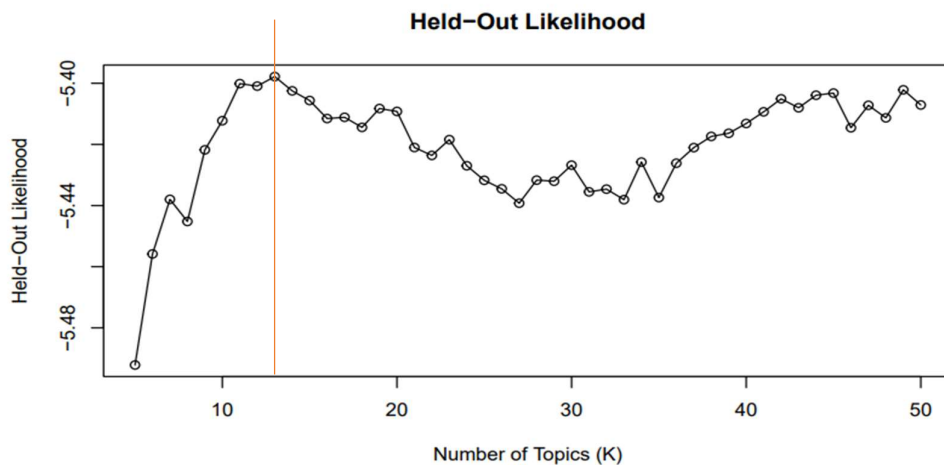


Figura 17- Valori di held-out likelihood ottenuti per il caso di studio

Come si può osservare dal grafico, il numero di topics che consente all'algoritmo di avere le performance migliori è K=13.

Individuato il numero ottimale di topics, questo è stato dato in input all'algoritmo e sono stati ottenuti i seguenti risultati in figura 18.

Top Topics

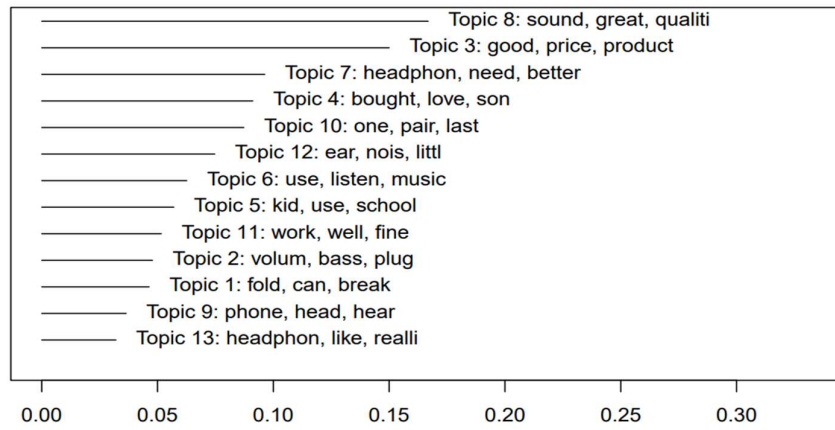


Figura 18- Percentuale dei topics all'interno dell'insieme di recensioni del prodotto Sony ZX Series Wired On-Ear Headphones, Black MDR-ZX110.

6 Metodo proposto per l'estrazione dei requisiti

Una volta che è stato applicato lo Structural Topic Modeling, è stato proposto un metodo che, partendo dall'output dell'algoritmo, consente di estrarre dall'insieme delle recensioni in input un campione da cui estrarre i requisiti manualmente. Prima di spiegare i vari passi di questo metodo è bene esplicitare alcune considerazioni che sono state fatte. In primo luogo, l'STM non estrae i requisiti dai documenti per cui sono sicuramente necessari passaggi ulteriori per avvicinarsi all'obiettivo. Infatti, un topic può contenere da 0 a N requisiti e non si può conoscere a priori quest'informazione. Per questa ragione, non sapendo dove si celano i requisiti è necessario considerare tutti i topics.

In secondo luogo, la procedura che è stata proposta si basa sull'idea che le parole caratteristiche di un topic, il topical content, possono avere a che fare con zero o più requisiti in quanto si tratta di parole stemmate. Di conseguenza, si è pensato che sarà poi l'essere umano ad estrarre i requisiti dal campione in quanto è in grado di cogliere i diversi contesti a cui una parola stemmata può far riferimento. L'estrazione manuale, inoltre, oltre a consentire il riconoscimento del contesto in cui le parole stemmate sono utilizzate, consente anche di cogliere ulteriori aspetti della recensione poiché all'interno di ogni recensione possono essere discussi più topics e conseguentemente possono essere presenti anche requisiti associati ad altri topics.

I passi della procedura sono i seguenti:

1. Output del Structural Topic Modeling (Topical Content, Topical prevalence)
R = insieme totale delle recensioni
T = insieme dei topics
2. $\forall recensione \in R$
Definizione della soglia dinamica
3. $\forall topic i \in T$
 - 3.1- Definizione dell'insieme delle recensioni rilevanti RR_i
 - 3.2- Creazione di insiemi di parole che possono avere a che fare con i requisiti partendo dal topical content del topic i e calcolo della percentuale di recensioni rilevanti in cui ogni insieme è presente
4. Scelta della dimensione del campione da analizzare manualmente
Definizione del numero di insiemi di parole n e, per ciascuno di questi, del numero di recensioni rilevanti in cui ognuno è contenuto

5. $\forall \text{topic } i \in T$

5.1- Selezione degli n insiemi di parole con percentuale maggiore che compaiono in RRi

$\forall \text{insieme di parole} \in n$

5.2- Estrazione delle m recensioni $\in RRi$ in cui l'insieme di parole è contenuto e con valore di `topical content` per il topic i maggiore.

Definiti i vari steps necessari per la composizione del campione, sarà illustrato approfonditamente il contenuto di ogni step:

- Output del Structural Topic Modeling

Il punto di partenza è stato l'output dello Structural Topic Modeling, ovvero la `topical prevalence` (distribuzione multinomiale di probabilità per ogni documento che indica la percentuale dei vari topics trattati all'interno del documento) e il `topical content` (indica da quali parole è composto un topic e in quali proporzioni).

- Definizione della soglia dinamica

In questo step si definisce il valore più opportuno di una soglia che servirà a comprendere se un determinato topic è rappresentativo all'interno di una recensione, ovvero se all'interno della recensione un topic è sufficientemente discusso e per cui tale recensione può essere considerata rilevante per quel topic.

I metodi presi in considerazione per far fronte a questo problema sono stati tre ed è stato deciso di adottare l'ultimo in quanto considerato il più opportuno (Mastrogiacomo, Barravecchia, e Franceschini 2021).

Il primo metodo sarebbe stato quello di definire come rappresentativo per un documento il topic che presenta valore massimo di `topical prevalence`. Questa soluzione non è ottimale in quanto, visto che all'interno di ciascun documento si possono trattare più topic, potrebbe escludere alcuni topics che, nonostante non possiedano il massimo valore di `topical prevalence`, sono comunque sufficientemente discussi.

La seconda alternativa comprenderebbe l'utilizzo di una soglia comune a tutti i documenti la quale, però, presenta un grande difetto. Vista l'unicità di ciascuna recensione e di conseguenza l'unicità della `topical prevalence` relativa a ciascun documento, un unico valore soglia comune a tutti i

documenti potrebbe non indicare come rappresentativi alcuni topic che lo sono oppure indicare come rappresentativi topic marginali.

Il terzo metodo è quello di definire un'opportuna soglia per ciascun documento in modo da considerare l'unicità del documento e poter, eventualmente, considerare più topic come rappresentativi per quello specifico documento. Di conseguenza, superando i limiti presenti nei primi due metodi, si è scelto di utilizzare quest'ultimo.

Questo metodo utilizza una soglia specifica per ogni documento e, dal momento che non è una soglia fissa, è detta soglia dinamica. Analiticamente, questa soglia dinamica è calcolata utilizzando il Tukey fence method per l'identificazione degli outlier, poiché, oltre a consentire il calcolo di una soglia opportuna per ogni documento, non fa ipotesi sulla distribuzione dei dati a differenza di altri metodi utilizzati che ipotizzano una distribuzione di tipo normale. Questa metodologia definisce come outliers i punti che cadono più distanti dal primo e dal terzo quartile di $1,5 \times (Q3_i - Q1_i)$. In generale, $Q3_i$ e $Q1_i$ rappresentano rispettivamente il terzo e il primo quartile della distribuzione di probabilità e $(Q3_i - Q1_i)$ è definito come range interquartile RIQ . Quindi, sono considerati outliers i punti che cadono al di fuori dell'intervallo:

$$(Q1_i - 1,5 RIQ; Q3_i + 1,5 RIQ). \quad (6.1)$$

Nel caso in esame, l'estremo sinistro dell'intervallo non è di alcun aiuto in quanto si stanno cercando valori che superano una certa soglia e di conseguenza sarà utilizzato esclusivamente l'estremo destro, che coinciderà proprio con la soglia dinamica:

$$Soglia\ dinamica_i = Q3_i + 1,5 \times (Q3_i - Q1_i). \quad (6.2)$$

I due quartili sono stati calcolati per ogni documento sulla base della propria topical prevalence.

Un esempio illustrativo di calcolo della soglia dinamica è mostrato in tabella 1 per due recensioni generiche R1 e R2 e 13 topics (T1,T2,...,T13). Come si può osservare ogni recensione ha il proprio valore di soglia dinamica che tiene in considerazione l'unicità della topical prevalence associata ad ogni documento. I due quartili Q1 e Q3 sono calcolati tramite l'utilizzo della funzione *quartile* di Rstudio in grado di calcolare i quantili desiderati di un insieme di valori assunti da una variabile casuale. A questa funzione sono stati quindi passati come parametri i valori della topical prevalence e i relativi quantili di interesse, ovvero il venticinquesimo e il settantacinquesimo quantile, rispettivamente primo e terzo quartile (Q1 e Q3). Una volta ottenuti Q1 e Q3, è stata applicata la (6.2) ed è stata calcolata la soglia dinamica per il documento.

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	Q1	Q3	Soglia dinamica
R1	0,05	0,037	0,035	0,026	0,011	0,019	0,032	0,056	0,235	0,350	0,073	0,058	0,021	0,026	0,058	0,107
R2	0,11	0,194	0,016	0,020	0,065	0,019	0,044	0,019	0,060	0,348	0,073	0,019	0,013	0,019	0,073	0,153

Tabella 1- Esempio di calcolo soglia dinamica

- Definizione dell'insieme delle recensioni rilevanti RR_i (step 3.1)

Una volta definita la soglia per ciascun documento, si verifica per ogni recensione appartenente all'insieme R se il valore della topical prevalence per il topic i è maggiore o uguale alla soglia dinamica. Se così fosse, quel topic sarebbe rappresentativo all'interno della recensione e questa sarà considerata rilevante ed aggiunta all'insieme RR_i . Dopo aver effettuato questo confronto per ogni recensione appartenente all'insieme R si otterrà l'insieme delle recensioni rilevanti per il topic i (RR_i).

In tabella 2 è mostrato un esempio di come si definisce la rilevanza di un documento rispetto ad un topic. Come si può osservare, all'interno di R1 sono rilevanti il topic 9 e il topic 10 poiché il valore di topical prevalence per questi due topic è maggiore o uguale al valore della soglia dinamica definita per il documento. Invece, all'interno di R2 sono rilevanti il topic 2 e il topic 10. Di conseguenza, R1 contribuirà a comporre l'insieme delle recensioni rilevanti per il topic 9 e per il topic 10. Invece, R2 contribuirà a comporre l'insieme delle recensioni rilevanti per il topic 2 e per il topic 10.

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	Q1	Q3	Soglia dinamica	Topic rilevanti
R1	0,05	0,037	0,035	0,026	0,011	0,019	0,032	0,056	0,235	0,350	0,073	0,058	0,021	0,026	0,058	0,107	T9, T10
R2	0,11	0,194	0,016	0,020	0,065	0,019	0,044	0,019	0,060	0,348	0,073	0,019	0,013	0,019	0,073	0,153	T2, T10

Tabella 2- Esempio di come si definiscono i topics rilevanti per un documento

- Creazione degli insiemi di parole e calcolo della loro percentuale (step 3.2)

Definito l'insieme di documenti per cui il topic i è rappresentativo, RR_i , si passa all'utilizzo del topical content. Alla base del suo utilizzo c'è l'ipotesi che le parole da cui un topic è definito possono avere a che fare con i requisiti del prodotto.

Il topical content è servito a definire quali parole del vocabolario definiscono un topic. In particolare, definiscono un topic tutte quelle parole che hanno un valore di topical content per il topic i maggiore di zero.

Nel caso in analisi, si è scelto di valorizzare per la ricerca dei requisiti un insieme di parole composto almeno da due parole per le seguenti ragioni. In primo luogo, le parole presenti nel topical content possono essere nomi, verbi, aggettivi o avverbi e questi ultimi due assumono valore se contestualizzati. In secondo luogo, si presume che, se un requisito è ampiamente condiviso sarà nominato diverse volte e, per la maggior parte di queste volte, le parole che riguardano il punto chiave del requisito siano sempre le stesse. Per queste ragioni è stato scelto di utilizzare un insieme di minimo due parole per la ricerca dei requisiti. Ad esempio, se $W1$, $W2$, $W3$ sono parole appartenenti al topical content del topic i , “ $W1 W2$ ” è un insieme di due parole, “ $W1 W2 W3$ ” un insieme di 3 parole.

Per ogni insieme di parole è stata calcolata la proporzione di recensioni rilevanti in cui l’insieme è completamente presente. Analiticamente, la proporzione è stata calcolata nel seguente modo:

$$Proporzione = \frac{\text{Numeri di recensioni rilevanti in cui è presente l'insieme di parole}}{\text{Numero recensioni rilevanti}} \quad (6.3)$$

Per quanto riguarda il numeratore, si conta un’occorrenza dell’insieme di parole se tutte le parole che compongono l’insieme sono presenti all’interno dello stesso documento.

In tabella 3 è riportato un esempio esplicativo di come è stato effettuato questo calcolo definito un topical content generico ($W1,W2,W3$) che contribuisce a definire il topic i e definito un insieme generico di recensioni rilevanti ($R1,R2,R3,R4$) per il topic i . Visto che si considerano insiemi di minimo due parole, gli insiemi che si possono generare sono ($W1,W2$), ($W1,W3$), ($W2,W3$) e ($W1,W2,W3$). Per ciascun insieme, si conta come un’occorrenza ogni recensione che contiene contemporaneamente tutte le parole appartenenti all’insieme.

Un esempio, del risultato che si ottiene a valle di questo passaggio (3.2) è mostrato in figura 19.

Al termine di questo step saranno disponibili, per il topic i , tutti gli insiemi di parole insieme alle percentuali con cui questi si presentano all’interno delle recensioni rilevanti.

Al termine del punto 3, quindi dopo che i passaggi 3.1 e 3.2 sono stati eseguiti per tutti i topics, sarà possibile risalire, per ogni topic, agli insiemi di parole caratteristici, con relativa proporzione in cui sono presenti nelle recensioni rilevanti, ed alle recensioni rilevanti in cui questi sono contenuti.

	Insieme di parole	Presenza	Occorrenza	Insieme di parole	%
R1	(W1,W2)	(1,1)	1	(W1,W2)	1/4
R2	(W1,W2)	(1,0)	0		
R3	(W1,W2)	(1,0)	0		
R4	(W1,W2)	(1,1)	0		
R1	(W1,W3)	(1,0)	0	(W1,W3)	2/4
R2	(W1,W3)	(1,1)	1		
R3	(W1,W3)	(1,0)	0		
R4	(W1,W3)	(1,1)	1		
R1	(W2,W3)	(1,0)	1	(W2,W3)	1/4
R2	(W2,W3)	(0,1)	1		
R3	(W2,W3)	(0,0)	0		
R4	(W2,W3)	(1,1)	1		
R1	(W1,W2,W3)	(1,0,0)	0	(W1,W2,W3)	1/4
R2	(W1,W2,W3)	(1,0,1)	0		
R3	(W1,W2,W3)	(1,0,0)	0		
R4	(W1,W2,W3)	(1,1,1)	1		

Tabella 3- esempio di calcolo della percentuale di un insieme di parole

Parola		
sound	sound great	0.163
great	sound comfort	0.121
qualiti	great comfort headphon price	0.010
comfort	great comfort good price	0.004
headphon		
good	sound comfort headphon good price	0.007
nice	great comfort headphon good price	0.002
price	sound great comfort headphon good price	0.002

Figura 19- Alcuni insiemi di parole che si possono ottenere partendo dalla lista di parole a sinistra. Per ciascun insieme è mostrata anche la proporzione con cui si presenta all'interno dell'insieme delle recensioni rilevanti. Ad esempio, "sound great" è presente nel 16,3% delle recensioni rilevanti del generico topic a cui gli insiemi fan riferimento.

- Definizione del numero recensioni da analizzare

Arrivati a questo punto, il soggetto interessato all'estrazione dei requisiti del prodotto dovrà decidere quante recensioni analizzare da cui estrarre i requisiti. Dalla seguente formula dovrà decidere i parametri n e m al fine di ottenere il numero di recensioni a cui è interessato:

$$\text{Dimensione del campione da analizzare} = n * m * k \quad (6.4)$$

dove:

- $n = \frac{\text{numero di insiemi di parole}}{\text{Topic}}$, cioè il numero di insiemi di parole per topic;

- $m = \frac{\text{Num di recensioni}}{\text{Insieme}}$, cioè il numero di recensioni rilevanti in cui è contenuto l'insieme di parole;

- $k = \text{numero di topic}$.

Il numero di topic non è da decidere in quanto è il numero ottimale di topic definito dal Structural Topic Modeling.

- Estrazione del campione da analizzare

Scelta la dimensione del campione da analizzare, l'algoritmo selezionerà, per ogni topic, gli n insiemi di parole più probabili poiché si presume che gli insiemi di parole con percentuale più alta possano avere realmente a che fare con i requisiti in quanto sistematicamente presenti contemporaneamente all'interno delle recensioni. In seguito, per ciascuno di essi saranno estratte le m recensioni che contengono l'insieme di parole e con la maggior topic prevalence per il topic i .

Ipotizzando che colui interessato all'analisi delle recensioni abbia deciso, al passo precedente, $n=2$ ed $m=1$ dovrà quindi analizzare un numero di recensioni pari a $(K*2*1)$ recensioni. Con questi parametri, per ognuno dei K topics saranno considerati i due insiemi di parole più probabili e per ciascuno di questi sarà estratta una singola recensione rilevante contenente l'insieme di parole. La recensione estratta per ogni insieme di parole sarà la recensione che contiene l'insieme di parole e che, tra le recensioni rilevanti contenenti l'insieme di parole, possiede il valore più alto di topic prevalence per il topic in questione. Tuttavia, può succedere che la recensione con valore più alto di topic prevalence per il topic in questione contenga più di un insieme di parole e quindi in due casi coincida con la recensione con massimo valore di topical prevalence. In questo caso, seguendo il processo, si inizia dal considerare l'insieme di parole più probabile e da questo sarà selezionata ed aggiunta al campione da analizzare la recensione con valore più alto di topic prevalence. Una volta

aggiunta la recensione contenente il primo insieme di parole si passa al prossimo insieme di parole e si selezionerà la recensione contenente il secondo insieme con valore più alto di topic prevalence per il topic i. In questo caso la recensione coinciderà con quella già selezionata dal primo insieme di parole, di conseguenza, essendo questa recensione già presente all'interno del campione da analizzare, si selezionerà la seconda recensione rilevante con topic prevalence maggiore per il topic i contenente il secondo insieme di parole e sarà aggiunta al campione da analizzare.

Discorso analogo vale anche nel caso in cui un insieme di parole di un generico topic j andrebbe a selezionare una recensione che è già stata selezionata da un insieme di parole del topic i. Anche in questo caso, si verifica se la recensione contenente l'insieme di parole del topic j che si aggiungerebbe al campione da analizzare non sia già presente. Se non è ancora presente sarà aggiunta, altrimenti si passa a selezionare la prossima recensione che contiene l'insieme di parole del topic j con topic prevalence maggiore dopo quella che sarebbe stata aggiunta.

Riassumendo, ogni volta che una recensione è candidata ad essere selezionata per essere aggiunta all'insieme di recensioni da analizzare si verifica che questa non sia già presente. Nel caso in cui non lo sia sarà aggiunta, altrimenti sarà selezionata la prima recensione più rappresentativa per il topic in questione dopo quella che sarebbe stata selezionata.

sound great	0.163
sound comfort	0.121
sound headphon	0.185
sound good	0.166
sound price	0.170
great comfort	0.056
great headphon	0.114
great good	0.049
great price	0.113
comfort headphon	0.078
comfort good	0.056
comfort price	0.056

Figura 20- Esempio di insiemi di parole

Per esempio, si faccia riferimento alla figura 20 e alla figura 21. Si ipotizzi anche che gli insiemi di parole in figura 20 siano relative al topic 10 (T10) in figura 21. Inoltre, si ipotizzi che sia stato scelto $n=2$ ed $m=1$, quindi per ogni topic saranno scelti i due insiemi di parole più probabili all'interno delle

recensioni rilevanti e sarà estratta una recensione per ognuno di essi. In questo esempio, saranno scelti gli insiemi “sound headphon” e “sound price”.

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	Soglia dinamica
R1	0,047	0,037	0,035	0,026	0,011	0,019	0,032	0,056	0,235	0,350	0,073	0,058	0,021	0,107
R2	0,109	0,194	0,016	0,020	0,065	0,019	0,044	0,019	0,060	0,348	0,073	0,019	0,013	0,153
R3	0,067	0,020	0,092	0,077	0,027	0,029	0,079	0,038	0,036	0,313	0,158	0,040	0,024	0,155
R4	0,057	0,023	0,032	0,018	0,011	0,014	0,029	0,081	0,024	0,324	0,042	0,328	0,017	0,116
R5	0,030	0,314	0,027	0,016	0,008	0,017	0,041	0,034	0,012	0,441	0,021	0,022	0,018	0,060
R6	0,186	0,067	0,060	0,014	0,009	0,016	0,051	0,055	0,178	0,033	0,071	0,235	0,025	0,140
R7	0,076	0,114	0,055	0,018	0,010	0,042	0,162	0,032	0,059	0,344	0,020	0,029	0,037	0,147
R8	0,073	0,028	0,378	0,028	0,022	0,027	0,067	0,151	0,014	0,119	0,021	0,049	0,023	0,147
R9	0,400	0,016	0,050	0,030	0,011	0,010	0,101	0,029	0,010	0,275	0,015	0,038	0,015	0,102
R10	0,108	0,050	0,095	0,037	0,024	0,075	0,063	0,083	0,024	0,210	0,094	0,109	0,029	0,181

Figura 21- Campione di recensioni con relative topical prevalence e soglia dinamica.

Il primo caso è quello in cui per due insiemi di parole differenti si selezionano recensioni rilevanti differenti quindi, si ipotizzi che i due insiemi di parole siano presenti all’interno delle 10 recensioni di figura 21 nel modo rappresentato in figura 22, dove 1 indica la presenza dell’insieme di parole all’interno della recensione e 0 l’assenza.

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
"sound headphone"	1	1	0	0	1	0	1	1	0	1
"sound price"	0	1	0	1	0	0	1	0	1	1

Figura 22-Presenza/non presenza dei due insiemi di parole all’interno delle 10 recensioni di figura 21 (caso1).

A questo punto, il metodo proposto andrà a cercare la recensione rilevante per il topic 10 con più alta topic prevalence e contenente l’insieme “sound headphone”. Le recensioni rilevanti per il topic 10, facendo riferimento alla figura 20, sono R1, R2, R3, R4, R5, R7, R9, R10 in quanto il valore di topic prevalence del topic 10 per ognuna di queste recensioni è superiore alla rispettiva soglia dinamica. L’insieme “sound headphone”, facendo riferimento alla figura 21, è presente in R1, R2, R5, R7, R8, R10. Per cui, l’estrazione di una recensione che allo stesso tempo sia rilevante per il topic 10 e che contenga l’insieme “sound headphone” sarà effettuato sull’insieme R1, R2, R5, R7, R10. Le recensioni R3, R4 e R9 non sono considerate perché non contengono “sound headphon”, invece R8 non è considerata perché non è rilevante per il topic 10, nonostante contenga l’insieme “sound

headphone". A questo punto, tra le recensioni rilevanti e contenenti l'insieme "sound headphone" (R1,R2,R5,R7,R10), sarà selezionata la recensione con topical prevalence maggiore per il topic 10, ovvero R5 che ha un valore di topical prevalence pari a 0,441. Ipotizzando che R5 non sia ancora presente all'interno del campione da analizzare sarà aggiunta ad esso. Successivamente, si cercherà la recensione rilevante e contenente "sound price" da aggiungere al campione di recensioni che saranno analizzate manualmente. Essendo il topic lo stesso, l'insieme di recensioni rilevanti non cambia ma cambiano le recensioni rilevanti contenenti "sound price". Quest'ultimo, facendo riferimento alla figura 22, è composto dalle recensioni R2, R4, R7, R9, R10. Come precedentemente, le recensioni considerate saranno quelle che sono sia rilevanti per il topic 10 e che contengono l'insieme "sound price", ovvero R2,R4,R7,R9,R10. A questo punto, sarà selezionata la recensione con topical prevalence maggiore per il topic 10 e contenente l'insieme "sound price", cioè R2. Ipotizzando che R2 non sia stata ancora presente nel campione da analizzare sarà aggiunta ad esso.

Il secondo caso è quello in cui un insieme di parole appartenente allo stesso topic selezioni la stessa recensione da aggiungere al campione quindi, si ipotizzi ora che la presenza degli insiemi "sound headphone" e "sound price" all'interno delle recensioni in figura 21 sia descritta dalla figura 23 mentre le altre condizioni rimangano invariate (n=2, m=1).

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
"sound headphone"	1	1	0	0	1	0	1	1	0	1
"sound price"	0	1	0	1	1	0	1	0	1	1

Figura 23- Presenza/non presenza dei due insiemi di parole all'interno delle 10 recensioni di figura 21 (caso2)

In questo secondo caso, rimane tutto invariato rispetto al caso 1 fino all'aggiunta di R5 per l'insieme "sound headphone". Dopo aver aggiunto R5, si passa alla selezione di una recensione rilevante che contenga " sound price". L'insieme delle recensioni rilevanti per il topic 10 è sempre R1, R2, R3, R4, R5, R7, R9, R10 mentre l'insieme di recensioni che contiene "sound price" questa volta è R2, R4, R5, R7, R9, R10. In questo caso, quindi, le recensioni che rispettano entrambe le condizioni sono R2, R4, R5, R7, R9, R10. Selezionando ora la recensione che tra queste ha il valore maggiore di topic prevalence per il topic 10 si dovrebbe selezionare R5. Tuttavia, questa è già presente all'interno del campione di recensioni da analizzare per cui non sarà considerata e si considereranno le rimanenti. Di conseguenza, tra le recensioni candidate ad essere aggiunte al campione rimangono R2, R4, R7, R9, R10 e tra queste sarà scelta la recensione con valora maggiore di topical prevalence per il topic

10, ovvero R2. Ipotizzando che questa non sia ancora presente, sarà aggiunta al campione e di fatto è la seconda recensione con topic prevalence maggiore per il topic 10 che rispetta le condizioni. Altrimenti, se R2 fosse già presente, si ripeterà questa procedura: R2 sarà tolta dalle recensioni candidate a finire nel campione e sarà estratta la recensione con maggiore topical prevalence per il topic 10 tra R4, R7, R9, R10.

Il terzo caso, trattato nella stessa maniera del secondo, è il caso in cui una recensione è già stata selezionata per un insieme di parole di un topic e sarebbe selezionata anche per un insieme di parole appartenente ad un altro topic. Un esempio di questa situazione è rappresentato in figura 24 e le recensioni si farà riferimento sono di nuovo quelle della figura 21.

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
"sound headphone"	1	1	0	0	1	0	1	1	0	1
"sound price"	0	1	0	1	0	0	1	0	1	1
"Insieme 1 topic 2"	0	0	1	0	1	0	0	0	1	0
"Insieme 2 topic 2"	1	1	0	0	0	0	0	0	0	1

Figura 24- Presenza/non presenza dei due insiemi di parole all'interno delle 10 recensioni di figura 21 (caso 3)

Facendo riferimento alla situazione rappresentata in figura 24, si immagini di partire dall'aggiunta al campione delle recensioni appartenenti al topic 2.

Le recensioni rilevanti per il topic 2 sono R2 e R5. L' "Insieme 1 topic 2" è contenuto solo in R5 e questa recensione, ipotizzando che non sia ancora presente all'interno del campione sarà aggiunta. Successivamente, relativamente all' "Insieme 2 topic 2" sarà aggiunta R2, sempre ipotizzando che non sia ancora presente all'interno del campione. terminate le recensioni relative al topic 2, si procederà con l'aggiunta delle recensioni dei vari topic fino ad arrivare al topic 10. Arrivati a questo punto, per l'insieme "sound headphon" tra le recensioni che rispettano entrambe le condizioni ci sono R1, R2, R5, R7, R10 e si dovrebbe aggiungere al campione R5. Tuttavia, questa recensione è già presente all'interno del campione da analizzare poiché è già stata aggiunta relativamente all' "Insieme 2 topic 2". Di conseguenza, R5 sarà tolta dalle recensioni candidate a comporre il campione e sarà quindi selezionata la recensione rilevante per il topic 10 e relativa a "sound headphon" con topic prevalence maggiore per il topic 10 dopo R5, ovvero R1. Anche quest'ultima è già presente per cui si selezionerà la terza recensione tra R1, R2, R5, R7, R10 con topic prevalence maggiore, cioè R7. Per quanto riguarda, l'insieme "sound price, tra le recensioni rilevanti e contenenti l'insieme ci sono

R1, R4, R7, R9, R10. Tra queste la recensione con maggiore topical prevalence per il topic 10 è R1 ma è già presente all'interno del campione. Quindi, sarà selezionata la recensione tra R1, R4, R7, R9, R10 con la seconda topical prevalence maggiore, ovvero R4. Non essendo R4 ancora presente all'interno del campione, sarà aggiunta.

Riassumendo, l' "Insieme 1 topic 2" ha aggiunto al campione da analizzare la recensione R5, l' "Insieme 2 topic 2" ha aggiunto la recensione R2, "sound headphone" ha aggiunto la recensione R1 e "sound price ha aggiunto la recensione R4.

Una volta che la procedura avrà estratto le opportune recensioni da tutti i topic, sarà disponibile un sottoinsieme di recensioni rilevanti da cui un esperto potrà estrarre i requisiti manualmente, ovvero leggendo ogni recensione ed estraendone i requisiti. Un esempio di come si presenta l'output della procedura all'esperto è mostrato in tabella 4. Come si può osservare, sarà presente una lista di recensioni e per ognuna di queste c'è l'insieme di parole che l'ha selezionata.

Recensioni	Insieme di parole
I like that the headphones can fold up to save space.	"fold like"
Low volume, sound is very muffled and distant, not "clean" sound at all and a lot of statics, short cable so not suitable for computer. Returned.	"volum sound"
The Sony headphones are excellent for the price I paid for them, very good value for money. I would buy them again.	"good price"
Bought for kids to use at school. Held up thus far	"kid school"
I liked the sound quality of the headpones, but they seemed heavier than I thought and also they seemed tight. It would seem that they could have made the device a lot lighter. I will use this device, but I am already looking for an alternative. set.	"like sound"
It's great to having spare non bluetooth wired headphones I'm always wearing out Bluetooth earbuds. These wired headphones is very comfortable to wear. Great for meditation/journaling while listening to ocean sounds and classical music.	"headphon sound"

Tabella 4- Esempio del risultato finale

7 Confronto dei risultati ottenuti tramite estrazione manuale, estrazione tramite il metodo delle parole e estrazione tramite ChatGPT

Successivamente all'implementazione del metodo è stata misurata la capacità dell'insieme di parole di suggerire effettivamente l'esistenza dei requisiti all'interno delle recensioni rispetto ai requisiti estratti manualmente. Per rendere più fluida la lettura, l'estrazione di requisiti tramite l'utilizzo degli insiemi di parole sarà indentificata come l'estrazione tramite il metodo delle parole.

Al fine di eseguire questo confronto, sono stati estratti 3 campioni di recensioni del prodotto Sony ZX Series Wired On-Ear Headphones, Black MDR-ZX110, utilizzando la procedura descritta nella sezione precedente e facendo variare i parametri n ed m ma mantenendo la dimensione del campione costante, uguale a 104 recensioni. Inizialmente si è scelto di estrarre campioni composti da 100 recensioni ma con 13 topics il prodotto $n*m$ risultava decimale. Visto che n ed m sono entrambi interi, si è deciso di arrotondare per eccesso il prodotto $n*m$, ottenendo 8 recensioni estratte per topic. Queste 8 recensioni per topic sono state composte, rispetto ai parametri n ed m , arbitrariamente all'interno dei campioni nel modo seguente:

1. Campione 1: $n=4$, $m=2$;
2. Campione 2: $n=8$, $m=1$;
3. Campione 3: $n=1$, $m=8$.

Quindi, per ogni campione sono stati estratti i requisiti manualmente dalle singole recensioni e poi sono stati considerati come riferimento per valutare la capacità del metodo delle parole di suggerire la presenza dei requisiti estratti manualmente.

Dal confronto possono emergere 4 casi:

- VERI POSITIVI: rappresentano i requisiti assegnati alla recensione sia manualmente che dal metodo delle parole.
- FALSI POSITIVI: rappresentano i requisiti che sono stati assegnati alla recensione dal metodo delle parole ma che non sono stati assegnati alla recensione manualmente.
- VERI NEGATIVI: rappresentano i requisiti che non sono stati assegnati alla recensione né manualmente né dal metodo delle parole.
- FALSI NEGATIVI: rappresentano i requisiti che sono stati assegnati alla recensione manualmente ma che non sono stati assegnati dal metodo delle parole.

Veri positivi e veri negativi rappresentano i due casi in cui il metodo delle parole identifica correttamente i requisiti in quanto è in linea con l'assegnazione manuale. Invece, i due casi rimanenti rappresentano un disallineamento tra l'estrazione manuale dei requisiti e il metodo delle parole, ovvero una mancata capacità del metodo delle parole di assegnare correttamente i requisiti alla recensione.

Per poter definire il numero di veri positivi, veri negativi, falsi positivi e falsi negativi è stato considerato un insieme totale di 50 requisiti per il campione 1, 42 requisiti per il campione 2 e 27 per il campione 3 che rispettivamente sono i requisiti totali estratti manualmente all'interno del primo, del secondo e del terzo campione. Per ogni recensione, i requisiti totali estratti manualmente per il campione a cui appartiene sono stati distribuiti tra veri positivi, falsi positivi, veri negativi e falsi negativi. Ad esempio, per una recensione del primo campione sono stati distribuiti i 50 requisiti tra veri positivi, falsi positivi, veri negativi e falsi negativi a seconda dei requisiti che il metodo delle parole ha assegnato alla recensione. Per una recensione del campione 2 sono stati distribuiti sulle quattro categorie 42 requisiti e per una recensione del campione 3 sono stati distribuiti 27 requisiti.

La figura 25 mostra un esempio del risultato ottenuto a valle dell'assegnazione dei requisiti alle recensioni sia manualmente che dal metodo delle parole. Volendo spiegare come sono stati assegnati i requisiti di seguito saranno riportati alcuni esempi. Ad esempio, prendendo la prima recensione in figura 25 (*"I like that the headphones can fold up to save space"*) e simulando l'estrazione manuale è stato assegnato manualmente il requisito che "le cuffie devono potersi piegare". Allo stesso modo, il metodo delle parole ha a che fare con il fatto che devono potersi piegare.

Facendo riferimento alla quinta recensione in figura 25 (*"This is second pair I owned. My first pair last me 3 years and I only paid \$10.00 for it. I am hoping this one will last me at least 3 years as well :)"*) manualmente è stato estratto il requisito "Le cuffie devono essere resistenti", mentre il metodo manuale non suggerisce alcun requisito.

Facendo invece riferimento alla sesta recensione in figura 25 (*" Disposable headphones, right side stopped working after 4 months of use. Not worth the money paid."*) manualmente sono stati estratti i requisiti "Le cuffie devono essere resistenti", "Le cuffie devono garantire continuità di funzionamento da entrambi i lati", "Le cuffie devono valere il prezzo pagato". Invece, l'insieme di parole "headphone one" non suggerisce alcun requisito.

Un ultimo esempio di come sono stati assegnati i requisiti alla recensione manualmente e tramite il metodo delle parole si può fare considerando la settima recensione in figura 25 (*“May be good for a portable device but not good for a home system. The bass that I expected was not there at all and had a blaring sound even at a low volume. Treble tones were muffled and muddy.”*). Manualmente è stato assegnato il requisito *“Le cuffie devono suonare bene (qualità del suono)”*. Tramite il metodo delle parole sono stati assegnati i requisiti *“Le cuffie devono suonare bene (qualità del suono)”* e *“Le cuffie devono avere un range di volume che consenta volumi alti”* in quanto *“sound”* e *“volume”* hanno a che fare con questi due requisiti tra quelli trovati manualmente.

Recensioni	Insiemi di parole	Estrazione manuale	Metodo delle parole
Great quality headphones, sound great, fit great, and a great price too!	“sound headphon”	R5,R11,R12	R11
I like that the headphones can fold up to save space	“fold like”	R1	R1
Bought them for my 7 year old grand daughter for Christmas and she is enjoying them a lot.	old year	0	0
Great quality headphones, sound great, fit great, and a great price too!	“sound headphon”	R5,R11,R12	R11
This is second pair I owned. My first pair last me 3 years and I only paid \$10.00 for it. I am hoping this one will last me at least 3 years as well :)	“one pair”	R13	0
Disposable headphones, right side stopped working after 4 months of use. Not worth the money paid.	“headphon one”	R3,R13,R40	0
May be good for a portable device but not good for a home system. The bass that I expected was not there at all and had a blaring sound even at a low volume. Treble tones were muffled and muddy.	“volum sound”	R11	R11,R17
I really like these, they work well for wired headphones.	“headphon like”	0	0
Good quality headphones for a low price but Sony always produces good products.	“price headphon”	R3, R12	R3,R12

Figura 25- Esempio illustrativo di come è avvenuto il confronto. La colonna Estrazione manuale mostra i requisiti estratti manualmente. La colonna Metodo delle parole mostra i requisiti che gli insiemi di parole hanno permesso di individuare.

Una volta che sono stati assegnati i requisiti sia manualmente che con il metodo delle parole, è stato indicato per ogni recensione i veri positivi, i falsi positivi, i veri negativi e i falsi negativi considerando come totale dei requisiti assegnabili il totale dei requisiti estratti manualmente per quel campione. Un esempio esplicativo è mostrato in figura 26, considerando un totale di 50 requisiti estratti manualmente sull'intero campione e considerando le assegnazioni in figura 25. Ad esempio,

considerando la prima recensione in figura 25 a cui manualmente sono stati assegnati i requisiti R5,R11,R12 e il metodo delle parole ha assegnato solamente R11 risulta esserci un solo vero positivo (R11), due falsi negativi (R5,R12) e 47 veri negativi (i requisiti rimanenti). Un altro esempio può essere fatto considerando la terzultima recensione in tabella 3 a cui manualmente è stato assegnato R11 e il metodo delle parole le ha assegnato R11 ed R17. In questa situazione risulta esserci un vero positivo (R11), un falso positivo (R17), nessun falso negativo e 48 (il resto dei 50 requisiti) veri negativi. Il risultato complessivo di questa operazione relativo alle recensioni in tabella 3 è mostrato in figura 26.

VERI POSITIVI	FALSI POSITIVI	FALSO NEGATIVO	VERO NEGATIVO
1	0	2	47
1	0	0	49
0	0	0	50
1	0	2	47
0	0	1	49
0	0	3	47
1	1	0	48
0	0	0	50
2	0	0	48

Figura 26- Distribuzione dei veri positivi, falsi positivi, veri negativi e falsi positivi relativamente alla situazione mostrata in figura 25.

L'operazione di assegnazione dei veri positivi, falsi positivi, falsi negativi e veri negativi è stata eseguita per ogni recensione all'interno dei 3 campioni. Per ogni campione, a valle di questa assegnazione, sono stati sommati i valori di ogni categoria per poi successivamente calcolare alcune metriche al fine di comprendere la capacità del metodo delle parole di avere a che fare con i requisiti rispetto ai requisiti trovati manualmente. Come detto in precedenza, per ogni campione variava il numero totale dei requisiti estratti manualmente (50 per il primo, 42 per il secondo e 27 per il terzo) di conseguenza variava il numero dei requisiti assegnabili ad ogni recensione ed il totale.

I risultati ottenuti a valle di questa operazione sono mostrati in tabella 6 e sono stati ottenuti nel modo seguente:

- $VPT = \sum_{i=1}^{104} vp_i$, (7.1)

vp_i rappresenta il numero di veri positivi assegnati ad una generica recensione i del campione;

- $FPT = \sum_{i=1}^{104} fp_i$, (7.2)

fp_i rappresenta il numero di falsi positivi assegnati ad una generica recensione i del campione;

- $FNT = \sum_{i=1}^{104} fn_i,$ (7.3)

fn_i rappresenta il numero di falsi negativi assegnati ad una generica recensione i del campione;

- $VNT = \sum_{i=1}^{104} vn_i,$ (7.4)

vn_i rappresenta il numero di veri negativi assegnati ad una generica recensione i del campione.

	VPT	FPT	FNT	VNT
Campione 1	57	51	130	4962
Campione 2	59	51	131	4127
Campione 3	60	44	128	2576

*Tabella 5- Veri positivi totali, Falsi positivi totali, Falsi negativi totali, Veri negativi totali risultanti dal confronto tra estrazione dei requisiti manuale ed estrazione tramite il metodo delle parole. Ad esempio, essendoci 50 requisiti attribuibili per recensioni all'interno del primo campione, ci saranno 5200 (50*104) requisiti totali distribuiti sulle quattro casi di assegnazione per il campione 1. Stesso ragionamento vale per gli altri due campioni, considerando che nel secondo campione ci son 42 requisiti attribuibili per recensione e nel terzo campione 27.*

Infine, per ogni campione, è stata valutata la capacità indicare la presenza dei requisiti da parte del metodo delle parole attraverso il calcolo delle seguenti metriche (Costa e al., 2007; Franceschini e al., 2019; Maria Navin e Pankaja, 2016; Zaki e McColl-Kennedy, 2020; citati in Mastrogiacomo, Barravecchia, e Franceschini 2021):

- Accuratezza, è una misura dell'efficacia dell'estrazione dei requisiti con il metodo delle parole. Sono misurate in percentuale le assegnazioni corrette sul totale.

$$Accuratezza = \frac{\sum_{i=1}^{104} vp_i + \sum_{i=1}^{104} vn_i}{\sum_{i=1}^{104} vp_i + \sum_{i=1}^{104} fp_i + \sum_{i=1}^{104} vn_i + \sum_{i=1}^{104} fn_i} \quad (7.5)$$

- Precisione, misura la probabilità che l'assegnazione di un requisito ad una recensione con il metodo delle parole sia corretta.

$$Precisione = \frac{\sum_{i=1}^{104} vp_i}{\sum_{i=1}^{104} vp_i + \sum_{i=1}^{104} fp_i} \quad (7.6)$$

- Sensitività (o recall), misura la percentuale di requisiti assegnati dal metodo delle parole rispetto all'insieme dei requisiti assegnati manualmente.

$$Sensitività = \frac{\sum_{i=1}^{104} vp_i}{\sum_{i=1}^{104} vp_i + \sum_{i=1}^{104} fn_i} \quad (7.7)$$

- Fall-out, rappresenta la proporzione di requisiti assegnati erroneamente dal metodo delle parole sul totale dei requisiti non assegnati dal metodo delle parole, ovvero falsi positivi e veri negativi.

$$Fall - out = \frac{\sum_{i=1}^{104} fp_i}{\sum_{i=1}^{104} fp_i + \sum_{i=1}^{104} vn_i} \quad (7.8)$$

- Miss rate, rappresenta l'errore di seconda specie, ovvero la probabilità di accettare l'ipotesi nulla quando questa è falsa. In questo caso l'ipotesi nulla coincide con "il metodo non manuale non assegna i requisiti non assegnati dal metodo manuale".

$$Miss - rate = \frac{\sum_{i=1}^{104} fn_i}{\sum_{i=1}^{104} vp_i + \sum_{i=1}^{104} fn_i} \quad (7.9)$$

- Specificità, misura la percentuale di requisiti non assegnati correttamente dal metodo delle parole rispetto al totale dei requisiti non assegnati tramite estrazione manuale.

$$Specificità = \frac{\sum_{i=1}^{104} vn_i}{\sum_{i=1}^{104} fp_i + \sum_{i=1}^{104} vn_i} \quad (7.10)$$

- Negative predicted value misura la proporzione di requisiti non assegnati correttamente sull'insieme dei requisiti non assegnati dal metodo delle parole.

$$NPV = \frac{\sum_{i=1}^{104} vn_i}{\sum_{i=1}^{104} vn_i + \sum_{i=1}^{104} fn_i} \quad (7.11)$$

- False omission rate rappresenta la percentuale di requisiti non assegnati erroneamente dal metodo delle parole sul totale dei requisiti non assegnati.

$$FOR = \frac{\sum_{i=1}^{104} fn_i}{\sum_{i=1}^{104} vn_i + \sum_{i=1}^{104} fn_i} \quad (7.12)$$

- False discovery rate è la proporzione di requisiti assegnati correttamente rispetto al totale dei requisiti assegnati alla recensione dal metodo delle parole.

$$FDR = \frac{\sum_{i=1}^{104} fp_i}{\sum_{i=1}^{104} vp_i + \sum_{i=1}^{104} fp_i} \quad (7.13)$$

Queste metriche sono state calcolate su ciascun campione utilizzando i valori presenti in tabella 5 e sono stati ottenuti i risultati riportati in tabella 6.

	Accuratezza	Precisione	Recall	Fall-out	Miss-rate	Specificità	NPV	FOR	FDR
Campione 1	96,52%	52,78%	30,48%	1,02%	69,52%	98,98%	97,45%	2,55%	47,22%
Campione 2	95,83%	53,64%	31,05%	1,22%	68,95%	98,78%	96,92%	3,08%	46,36%
Campione 3	93,87%	57,69%	31,91%	1,68%	68,09%	98,32%	95,27%	4,73%	42,31%

Tabella 6- Valori delle metriche ottenuti nel confronto tra estrazione manuale ed estrazione tramite il metodo delle parole

Indicators	Range	Direction	Target values
Accuracy	[0;1]	High is good	>0.95
Recall	[0;1]	High is good	>0.70
Precision	[0;1]	High is good	>0.70
F ₁ score	[0;1]	High is good	>0.70
Miss-rate	[0;1]	Low is good	<0.20
Fall-out	[0;1]	Low is good	<0.05
Specificity	[0;1]	High is good	>0.90
False omission rate	[0;1]	Low is good	<0.05
False discovery rate	[0;1]	Low is good	<0.05
Negative predictive value	[0;1]	High is good	>0.90

Figura 27- Valori di riferimento delle metriche (Mastrogiacomo, Barravecchia, e Franceschini 2021)

7.1 Analisi dei risultati: estrazione manuale e metodo delle parole

Facendo riferimento alla figura 27 con i valori target, solamente accuratezza, specificità e NPV rispettano le soglie. Al contrario, precisione e recall sono abbondantemente lontani dai valori target. I valori dei singoli indicatori sono contrastanti. Questa contraddizione è dovuta a differenti ordini di grandezza nelle singole componenti degli indicatori. In particolare, osservando la tabella 6, si può immediatamente notare come il numero di veri negativi è di uno o due ordini di grandezza superiore rispetto alle altre quantità. Questo sbilanciamento è dovuto alla struttura del problema. In particolare, è dovuto alla grandezza dell'insieme di requisiti attribuibili a ciascuna recensione. I requisiti attribuiti ad ogni recensione sono pochi rispetto alla loro totalità e questo ha due conseguenze. La prima è che i falsi negativi sono anch'essi destinati ad essere pochi. La seconda è

che la maggior parte dei requisiti attribuibili è candidata a finire tra i falsi positivi o i veri negativi. Per quanto riguarda i falsi positivi, come sarà spiegato nei prossimi paragrafi, sono destinati ad essere pochi a causa dell'incapacità del metodo delle parole di suggerire tanti requisiti. Conseguentemente, la maggior parte dei requisiti è sempre destinata a finire nei veri negativi e la parte restante distribuito all'interno delle altre 3 categorie.

Questo fa sì che all'interno degli indicatori in cui sono considerati i veri negativi, i valori assunti da veri positivi, falsi positivi, falsi negativi e dalla loro somma, siano eclissati. Per questa ragione, si osservano valori di accuratezza, specificità ed NPV molto superiori alla soglia di accettabilità. Al contrario, precisione e recall sono molto lontani dalla loro rispettiva soglia di accettabilità, infatti nella loro computazione i veri negativi non sono considerati. Di conseguenza, avendo i componenti degli indicatori lo stesso ordine di grandezza all'interno di precisione e recall, questi raccontano una storia diversa e rispetto alla capacità del metodo delle parole di suggerire i requisiti.

Dai singoli valori assunti da precisione e recall si può dire che:

- tra il 50% e il 60% dei requisiti assegnati dal metodo delle parole alla recensione è corretto (precisione);
- il metodo delle parole è in grado di individuare circa il 31% dei requisiti trovati con il metodo manuale (recall).

7.2 Estrazione dei requisiti con ChatGPT e confronto con estrazione manuale

I requisiti estratti manualmente sono stati utilizzati come riferimento anche per valutare la capacità di ChatGPT di estrazione dei requisiti in modo da poter successivamente mettere a confronto la capacità di estrarre i requisiti tramite l'utilizzo di un modello di machine learning avanzato, ChatGPT, e tramite il metodo delle parole.

In questo caso, è stato chiesto a ChatGPT di estrarre i requisiti del prodotto dalle recensioni presenti nei campioni. L'estrazione, anche in questo caso, è stata effettuata singolarmente per ciascuna recensione utilizzando sempre una nuova chat in modo che lo strumento non avesse dei riferimenti che potessero condizionare la risposta e ponendo la richiesta di estrazione sempre allo stesso modo. Un esempio esplicativo di come stata posta avvenuta la richiesta di estrazione dei requisiti delle recensioni a ChatGPT è mostrata in figura 28. In particolare, nella figura 28, è stato chiesto a Chat GPT di estrarre i requisiti di prodotto dalla quinta recensione presente in figura 25.

ST

Extract product requirements from the following review related to a pair of headphone: "This is second pair I owned. My first pair last me 3 years and I only paid \$10.00 for it. I am hoping this one will last me at least 3 years as well :)"



The review provides some information about the user's experience with a pair of headphones. From the review, we can extract the following potential product requirements:

1. **Durability:** The user mentions that their first pair lasted for 3 years. This suggests a requirement for durability in the headphones.
2. **Longevity:** The user expresses a desire for the second pair to last at least 3 years as well. This implies a product requirement for a long lifespan.
3. **Cost-Effectiveness:** The user highlights that they only paid \$10.00 for their first pair. This suggests a requirement for cost-effectiveness or affordability.

These extracted requirements indicate that the user values headphones that are durable, have a long lifespan, and are reasonably priced.

Figura 28- Esempio di richiesta di estrazione dei requisiti a ChatGPT e output ottenuto.

Una volta che è stato chiesto Chat GPT di estrarre i requisiti da tutte le recensioni di tutti e tre i campioni sono stati assegnati ad ogni recensione i veri positivi, i falsi positivi, i falsi negativi ed i veri negativi, come è stato fatto precedentemente nel confronto tra l'estrazione manuale ed il metodo delle parole. A valle di questa assegnazione i valori sono stati aggregati per campione al fine di calcolare le metriche utilizzate anche nel confronto precedente.

Essendo i campioni gli stessi, ad ogni recensione si potevano assegnare alle categorie veri positivi, falsi positivi, veri negativi e falsi negativi 50 requisiti per il campione 1, 42 requisiti per il campione 2 e 27 requisiti per il campione 3. Di conseguenza per il campione uno si possono osservare 5200 (50 requisiti attribuibili *104 recensioni) requisiti distribuiti tra veri positivi, falsi positivi, veri negativi e falsi negativi, 4368 (42*104) per il campione 2 e 2808 (27*104). I risultati ottenuti dall'aggregazione dei valori per ogni categoria per ogni campione sono mostrati tabella 7. Dai dati presenti in tabella 7 sono stati calcolati i valori delle metriche mostrati in tabella 8.

	VPT	FPT	FNT	VNT
Campione 1	166	113	21	4900
Campione 2	175	89	15	4089
Campione 3	174	74	14	2546

Tabella 7- Veri positivi, Falsi positivi, Falsi negativi, Veri negativi risultanti dal confronto tra estrazione dei requisiti manuale e tramite ChatGPT

	Accuratezza	Precisione	Recall	Fall-out	Miss-rate	Specificità	NPV	FOR	FDR
Campione 1	97,42%	59,50%	88,77%	2,25%	11,23%	97,75%	99,57%	0,43%	40,50%
Campione 2	97,62%	66,29%	92,11%	2,13%	7,89%	97,87%	99,63%	0,37%	33,71%
Campione 3	96,87%	70,16%	92,55%	2,82%	7,45%	97,18%	99,45%	0,55%	29,84%

Tabella 8- Valori delle metriche ottenuti nel confronto tra estrazione dei requisiti manuale ed tramite ChatGPT

7.2.1 Analisi dei risultati: metodo manuale e ChatGPT

A differenza del metodo delle parole, ChatGPT presenta risultati migliori, infatti, l'output risulta essere molto più allineato con l'output prodotto manualmente.

Anche in questo caso a causa della dimensione dell'insieme dei requisiti attribuibili, i veri negativi sono di uno, alcune volte di due ordini di grandezza superiori rispetto ai valori delle altre categorie e di conseguenza i valori di accuratezza, specificità e NPV sono superiori ai valori soglia. Tuttavia, i valori di precisione e recall, a differenza del caso precedente, risultano essere più coerenti con accuratezza, specificità ed NPV. Inoltre, l'indicatore recall è ampiamente oltre la soglia di 0.7 ed indica proprio la grande capacità di ChatGPT di individuare una buona parte dei requisiti che l'umano trova all'interno delle recensioni. La precisione, invece, non è sopra la soglia nei primi due campioni mentre è poco più grande nel terzo, tuttavia, è comunque migliore rispetto alla precisione del metodo delle parole.

7.3 Confronto tra ChatGPT e metodo delle parole

Dopo aver valutato la capacità di estrazione dei requisiti del metodo delle parole e di Chat GPT rispetto all'estrazione manuale, sono stati messe a confronto le metriche ottenute coi l'utilizzo dei due metodi in tabella 9.

Come già detto in precedenza, Chat GPT presenta complessivamente risultati migliori rispetto al metodo delle parole. L'unica metrica in cui il metodo delle parole produce risultati di poco migliori è

la specificità. Il metodo delle parole ha quindi una miglior capacità rispetto a ChatGPT di non attribuire correttamente i requisiti ad una recensione.

ChatGPT
Metodo delle parole

Legenda

	Accuratezza	Precisione	Recall	Fall-out	Miss-rate	Specificità	NPV	FOR	FDR
Campione 1	97,42%	59,50%	88,77%	2,25%	11,23%	97,75%	99,57%	0,43%	40,50%
	96,52%	52,78%	30,48%	1,02%	69,52%	98,98%	97,45%	2,55%	47,22%
Campione 2	97,62%	66,29%	92,11%	2,13%	7,89%	97,87%	99,63%	0,37%	33,71%
	95,83%	53,64%	31,05%	1,22%	68,95%	98,78%	96,92%	3,08%	46,36%
Campione 3	96,87%	70,16%	92,55%	2,82%	7,45%	97,18%	99,45%	0,55%	29,84%
	93,87%	57,69%	31,91%	1,68%	68,09%	98,32%	95,27%	4,73%	42,31%

Tabella 9- Confronto tra le metriche ottenute tramite l'utilizzo di ChatGPT e tramite l'utilizzo del metodo delle parole. Per ogni cella, il primo valore si riferisce al valore della metrica di ChatGPT e il secondo al valore della metrica ottenuto tramite il metodo delle parole. Prendendo ad esempio la cella relativa all'accuratezza del campione 1: 97,42 fa riferimento all'accuratezza di ChatGPT e96,52 all'accuratezza del metodo delle parole.

Vale la pena però osservare più da vicino i valori di questi indicatori, andando ad osservare le loro componenti e come queste componenti sono distribuiti nei diversi campioni con i due diversi metodi.

Dalle tabelle 5 e 7 si può notare come i valori di veri positivi, falsi positivi e falsi negativi sono distribuiti diversamente nei due metodi. Tuttavia, osservando le distribuzioni di questi tre valori all'interno dei campioni di ciascun metodo, queste mostrano distribuzioni simili (tabelle 8 e 6). L'utilizzo del metodo delle parole, facendo riferimento alla tabella 6, presenta una distribuzione in cui i falsi negativi sono preponderanti (mediamente 129 a campione), mentre i veri positivi e i falsi positivi presentano una numerosità simile (mediamente 59 veri positivi e 49 falsi positivi a campione). Contrariamente, ChatGPT, facendo riferimento alla tabella 7, presenta una distribuzione in cui sono preponderanti i veri positivi, (mediamente 172 a campione) i falsi positivi sono inferiori ai veri positivi (mediamente 92 a campione) ed infine i falsi negativi (17 a campione in media) che sono i meno numerosi.

Queste differenti distribuzioni dei valori dei veri positivi, falsi positivi e falsi negativi, rispetto ai due metodi confrontati si riflettono all'interno dei valori delle metriche di validazione, in particolare nella precisione e nella recall in quanto non sono considerati i veri negativi a differenza di accuratezza, specificità ed NPV. Facendo riferimento alla tabella 9, si può osservare che i valori della precisione sono superiori per ChatGPT in tutti e tre i campioni poiché nonostante in termini assoluti produca più falsi positivi è anche in grado di trovare un numero requisiti veri positivi che compensa il maggior numero di falsi positivi. Tuttavia, solamente nell'ultimo campione il valore della metrica precisione è di poco superiore alla soglia di accettabilità mentre è inferiore nel campione 1 e nel campione 2. Invece, facendo riferimento sempre alla tabella 9, per quanto riguarda l'indicatore recall ChatGPT surclassa il metodo delle parole, infatti a differenza di quest'ultimo è anche abbondantemente sopra la soglia di accettabilità. Per esempio, basandosi sull'indicatore recall, fatto 100 i requisiti attribuiti dall'umano ad una recensione, ci si aspetta che ChatGPT ne attribuisca correttamente 90 mentre con il metodo delle parole ci si aspetta suggerisca l'esistenza di solamente 30 requisiti sul totale dei 100 attribuiti manualmente.

La differente distribuzione di veri positivi, falsi positivi e falsi negativi nei due metodi può essere dovuta alle differenti modalità di funzionamento.

Ad esempio, la dimensione dell'insieme di parole può condizionare l'insieme di requisiti con cui può avere a che fare in quanto più sono le parole all'interno dell'insieme e più sono i requisiti a cui queste possono fare riferimento. Tuttavia, per come è stato costruito l'algoritmo difficilmente l'insieme di parole sarà costituito da più di due parole a causa del fatto che aumentando le parole nell'insieme si aggiungono vincoli nel calcolo della proporzione di recensioni rilevanti in cui questo è presente. Di conseguenza, saranno privilegiati insiemi di parole più piccoli. In particolare, nel presente lavoro gli insiemi di parole a cui sono state legate le recensioni rilevanti erano composti tutti da due parole.

Una diretta conseguenza all'uso di insieme di parole composto da poche parole è che non può suggerire molti requisiti. Da questo, molto probabilmente, deriva il basso numero di veri positivi, dal momento in cui nelle recensioni più lunghe sono presenti più requisiti anche non legati all'insieme di parole. Inoltre, la ridotta dimensione dell'insieme di parole potrebbe anche essere causa del basso numero di falsi positivi, in quanto anche se l'insieme di parole indica requisiti che non sono presenti nella recensione questi sono comunque pochi per la stessa ragione per cui sono pochi i veri positivi.

Al contrario, ChatGPT, integrando tecniche statistiche e di machine learning più avanzate, riesce a contestualizzare bene e a cogliere le sfumature della recensione rilevando una buona parte dei

requisiti che trova anche un umano. Conseguentemente i falsi negativi sono in numero ridotto. Tuttavia, per qualche ragione, propone anche requisiti che l'umano non trova infatti ha un numero di falsi positivi piuttosto elevato.

8 Conclusione

A valle del lavoro svolto, è possibile rendersi conto di quanto si è ancora lontani dall'aver strumenti in grado di sfruttare totalmente l'informazione presente sul web per l'estrazione dei requisiti di un prodotto. Gli algoritmi di machine learning che utilizzano tecniche di text mining sono sicuramente lo strumento principale che si ha a disposizione per l'analisi di grandi quantità di dati in forma testuale ma non sono ancora in grado di fornire risultati che siano vicini all'idealità.

Un risultato inaspettato è stato l'elevato numero di requisiti proposti da ChatGPT non assegnati alle recensioni dall'umano. Benché fosse la versione gratuita, ChatGPT dovrebbe essere in grado di analizzare semplici dati in input e generare una risposta coerente e il più precise possibili rispetto alla richiesta. Tuttavia, nonostante generasse contenuto coerente, le risposte alla richiesta di estrazione dei requisiti dalle singole recensioni avevano tra il 30 e il 40% del contenuto "inventato". Per questa ragione, non è possibile individuare quali siano effettivamente i requisiti presenti all'interno di una recensione e quali invece sono stati "inventati". Di conseguenza, utilizzare ChatGPT per l'estrazione dei requisiti dalle recensioni è molto rischioso poiché, sia nel caso dello sviluppo di un nuovo prodotto/servizio sia nel caso di miglioramento di un prodotto/servizio già sul mercato, il rischio di investire nello sviluppo di caratteristiche tecniche che consentono il soddisfacimento di falsi requisiti è molto alto. Tuttavia, c'è anche da considerare che ChatGPT non è stato pensato per questo fine. A valle di queste considerazioni si può concludere che utilizzare ChatGPT (versione gratuita) per l'estrazione dei requisiti non conviene a causa della difficoltà con cui non si riescono a distinguere i veri requisiti da quelli falsi.

Per quanto riguarda il metodo proposto, invece, si può dire che l'utilizzo di insiemi di parole come indicatore della presenza di requisiti non è una strada promettente in quanto, osservando i valori delle metriche, le prestazioni sono pessime. Per come è stata costruita la procedura proposta, se si vuole analizzare un campione di dimensione limitata, difficilmente gli insiemi di parole saranno composti da più di due parole. Di conseguenza, i requisiti con cui può avere a che fare un insieme di parole composto da due parole sono veramente pochi, come spiegato nella sezione precedente. Tuttavia, tale procedura prevede l'analisi manuale delle recensioni del campione e ciò consente di considerare tutti i vari aspetti presenti all'interno di una recensione, indipendentemente dal fatto che questi siano legati o meno all'insieme di parole. Inoltre, sono considerate recensioni che coprono tutti i topics avendo così la possibilità di estrarre i requisiti anche relativi ad alcuni

determinanti la qualità che magari mediante l'utilizzo di questionari o interviste non sarebbe stato possibile in quanto sottovalutati dal soggetto che ha progettato le domande.

Tuttavia, quello che ancora è da conoscere è se dall'insieme di recensioni estratte attraverso questa procedura è possibile ottenere una buona parte dei requisiti che si estrarrebbero se si analizzasse l'intero insieme di recensioni o almeno i requisiti più importanti. Inoltre, ci sarebbe anche da stabilire come i parametri n ed m influenzano la creazione di un campione che sia esaustivo dal punto di vista dei requisiti estratti. Una strada alternativa alla procedura proposta potrebbe essere quella di far scegliere a colui interessato all'estrazione dei requisiti gli insiemi di parole considerati da lui rilevanti e comporre così il campione con le recensioni contenenti gli insiemi scelti, invece di far scegliere la dimensione del campione da analizzare in funzione dei parametri n ed m . In questo modo, però, bisogna essere consapevoli che, se si scelgono troppi insiemi di parole, si rischia di analizzare un campione di recensioni di elevata numerosità e l'utilizzo del metodo potrebbe perdere la sua funzione, ovvero quella di consentire l'analisi un numero ridotto di recensioni al posto della loro totalità.

Osservando alle caratteristiche del metodo proposto, ovvero la rapidità con cui si possono ottenere i requisiti di prodotto e l'incertezza legata all'esaustività dei requisiti trovati questo può trovare applicazione in contesti in cui la ricerca di mercato deve avvenire rapidamente e/o con la possibilità di adattare il prodotto in tempi futuri senza incorrere in costi troppo elevati. Un esempio può essere il caso della progettazione di un MVP, in cui l'incertezza legata al mercato è molto alta e di conseguenza è difficile comprendere come progettare il prodotto. In questo caso, il metodo proposto potrebbe essere utilizzato per estrarre i requisiti da prodotti concorrenti ed effettuare un'analisi della concorrenza al fine di avere un'idea su quali aspetti i prodotti concorrenti sono da migliorare e su quali, invece, sono sufficientemente apprezzati. Una volta estratti i requisiti dai prodotti concorrenti, è possibile comprendere meglio come posizionare il proprio prodotto e magari quali caratteristiche incorporare per soddisfare i bisogni primari. Inoltre, oltre ad analizzare la concorrenza in tempi brevi, potrebbe aiutare eventualmente ad indirizzare con più precisione il target di eventuali questionari ed interviste utilizzati per comprendere le caratteristiche dell'MVP in quanto, avendo acquisito conoscenza sulla concorrenza, potrebbe rendere più chiaro a quale target di mercato rivolgersi.

Infine, come è stato anticipato nell'introduzione, la procedura proposta può essere utilizzata per l'estrazione dei requisiti al fine della costruzione del QFD di un prodotto. Dal momento che analizza

recensioni relative a prodotti già presenti sul mercato i requisiti estratti sono requisiti che possono essere utilizzati per la costruzione di un QFD che ha lo scopo di apportare miglioramenti.

9 Bibliografia e sitografia

Bibliografia

Barravecchia, F., Masrogiacomo, L. e Franceschini, F. (2020), *“Categorizing Quality Determinants in Mining User-Generated Contents”*, Sustainability 2020, 12(23), 9944, doi: 10.3390/su12239944

Mastrogiacomo, L., Barravecchia, F., Franceschini, F. e Marimon, F. (2021), *“Mining quality determinants of product-service systems from user-generated contents”*, Quality Engineering, Vol. 33, pp. 425-442, doi: 10.1080/08982112.2021.1877305

Barravecchia, F., Masrogiacomo, L. e Franceschini, F. (2021), *“Digital voice-of-customer processing by topic modeling algorithms: insights to validate empirical results”*, International Journal of Quality & Reliability Management, Vol. 39 No. 6, pp. 1453-1470, doi: 10.1108/IJQRM-07-2021-0217

Barravecchia, F., Masrogiacomo, L. e Franceschini, F. (2022), *“KA-VoC Map: Classifying product Key-Attributes from digital Voice-of-Customer”*, Quality Engineering, Vol. 34, pp. 344-358, doi: 10.1080/08982112.2022.2057805

Sitografia

<https://www.ibm.com/it-it/topics/text-mining>

https://en.wikipedia.org/wiki/Text_mining

<https://www.geeksforgeeks.org/text-mining-in-data-mining/>

https://blog.osservatori.net/it_it/chatgpt-come-funziona-cosa-puo-fare-limiti-opportunita

<https://www.partitaiva.it/chatgpt/>

<https://www.ilmioposizionamento.it/chatgpt-caratteristiche-funzionamento/>

