Politecnico di Torino

Dipartimento di Ingegneria Gestionale e della Produzione

Corso di Laurea Magistrale
in Ingegneria Gestionale

Tesi di Laurea Magistrale

Predictive Analytics applicati alla Digital Voice of Customer

Relatore:
Federico Barravecchia

Correlatore:
Luca Mastrogiacomo

Candidato:
Noemi Collazzo
Matricola s314664

Anno accademico 2024/2025

Sommario

Indice figure

Indice tabelle

Introduzione

Con l'avvento di Internet, i clienti esprimono liberamente la loro opinione su un prodotto o servizio attraverso contenuti pubblicati su social media, forum e blog [1]. Tali contenuti prendono il nome di *"Digital Voice of Customer"* (Digital VoC) e rappresentano un mezzo economico, utile alle aziende, per comprendere le aspettative e le percezioni dei clienti nei riguardi di un prodotto o servizio offerto [1].

Sebbene tali dati siano presenti in grandi quantità e facilmente reperibili online, sono non strutturati e per questo la loro analisi richiede l'applicazione di opportuni algoritmi detti *topic modelling*, che permettono di estrapolare gli argomenti rilevanti all'interno del corpus dei documenti [3]. Tali argomenti vengono identificati come le determinanti di qualità del prodotto o servizio offerto di cui si vuole monitorare la qualità nel tempo [3].

Rispetto ai metodi tradizionali di monitoraggio della qualità, come, ad esempio, i questionari e le interviste, l'analisi della Digital VoC supera i limiti di tali metodi, i quali richiedono tempo e risorse per ottenere informazioni circa la soddisfazione o insoddisfazione di un campione limitato di clienti [2].

La domanda di ricerca su cui si sviluppa la tesi è:*" Come è possibile, partendo dai risultati dei topic modelling applicati alla Digital VoC, condurre predictive analytics per migliorare il monitoraggio della qualità e la soddisfazione del cliente?"*.

Le aziende hanno accesso ad una grande mole di dati, e per questo risulta essenziale ottenere informazioni rilevanti e condurre analisi predittive sui risultati ottenuti per prendere decisioni future volte al miglioramento della qualità del prodotto o servizio.

Analizzando la letteratura attuale, ci sono studi che partendo dai risultati dei topic modelling, valutano la tendenza della discussione media dei topic (crescente, decrescente o stazionaria) [3], ma non applicano modelli sofisticati per studiare la serie ed apprendere il suo andamento in modo da condurre accurate previsione future.

Una volta ottenuti i risultati dell'algoritmo di topic modelling, si valuta come varia l'"*Interval Mean Topical Prevalence"*(IMTP) degli argomenti rilevanti estratti, ovvero quanto viene discusso mediamente ciascun argomento o topic in un istante temporale specifico [3].

L'obiettivo è di addestrare modelli in grado di apprendere le relazioni sottese alla serie temporale e condurre predictive analytics. L'analisi è stata applicata a due casi studio: Disneyland e Ryanair.

Dopo una panoramica sui modelli che sono attualmente impiegati per condurre analisi predittive su serie storiche, si è scelto, di utilizzare nelle analisi:

i. *"Seasonal Auto-Regressive Integrated Moving Average"* (SARIMA) [23] e
ii. *"Long Short-Term Memory"* (LSTM) [18].

Tali modelli sono attualmente applicati dalle aziende, ma in ambiti differenti rispetto al customer service, come, ad esempio, nelle previsioni del prezzo delle azioni o meteorologiche [42-44].

Il seguente elaborato di tesi mira a colmare il gap presente nella letteratura attuale, proponendo una procedura che il management della qualità può applicare per valutare la discussione nel tempo degli argomenti nelle recensioni.

In questo modo le aziende, utilizzando strumenti capaci di anticipare con una certa accuratezza eventuali andamenti futuri, sono competitive sul mercato ed offrono un prodotto o servizio che si adatta alle richieste del cliente.

L'elaborato è strutturato in cinque capitoli:

1. Capitolo 1 – Analizza le metodologie tradizionali per il monitoraggio della qualità e il valore aggiunto apportato dall'analisi della Digital VoC. Introduce il concetto di *topic model* e descrive l'algoritmo "*Structural Topic Model*" (STM), utilizzato per individuare i temi più discussi nelle recensioni [1].

2. Capitolo 2 – Presenta una panoramica dei modelli predittivi applicabili alle serie storiche e la scelta dei modelli LSTM e SARIMA per l'analisi dell'IMTP.

3. Capitolo 3 – Descrive le analisi condotte sul caso studio Disneyland. Dopo aver applicato STM, vengono testati i modelli predittivi sull'IMTP di tre topic selezionati e valutata l'accuratezza delle previsioni.

4. Capitolo 4 – Descrive le analisi condotte sul caso studio Ryanair. Dopo aver applicato STM, vengono testati i modelli predittivi sull'IMTP di tre topic selezionati e valutata l'accuratezza delle previsioni.

5. Capitolo 5 – Riassume i principali risultati della ricerca, evidenziando come il modello LSTM abbia mostrato una maggiore capacità di catturare l'andamento della serie temporale rispetto SARIMA. Inoltre, vengono proposte due possibili soluzioni

applicative per supportare il management aziendale nel monitoraggio della qualità del prodotto o servizio.

Attraverso questo studio, si intende fornire un contributo concreto all'integrazione delle tecniche di *machine learning* nei processi di gestione della qualità, dimostrando come l'analisi predittiva dei temi discussi estrapolati della Digital VoC possa diventare uno strumento strategico per migliorare l'esperienza del cliente e l'efficienza aziendale.

1 Capitolo 1: revisione della letteratura

1.1 Analisi della Digital VoC

La gestione della qualità ha subito un'evoluzione nel tempo e come per l'industria si è avuta la quarta rivoluzione industriale, così anche per la qualità si parla di "Qualità 4.0" [3]. La digitalizzazione aziendale ed i "big data" hanno aperto nuove prospettive per la gestione della qualità di prodotti e servizi [1]. Le aziende, negli ultimi anni hanno iniziato a comprendere l'importanza di tali dati e per questo hanno introdotto nuovi metodi per gestirli e per ricavare informazioni importanti per la gestione della qualità di prodotti e servizi [3].

La "Digital Voice of Costumer"(Digital VoC) può essere definita come l'insieme di opinioni, recensioni o feedback su prodotti o servizi che i consumatori pubblicano liberamente su piattaforme digitali accessibili al pubblico [2]. Sebbene tali contenuti possano essere trovati in una molteplicità di formati, come foto o video, la maggior parte è composta da testi non strutturati pubblicati su blog, forum, social network o piattaforme di e-commerce [2].

La domanda che si sono poste le aziende ed i ricercatori negli ultimi anni è stata: "*Come è possibile utilizzare la Digital VoC per implementare un sistema di monitoraggio della qualità di un prodotto o servizio?"* [2].

Il problema principale quando si tratta di utilizzare la Digital VoC per il monitoraggio della qualità è che è spesso costituito da testi non strutturati [2]. I documenti non strutturati non hanno un formato predefinito e per questo contengono dati provenienti da diversi siti di recensioni o social media, ma anche video o immagini [5]. I dati strutturati, invece, presentano un formato standardizzato rendendone semplice l'analisi [5]. Per indagare le percezioni dei clienti e valutare la qualità di un prodotto o servizio è necessario applicare delle tecniche dette di text mining, in particolare gli algoritmi di topic modelling, capaci di estrapolare gli argomenti maggiormente discussi dal corpus testuale [2].

Le aziende, tradizionalmente, utilizzano una varietà di metodi per monitorare l'evoluzione della qualità dei loro prodotti e servizi nel tempo, identificare anomalie o criticità ed individuare possibili aree di miglioramento [1].

I metodi più utilizzati per la raccolta della voce dei clienti sono:

Metodi tradizionali per il monitoraggio della qualità	Descrizione
Interviste	Sono utilizzate quando è necessario raccogliere informazioni approfondite su un particolare cliente. Possono essere strutturate, semi-strutturate o non strutturate nel caso in cui non ci sia uno schema di domande preciso da seguire.
Focus Group	Intervista di gruppo in cui si valutano sia le risposte tra i partecipanti che l'interazione tra essi.
Questionari	Permette di raccogliere dati da un numero abbastanza elevato di clienti con una struttura che presenta sia domande chiuse che aperte.

Tabella 1 Descrizione dei metodi tradizionali di monitoraggio della qualità (Adattato [4])

Ciascun metodo presenta dei vantaggi e degli svantaggi nell'applicazione:

Metodi tradizionali per il monitoraggio della qualità	Vantaggi	Svantaggi
Interviste	• Strumento flessibile che consente di interagire con il cliente.	• Complesse da analizzare e tradurre i risultati raccolti. • Rivolta ad un singolo intervistato.
Focus Group	• Strumento flessibile capace di investigare un argomento in profondità.	• Complesso da organizzare ed analizzare i dati raccolti. • Rivolto ad un gruppo ristretto di clienti.
Questionari	• È un metodo per condurre analisi strutturate su una tematica di ricerca e permette di raggiungere un numero abbastanza elevato di clienti.	• Richiede tempo e risorse per essere strutturato opportunatamente, in modo da non essere percepito come troppo lungo dal cliente.

Tabella 2 Vantaggi e svantaggi dei metodi tradizionali di monitoraggio della qualità (Adattato [4])

Per monitorare la qualità nel tempo viene raccolta la voce del cliente in diversi istanti temporali [6]:

i. valutazione post-acquisto del prodotto/servizio, richiede ai consumatori la valutazione di un bene o un servizio acquistato solo dopo averlo utilizzato;

ii. sondaggi periodici consistono in questionari o interviste per ottenere informazioni periodiche dai clienti;
iii. monitoraggio continuo della qualità questionari o interviste sottoposte ai clienti in maniera continua.

Tali metodi permettono di indagare la soddisfazione o insoddisfazione del cliente nei riguardi degli attributi chiave o determinanti di qualità di un prodotto o servizio [2].

La raccolta e l'analisi di Digital VoC, però, è un metodo più affidabile rispetto ai tradizionali per comprendere il feedback dei clienti, in quanto, analizza direttamente le informazioni auto rilasciate sulla loro esperienza riguardo prodotti e servizi [2].

Si può pensare di valutare lo stesso servizio utilizzando un questionario strutturato o l'analisi della Digital VoC e fare un paragone tra i due metodi [2].

La Tabella 3 riassume i pro e contro dell'utilizzo dei distinti metodi di analisi.

Monitoraggio della qualità	**Vantaggi**	**Svantaggi**
Questionario	• Dati strutturati	• Tempo e risorse per essere opportunamente strutturato • Campione limitato di clienti
Digital VoC	• Contenuti accessibili online e gratuiti • Campione illimitato di clienti • Flusso di informazioni continuo	• Dati non strutturati

Tabella 3 Pro e contro dell'utilizzo dei questionari e della Digital VoC per monitorare la qualità di un prodotto o servizio (Adattato e tradotto [2-3])

I limiti riscontrati nel monitoraggio tradizionale della qualità sono ampiamente superati dall'analisi della Digital VoC, in quanto [2]:

- il flusso di informazioni è continuo;
- non risultano strumenti intrusivi come i questionari poiché i clienti esprimono liberamente la loro opinione sul prodotto o servizio e

- sono contenuti gratuiti e facilmente accessibili.

La Tabella 4 mostra un confronto schematico tra i metodi tradizionali di monitoraggio della qualità e il monitoraggio basato sull'analisi della Digital VoC [2].

	Monitoraggio tradizionale della qualità	**Monitoraggio della qualità basato sulla Digital VoC**
Fonte delle informazioni	Interviste, focus group, questionari strutturati.	Recensioni dei clienti, post pubblicati sui social media, forum.
Cosa viene valutato?	Caratteristiche di prodotto o servizi.	Determinanti latenti di qualità.
Variabile previste	Le variabili chiave misurate sono considerate note.	Le determinanti latenti di qualità sono estratte dalla Digital VoCs
Focus	Identificare le caratteristiche critiche di prodotti/servizi per ottenere miglioramenti.	Identificazione delle determinanti della qualità maggiormente percepiti dai clienti per guidare le attività di progettazione e miglioramento continuo.
Aggiornamento delle informazioni	Periodiche e guidate.	Costanti e non guidate.
Tecniche di analisi	Analisi statistica dei dati.	• Text mining • Analisi statistica dei dati

Tabella 4 Confronto tra il metodo tradizionale di monitoraggio della qualità e il monitoraggio basato sulla Digital VoC (Adattato e Tradotto [2])

I miglioramenti nelle tecniche di text mining, come verrà approfondito nel prossimo paragrafo, hanno reso possibile l'utilizzo di strumenti innovativi per estrarre gli argomenti latenti dalla Digital VoC [2].

1.2 Text mining

Il text mining è una branca del data mining che comprende algoritmi che permettono di analizzare documenti non strutturati, per estrarre da quest'ultimi informazioni rilevanti per il management aziendale [3].

Il text mining è una tecnica che generalmente consiste di cinque fasi [3]:

(1) collezione di documenti,

(2) preparazione e selezione dei dati,

(3) estrazione delle informazioni,
(4) valutazione ed
(5) interpretazione dei risultati.

L'estrazione delle informazioni dai documenti testuali può avvenire utilizzando diverse tecniche di text mining [5].

Tecniche di text mining	Descrizione
Analisi del sentiment	Classificazione dei contenuti testuali in base alle percezioni rilevate dei clienti.
Modellazione degli argomenti	Rileva gli argomenti più discussi all'interno della raccolta dei testi.
Riconoscimento delle entità	Estrae le informazioni rilevanti dei testi e le classifica in categorie.
Classificazione del testo	Suddivisione dei testi in categorie predefinite.
Estrazione delle regole di associazione	Indaga le relazioni o associazioni presenti nei contenuti analizzati.

Tabella 5 Tecniche di text mining (Adattato [5])

In base all'obiettivo dello studio di estrarre gli argomenti latenti dal testo delle recensioni auto rilasciate dai clienti, il modello scelto per l'analisi è la modellazione degli argomenti, ovvero il "topic modelling" [3].

1.2.1 Topic modelling

All'interno della famiglia degli algoritmi di text mining il topic modelling è una tecnica di modellazione statistica in grado di estrarre da testi non strutturati gli argomenti discussi [2]. I topic modelling sono modelli di apprendimento che permettono di analizzare grandi moli di dati in modo semplice e facilmente replicabile [2].

I dati non strutturati presi in input dal modello vengono elaborati come segue [2]:

I. il corpus dei documenti viene analizzato e generato il vocabolario che contiene tutte le parole uniche presenti all'interno dei documenti;

II. ciascun documento viene rappresentato come un "bag of words", ovvero un elenco di parole presenti nel vocabolario;
III. ogni documento viene assegnato ad una distribuzione iniziale di probabilità rispetto i topic/argomenti che verranno estratti e ciascun argomento viene associato ad una distribuzione di probabilità iniziale rispetto alle parole presenti nel vocabolario;
IV. i topic più rilevanti vengono identificati per ciascun documento e le parole chiave per ciascun topic;
V. viene applicato l'algoritmo e generato il modello.

Gli output ottenuti mediante l'applicazione dell'algoritmo sono [6]:

1. gli argomenti in grado di descrivere in modo dettagliato la raccolta di documenti;
2. la distribuzione delle parole chiave per ciascun topic (matrice topical content (φ) $TC_{v,t}$) dove:
   i. v∈{1,…,V} sono le parole del vocabolario dei documenti analizzati;
   ii. V numero totale di parole contenute nel vocabolario della Digital VoC;
   iii. t∈{1,..,T} sono i topic identificati dall'algoritmo di topic modelling;
   iv. T sono il numero totale di topic identificati dall'algoritmo.
3. la distribuzione degli argomenti per ciascun documento (matrice topical prevalence (ϑ) $TP_{j,t}$) dove:
   i. j∈{1,…,J} sono i documenti della Digital VoC;
   ii. J è il numero totale di documenti analizzati della specifica VoC.

La topical prevalence (ϑ) rappresenta la distribuzione multinomiale di probabilità dei topic all'interno dei documenti [6]. La somma della topical prevalence per ciascun topic, considerando un documento J presente nel corpus, è pari ad uno [6]. Nello specifico [6]:

$$\sum_{t=1}^{T} TPj = 1 \quad , t = 1 \; \forall j \tag{1.1}$$

La Figura 1 mostra gli output ottenuti con l'applicazione dell'algoritmo.

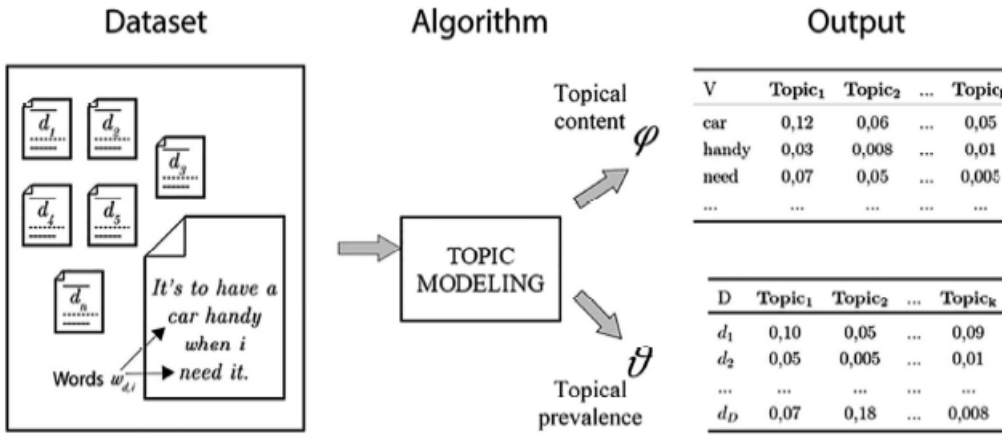

Figura 1 Output di un algoritmo di topic modelling (Adattato [3])

Analizzando la topical content, ovvero la distribuzione delle parole che sono presenti all'interno di uno topic è possibile fare alcune considerazioni su come lavorano questi algoritmi [7]. In particolare, le parole che sono presenti all'interno dei topic risolvono due problemi del text mining [7]:

- sinonimia: l'algoritmo permette di determinare che due parole appartengono allo stesso topic, suggerendo che ci sia una relazione semantica tra di esse;
- polinesia: si riferisce al fenomeno di una singola parola che ha più significati diversi, e per questo può appartenere contemporaneamente a più topic.

I primi approcci di topic model risalgono al 1999, quando, Hoffman formula il "*Probabilistic Latent Semantic Analysis"* (PLSA), il quale, rappresenta il primo modello probabilistico per identificare gli argomenti latenti in un insieme di documenti [8]. Introduce il concetto che ciascun documento del corpus ha una distribuzione multinomiale di argomenti e ciascun argomento è costituito da una distribuzione multinomiale di parole [7].

Agli inizi degli anni 2000, David M. Blei, Andrew Y. Ng e Michael I. Jordan hanno introdotto, il "*Latent Dirichlet Allocation"* (LDA) [9]. L'LDA è un'estensione del PLSA, che considera la distribuzione degli argomenti nei documenti come una distribuzione detta di "Dirichlet" che si adatta nel momento in cui vengono aggiunti nuovi documenti al corpus testuale [7].

L'LDA presenta, però delle restrizioni [10]:

i. gli argomenti identificati dall'algoritmo in un documento sono indipendenti tra di loro;

ii. la distribuzione di parole per ciascun argomento è identica per tutti i documenti che trattano di quello specifico argomento;
iii. considera unicamente il contenuto testuale dei documenti e non altre informazioni associate ai documenti quali, ad esempio, l'autore, la data, la fonte.

Il limite dell'indipendenza degli argomenti viene superato con l'introduzione del "Correlated Topic Model" (CTM) da parte di Blei, D., e Lafferty, J. nel 2005 [11].

Nel 2016 è stato introdotto il "Structural Topic Model" (STM) che supera l'LDA in quanto suppone che la distribuzione degli argomenti nei documenti e delle parole nei documenti dipendano non solo dal contenuto testuali, ma tiene conto dei metadati o covariate [12-13]. I metadati sono variabili accessorie che forniscono informazioni aggiuntive al contenuto testuale dei documenti come ad esempio, la data, il luogo, la fonte e tali informazioni hanno impatto sia sulla topical content che sulla topical prevalence [14]. Nelle analisi si è scelto di utilizzare l'algoritmo STM.

1.3 Algoritmo "Structural Topic Modelling" (STM)

Partendo dal campione di Digital VoC si eseguiranno le fasi presentate nel diagramma di flusso rappresentato in Figura 2 [1].

Figura 2 Diagramma di flusso dello Structural Topic Model (Adattato e tradotto [1])

Nell'Allegato 1 viene approfondito il codice sviluppato nell'ambiente R per l'applicazione di STM [12].

1.3.1 Pre- processamento

Al fine di migliorare l'efficienza dell'algoritmo di topic model, il corpus testuale [1]:

i. è convertito in minuscolo per evitare ambiguità con le medesime parole scritte in maiuscolo;
ii. vengono rimossi i numeri, la punteggiatura e le parole che non aiutano a identificare i topic, quali ad esempio: "who","an","that";

iii. vengono eliminate le parole più brevi di due e più lunghe di quindici caratteri e quelle con frequenza minore di 15 all'interno del testo;
iv. le parole simili vengono ridotte ad un unico termine come, ad esempio, le parole "loved", "lovely", "loves" e "lover" sono ridotte al suffisso "love";
v. le "stopword", ovvero quelle parole che non hanno alcun rapporto con il contenuto del dataset, come "over", "more", "solving", "false" e "too", vengono anch'esse eliminate;
vi. gli n-grammi, che sono le sequenze formate da più parole con un significato specifico vengono trasformate in un unico termine, come ad esempio, "customer-service" viene sostituito da "costumerservice".

1.3.2 Identificazione del numero ottimale di topic

È necessario definire il numero di argomenti da estrarre al fine di identificare correttamente il numero di topic che descrive il corpus testuale [3]. La funzione "search-k", permette di individuarne il numero ottimale [3]. L'output di tale funzione è una metrica detta "Held-out likelihood", utilizzata come misura delle prestazioni del modello. La metrica misura la probabilità che alcuni documenti di testo non visti siano forniti dal modello opportunatamente addestrato [3]. In genere, il 90% dei documenti disponibili fa parte del set di "training" e il restante 10% fa parte del set di "test" [3]. L'intervallo di tale metrica è $(-\infty,0]$, più alto è questo valore e più è statisticamente forte il modello di topic modelling sviluppato, in generale [3]:

- un valore elevato indica che il modello è capace di generalizzare bene da dati non visti, suggerendo che ha appreso rappresentazioni efficaci dai dati di addestramento;
- un valore basso, potrebbe indicare over-fitting, cioè il modello ha "imparato a memoria" i dati di addestramento senza catturarne le strutture sottostanti, oppure under-fitting ovvero il modello è troppo semplice per catturare la complessità dei dati.

Il vantaggio principale di queste valutazioni è che possono essere calcolate automaticamente senza l'aiuto umano [3]. Il numero ottimale di topic coincide, teoricamente, con il valore maggiore di held-out likelihood, dal momento che questo valore presenta un andamento asintotico al progredire del numero dei topic, si sceglie un "*k*" posto all'inizio dell'andamento quasi-stazionario [3]. Ciò consente di impostare i parametri di input dei topic model e misurare automaticamente la qualità dell'output per scegliere i valori migliori [3].

Nella Figura 3, è mostrato un tipico esempio dell'andamento della metrica applicato al caso studio Disneyland nell'ambiente R.

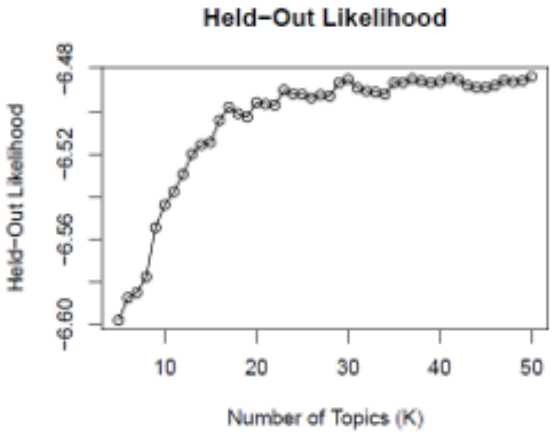

Figura 3 Output della funzione search-k appplicata al caso studio Disneyland

1.3.3 Applicazione algoritmo STM

L'applicazione dell'algoritmo restituisce come output la matrice review-topic, ovvero la distribuzione multinomiale dei topic negli argomenti che rappresenta l'input per la fase di convalida ed analisi dei risultati [14].

1.3.4 Etichettatura

Per ogni argomento si identificano le parole chiave più rilevanti. Tuttavia, per generare un'etichetta pertinente per ciascun topic, il metodo richiede la supervisione umana [14]. L'elenco delle parole chiave più importanti che descrivono un argomento da utilizzare per la fase di etichettatura può essere ottenuto tenendo conto di [14]:

1. High probability: parole con la più alta probabilità all'interno di ogni argomento;
2. Frex: parole che sono frequenti ed esclusive.

Utilizzando entrambi gli approcci si rende più semplice l'etichettatura degli argomenti offrendo una caratterizzazione del contenuto del topic in analisi [14]. Oltre all'elenco delle parole chiave, le recensioni più rilevanti di ogni argomento, ovvero le recensioni con peso più alto, possono essere considerate per l'etichettatura [14]. Ad oggi, non sono ancora state implementate tecniche di etichettatura automatica [14]. Gli Allegati 2 e 3 mostrano gli elenchi di parole chiave ottenute come output dell'algoritmo ed utilizzate nella fase di labeling nei due casi studio in analisi. Nei prossimi capitoli si vedrà l'applicazione dell'STM a due casi studio: Disneyland e Ryanair.

1.3.5 Validazione del modello

Ottenuti i risultati dell'algoritmo STM è necessario validare il modello in modo da verificare che i risultati siano affidabili [3]. Nel tempo sono state proposte diverse metriche ma, manca una procedura standardizzata per la validazione del modello. È possibile applicare due differenti approcci [3]:

- le metriche "automatiche" sono un metodo di convalida applicato quando vi sono tempi brevi di validazione;
- le metriche "supervisionate" richiedono la presenza umana e si applicano quando è richiesta una prova sulla qualità del modello.

Le metriche automatiche non risultano come i migliori candidati in quanto non tengono conto del significato semantico degli argomenti [3]. L'approccio supervisionato richiede molte risorse e tempo proponendo una serie di metriche per valutare le prestazioni degli algoritmi [3]. Tale metodologia riconduce il problema dell'identificazione corretta tra i topic ed i documenti ad un problema di classificazione binaria identificando o meno la presenza della discussione di uno specifico topic [3]. Sulla base di queste valutazioni supervisionate, è possibile calcolare metriche di prestazione come il tasso di accuratezza, tasso di errore, precisione, specificità [3].

1.3.5.1 Modello di validazione supervisionato

La convalida supervisionata è stata applicata come metodo di valutazione nello studio in esame e comprende le seguenti fasi [3]:

(1) estrazione del campione e assegnazione umana degli argomenti,
(2) assegnazione automatica degli argomenti,
(3) confronto dei risultati e
(4) calcolo delle metriche.

La Figura 4 riassume i principali input e output di queste fasi [3].

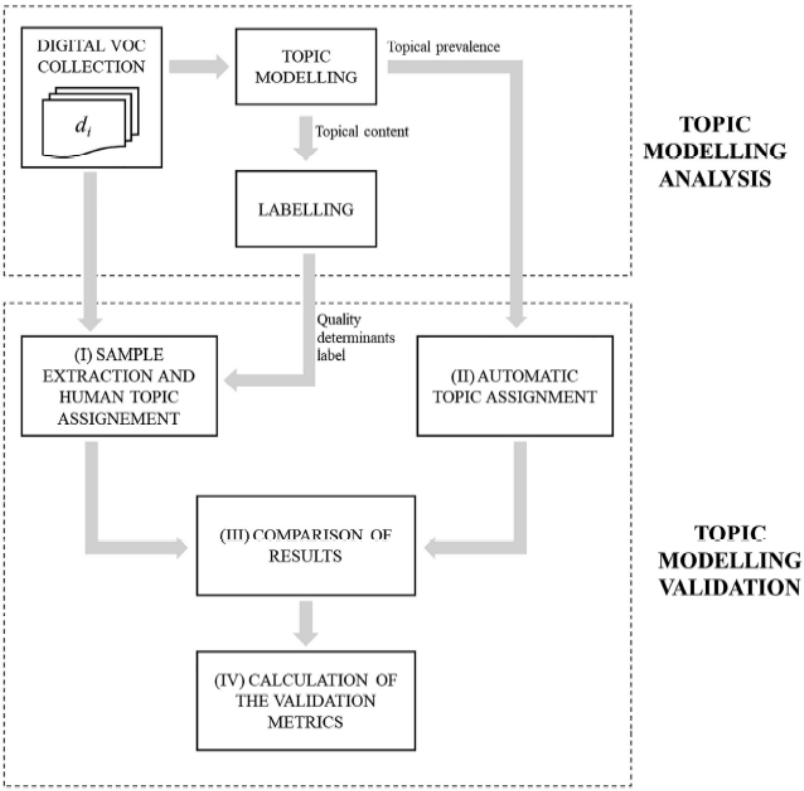

Figura 4 Passi della procedura di validazione del modello di topic model (Adattato [3])

1.3.5.1.1 Assegnazione umana dei topic

Un campione casuale di *n* documenti viene estratto e classificato da un valutatore umano. Il valore di *n* dovrebbe essere scelto in modo da ottenere risultati affidabili [3]. Tale valore non deve essere troppo alto tale da necessitare uno sforzo elevato dei valutatori [3].

Solitamente, viene scelto un valore pari a 100 in quanto considerato sufficiente per ottenere risultati soddisfacenti [3]. Nella pratica, ogni valutatore, deve leggere attentamente il campione estratto e classificare il contenuto di ogni documento in base alle etichette scelte per ciascun topic [3].

Il valutatore deve identificare uno o più argomenti all'interno delle recensioni per rispettare l'ipotesi alla base degli algoritmi di topic modelling, secondo cui ogni documento può contenere un mix di argomenti diversi [3].

1.3.5.1.2 Assegnazione automatica dei topic

L'assegnazione automatica dei topic ai documenti può essere svolta in modo differente in base al criterio di scelta della soglia [3]:

i. Più alta probabilità: la soglia è stabilita mediante la regola del massimo secondo cui un singolo argomento sarebbe rappresentativo di un documento. Tale metodo va contro il principio fondamentale degli algoritmi di topic modelling, secondo cui ogni documento può trattare un mix di argomenti.
ii. Soglia statica: un topic è rappresentativo se si trova al di sopra di un valore definito a priori, ma con tale metodo gli argomenti che sono rilevanti potrebbero non essere noti o considerati marginali.
iii. Soglia dinamica: per ciascun documento viene calcolata la soglia in base alla distribuzione di probabilità dei topic nei singoli documenti.

Di seguito si applicherà il metodo "Tukey fence" per il calcolo della soglia dinamica, con tale metodo i valori che non rientrano nel limite superiore, sono considerati outliers e rappresentano gli argomenti più discussi nel documento [3]. La soglia dinamica viene calcolata come segue [3]:

$$DT_i = Q3_i + (1.5 * IQR_i) \quad (1.2)$$

Per ciascun documento i-esimo presente nel corpus, viene calcolato [3]:

- $Q1_i$ che rappresenta il primo quartile,
- $Q3_i$ che rappresenta il terzo quartile ed
- il range interquartile rappresentato da IQR_i.

Esempio di calcolo:

ID	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	Più alta probabilità	Soglia statica (0,15)	Soglia dinamica
1	0,47	0,01	0,01	0,06	0,02	0,01	0,01	0,04	0,01	0,01	T1	T1	0,0725
2	0,04	0,04	0,06	0,14	0,02	0,02	0,07	0,01	0,04	0,06	T4		0,1125
3	0,03	0,01	0,01	0,08	0,02	0,04	0,1	0,01	0,01	0,01	T7	T7	0,07875
4	0,02	0,01	0,02	0,01	0,01	0,01	0,01	0,02	0,01	0,04	T1	T1	0,035

Tabella 6 Esempio di calcolo della soglia dinamica (Adattato e tradotto [3])

1.3.5.1.3 Confronto dei risultati

Il confronto tra l'assegnazione umana e automatica è mostrato nella Tabella 7 [3].

		Assegnazione umana dei topic (condizione vera)	
		Presenza di Ti	Assenza di Ti
Assegnazione automatica dei topic	Presenza di Ti	**True Positive (*tp*)** Corretta inferenza Accordo tra assegnazione umana ed automatica.Entrambi riconoscono la presenza di un topic in una recensione.	**False Positive (*fp*)** Errore di I specie Non corrispondenza tra assegnazione umana ed automatica. L'algoritmo riconosce la presenza di un topic, mentre il valutatore no.
	Assenza di Ti	**False Negative (*fn*)** Errore di II specie Non corrispondenza tra assegnazione umana ed automatica. Il valutatore riconosce la presenza di un topic, mentre l'algoritmo no.	**True Negative (*tn*)** Corretta inferenza Accordo tra assegnazione umana ed automatica. Entrambi non riconoscono la presenza di un topic in una recensione.

Tabella 7 Matrice dei risultati dell'assegnazione umana combinati con i risultati dell'assegnazione automatica (Adattato e tradotto[3])

Per ciascun documento, è possibile verificare la corrispondenza della valutazione automatica e umana nell'assegnazione della discussione topic-documento come segue [3]:

- vero positivo ("true positive" (*tp*)), cioè accordo tra valutazione umana e automatica nell'assegnazione di un topic ad un documento;
- vero negativo ("true negative" (*tn*)), cioè accordo tra valutazione umana e automatica nel non assegnare un topic ad un documento;
- falso positivo ("false positive" (*fp*)) o errore di prima specie, ovvero disallineamento tra l'assegnazione di un topic al documento da parte dell'STM e la mancata assegnazione da parte dei valutatori;
- falso negativo ("false negative" (*fn*)) o errore di seconda specie, cioè disallineamento tra l'assegnazione di un argomento da parte dei valutatori e la non assegnazione da parte dell'algoritmo.

1.3.5.1.4 Calcolo delle metriche

Sono calcolate diverse metriche per valutare la bontà del modello di argomento nell'identificare correttamente i topic ai documenti [3].

Le metriche di valutazione sono elencate nella Tabella 8.

		Human topic assignment *(true condition)*			
		T_i *existence*	T_i *non-existence*	*Accuracy* $\frac{\sum_{i=1}^{n} tp_i + \sum_{i=1}^{n} tn_i}{\sum_{i=1}^{n} tp_i + \sum_{i=1}^{n} tn_i + \sum_{i=1}^{n} fp_i + \sum_{i=1}^{n} fn_i}$	
Automatic topic assignment	T_i *existence*	**True Positive (*tp*)** Correct inference	**False Positive (*fp*)** Type I error	*Precision* $\frac{\sum_{i=1}^{n} tp_i}{\sum_{i=1}^{n} tp_i + \sum_{i=1}^{n} fp_i}$	*False discovery rate* $\frac{\sum_{i=1}^{n} fp_i}{\sum_{i=1}^{n} tp_i + \sum_{i=1}^{n} fp_i}$
	T_i *non-existence*	**False Negative (*fn*)** Type II error	**True Negative (*tn*)** Correct inference	*False omission rate* $\frac{\sum_{i=1}^{n} fn_i}{\sum_{i=1}^{n} fn_i + \sum_{i=1}^{n} tn_i}$	*Negative predictive value* $\frac{\sum_{i=1}^{n} tn_i}{\sum_{i=1}^{n} fn_i + \sum_{i=1}^{n} tn_i}$
		Recall $\frac{\sum_{i=1}^{n} tp_i}{\sum_{i=1}^{n} tp_i + \sum_{i=1}^{n} fn_i}$	*Fall-out* $\frac{\sum_{i=1}^{n} fp_i}{\sum_{i=1}^{n} fp_i + \sum_{i=1}^{n} tn_i}$	*F1 Score* $2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$	
		Miss rate $\frac{\sum_{i=1}^{n} fn_i}{\sum_{i=1}^{n} tp_i + \sum_{i=1}^{n} fn_i}$	*Specificity* $\frac{\sum_{i=1}^{n} tn_i}{\sum_{i=1}^{n} fp_i + \sum_{i=1}^{n} tn_i}$		

Tabella 8 Metriche per la validazione dei topic model ([3])

L'accuratezza del modello indica la corretta assegnazione di ciascun topic come segue [3]:

$$Accurancy = \frac{\sum_{i=1}^{n} tp_i + \sum_{i=1}^{n} tn_i}{\sum_{i=1}^{n} tp_i + \sum_{i=1}^{n} tn_i + \sum_{i=1}^{n} fp_i + \sum_{i=1}^{n} fn_i} \quad (1.3)$$

Precision anche detta "valore positivo di predizione" è una stima della probabilità che una previsione positiva sia corretta [3]. È il rapporto tra le assegnazioni correttamente previste ed il totale delle assegnazioni dell'algoritmo [3]. Viene calcolato:

$$Precision = \frac{\sum_{i=1}^{n} tp_i}{\sum_{i=1}^{n} tp_i + \sum_{i=1}^{n} fpi} \quad (1.4)$$

Recall rappresenta il rapporto tra gli argomenti assegnati correttamente e il numero totale di argomenti che sono stati identificati dalla valutazione umana [3]. Viene calcolato:

$$Recall = \frac{\sum_{i=1}^{n} tp_i}{\sum_{i=1}^{n} tp_i + \sum_{i=1}^{n} fni} \quad (1.5)$$

F1 score, invece, misura l'accuratezza ed è calcolato come la media armonica tra *Precision* e *Recall* [3]. Questo valore tiene conto sia dei falsi positivi che dei falsi negativi [3]. Viene calcolato:

$$F1\ score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{1.6}$$

Fall-out è la proporzione di falsi positivi, cioè la probabilità che il valutatore non rilevi un argomento quando, invece, l'algoritmo l'ha identificato, rispetto il totale di argomenti non identificati dal valutatore [3]. Viene calcolato:

$$Fall - out = \frac{\sum_{i=1}^{n} fp_i}{\sum_{i=1}^{n} fp_i + \sum_{i=1}^{n} tni} \tag{1.7}$$

Complementarmente, *Miss rate* è la porzione di falsi negativi, ovvero la probabilità che il valutatore rilevi un argomento, quando, invece, l'algoritmo non l'ha identificato, rispetto il totale degli argomenti identificati dal valutatore [3]. Viene calcolato:

$$Miss\ rate = \frac{\sum_{i=1}^{n} fn_i}{\sum_{i=1}^{n} tp_i + \sum_{i=1}^{n} fni} \tag{1.8}$$

Specificity misura la percentuale di veri negativi, cioè la percentuale di argomenti che non sono realmente discussi all'interno dei documenti del campione rispetto il totale degli argomenti non individuati dal valutatore [3]. Viene calcolato:

$$Specificity = \frac{\sum_{i=1}^{n} tn_i}{\sum_{i=1}^{n} fp_i + \sum_{i=1}^{n} tni} \tag{1.9}$$

Negative predictive value è la probabilità che l'algoritmo di topic modelling non rilevi un argomento quando non è effettivamente presente, rispetto, al totale di argomenti non identificati dall'algoritmo [3]. Tale metrica viene calcolata:

$$Negative\ predictive\ value = \frac{\sum_{i=1}^{n} tn_i}{\sum_{i=1}^{n} fn_i + \sum_{i=1}^{n} tni} \tag{1.10}$$

Il complemento ad uno del *Negative predictive value* è il *False omission rate* che rappresenta la percentuale di argomenti non rilevati dall'algoritmo ma, identificati dal valutatore umano, rispetto il totale degli argomenti non identificati dall'algoritmo [3]. Viene calcolato:

$$False\ omission\ rate = \frac{\sum_{i=1}^{n} fn_i}{\sum_{i=1}^{n} fn_i + \sum_{i=1}^{n} tni} \tag{1.11}$$

False discovery rate è la percentuale di argomenti identificati erroneamente dall'algoritmo rispetto a tutti gli argomenti identificati [3]. Viene calcolato:

$$False\ discovery\ rate = \frac{\sum_{i=1}^{n} fp_i}{\sum_{i=1}^{n} tp_i + \sum_{i=1}^{n} fpi} \tag{1.12}$$

Una volta calcolate le metriche è possibile confrontarle con i valori target presenti nella Tabella 9.

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

Tabella 9 Valori di riferimento per gli indicatori di validazione dei topic model (Adattato [3])

1.3.6 Analisi dei risultati

L'STM combina l'analisi degli argomenti con i metadati, andando oltre i modelli convenzionali, come ad esempio l'LDA [2]. Si condurranno analisi sui risultati in base ai metadati associati alle recensioni prese in esame.

1.3.6.1 Calcolo del" Mean Topic Proportion" (MTP)

Partendo dai risultati dell'STM ovvero dalla matrice della topical prevalence è possibile ottenere informazioni rilevanti circa la discussione dei topic [2]. Il valore "*Mean Topic Proportion"*(MTP*)* mostra quanto frequentemente un argomento è discusso in media nelle recensioni del database [2]. Viene calcolato:

$$MTP_t = \sum_{j=1}^{N} \frac{TP_{j,t}}{N} \; \forall \; t \tag{1.13}$$

Dove *N* è il numero di recensioni considerate e *TPi,t* è la topical prevalance del *t-esimo* topic nella *j-esima* recensione [2].

La somma dei *MTPt* relativi a tutti i topic identificati è uguale a 1 [2]:

$$\sum_{t=1}^{T} MTP_t = 1 \tag{1.14}$$

Di seguito un esempio esplicativo del calcolo del MTP [2]:

Digital VoC record	Date	Sampling period (t)	Topical Prevalence ($TP_{j,d}$)		
			Quality determinant A	Quality determinant B	Quality determinant C
1	3 January 2022	1	0.8	0.15	0.05
2	15 January 2022		0.1	0.7	0.2
3	17 January 2022		0.8	0.15	0.05
4	11 February 2022	2	0.25	0.7	0.05
5	16 February 2022		0.45	0.15	0.4
6	18 February 2022		0.35	0.1	0.55
7	9 March 2022	3	0.15	0.65	0.2
8	13 March 2022		0.2	0.1	0.7
9	22 March 2022		0.1	0.3	0.6
		MTP	(0.8 + 0,1 + 0,8 + 0.25 + 0.45 +0.35 + 0.15 + 0.2 + 0.1)/9 = 0.36	(0.15 + 0.7 + 0.15 + 0.7 + 0.15 +0.1 + 0.65 + 0.1 + 0.3)/9 = 0.33	(0.05 + 0.2 + 0.05 + 0.05 + 0.4 +0.55 + 0.2 + 0.7 + 0.6)/9 = 0.31

Tabella 10 Esempio calcolo del valore del MTP (Adattato [2])

1.3.6.2 Calcolo dell'" Interval Mean Topical Prevalence" (IMTP)

Un altro importante aspetto potrebbe essere quello di monitorare la qualità dei prodotti o servizi nel tempo e per eseguire l'analisi proposta, è necessario introdurre una misura di distribuzione degli argomenti su periodi di campionamento specifici [2]. A tal fine, proponiamo il calcolo di "*Interval Mean Topical Prevalence"* (IMTP), che rappresenta quanto viene discusso in media il *d-esimo* topic correlato al *t-esimo* periodo di campionamento [2]. Viene calcolato:

$$IMTP_{d,t} = \frac{\sum_{j}^{R_t} TP_{j,d}}{|R_t|} \tag{1.15}$$

Dove:

- *Rt* è l'insieme di recensioni della Digital VoC nel periodo di campionamento *t-esimo*,
- |*Rt*| è la cardinalità dell'insieme *Rt*.

Per ogni periodo di campionamento *t-esimo*, la somma di *IMTPd,t* relativa a tutti i topic identificati è uguale a 1 [2]:

$$\sum_{d=1}^{D} IMTP_{d,t} = 1 \qquad \forall\, t \in (1, \dots, T) \tag{1.16}$$

Dove:

- *D* è il numero di argomenti identificati,
- *T* è il numero totale di periodi di campionamento.

Facendo riferimento alla Tabella 10, può essere calcolato il valore dell'IMTP su base annuale [2].

Sampling period (t)	$IMTP_{d,t}$		
	Quality determinant A	Quality determinant B	Quality determinant C
1	$IMTP_{A,1} = (0.8 + 0.1 + 0.8)/3 = 0.57$	0.33	0.1
2	$IMTP_{A,2} = (0.25 + 0.45 + 0.35)/3 = 0.35$	0.32	0.33
3	$IMTP_{A,3} = (0.15 + 0.2 + 0.1)/3 = 0.15$	0.35	0.5

Tabella 11 Esempio calcolo del valore dell'IMTP (Adattato [2])

1.3.6.3 Calcolo del "Mean Rating Proportion" (MRP)

Una valutazione globale della soddisfazione, il cosiddetto "rating", viene generalmente associato a ogni recensione [15]. Il rating è spesso espresso su una scala ordinale, che va da una stella, che indica massima insoddisfazione, a cinque stelle che invece, indica massima soddisfazione [15]. Il "*Mean Rating Proportion"* (MRP*)* rappresenta quanto mediamente un argomento viene discusso all'interno delle recensioni a cui è associato un punteggio che solitamente va da una a cinque stelle [15]. Viene calcolato come segue [15]:

$$MRP_{t,k} = \frac{\sum_{i \in R_k} TP_{i,k}}{|R_k|} \tag{1.17}$$

Dove:

- t identifica il topic;
- k è il livello della scala di valutazione;
- R_k è il sottoinsieme di recensioni associate a un livello di valutazione uguale a R_k;
- $TP_{i,k}$ è la topical prevalence del topic t nella recensione i;
- $|R_k|$ è la cardinalità di R_k.

Nota che la somma dei $MRP_{t,k}$ relativi a tutti i topic identificati a un livello di valutazione specifico è uguale a 1 [15]:

$$\sum_{t=1}^{T} MRP_{t,k} = 1 \; \forall \; k \tag{1.18}$$

Di seguito un esempio esplicativo che mostra il calcolo dell'indicatore per il terzo attributo chiave o terza determinate di qualità estratta dal modello [15]:

Review	Rating	Key-attribute 1	Key-attribute 2	Key-attribute 3
Review 1	1	0.7	0.3	**0**
Review 2	1	0.7	0.2	**0.1**
Review 3	2	0.9	0.1	**0**
Review 4	2	0	0.5	**0.5**
Review 5	3	0.4	0	**0.6**
Review 6	3	0.3	0.6	**0.1**
Review 7	4	0.1	0.6	**0.3**
Review 8	4	0.3	0	**0.7**
Review 9	5	0.1	0.3	**0.6**
Review 10	5	0	0.2	**0.8**

$MRP_{3,1} = \frac{0+0.1}{2} = 0.05$

$MRP_{3,2} = \frac{0+0.5}{2} = 0.25$

$MRP_{3,3} = \frac{0.5+0.6}{2} = 0.35$

$MRP_{3,4} = \frac{0.3+0.7}{2} = 0.5$

$MRP_{3,5} = \frac{0.6+0.8}{2} = 0.7$

Tabella 12 Esempio di calcolo del MRP (Adattato [15])

Nell'esempio si vede che l'attributo 3 è maggiormente discusso nelle recensioni che hanno una valutazione pari a 4-5 stelle, questo significa che l'argomento è legato ad aspetti che gli utenti identificano come positivi [15].

Tramite questo indicatore è possibile classificare le determinanti di qualità in [15]:

- positive: che generano soddisfazione del cliente, e per questo devono essere sviluppate e potenziate;
- negative: che causano insoddisfazione del cliente e per questo si devono mettere in atto misure di miglioramento della qualità per ridurne gli effetti;
- neutre: potrebbe sembrare che non abbiano un impatto sulla soddisfazione del cliente, tuttavia, questi elementi possono essere significativi per la percezione della qualità del prodotto o del servizio, in quanto i commenti degli utenti possono causare sia soddisfazione che insoddisfazione.

Per categorizzare i profili MRP, è possibile utilizzare il coefficiente di correlazione *"Spearman-Rho Ranked-Order"* (ρ_s), una misura non parametrica della correlazione tra i ranghi dei livelli di valutazione e i ranghi del MRP [15]. Il ρ_s può essere calcolato come segue [15]:

$$\rho_S = 1 - \frac{6 * \sum_{i=1}^{n} (R(X_i) - R(Y_i))^2}{n * (n^2 - 1)} \tag{1.19}$$

Dove:

- *R(Xi)* rappresentano i ranghi dei livelli di valutazione,
- *R(Yi)* rappresentano i ranghi del *MRPt,k*, ovvero i ranghi della proporzione media di un argomento con rating specifico,
- *n* è il numero di livelli di valutazione considerati.

Ogni determinante può essere classificata come segue [15]:

- se $\rho_s < -0{,}4$ profilo negativo;
- se $-0{,}4 \leq \rho_s \leq +0{,}4$ profilo neutro;
- se $\rho_s > +0{,}4$ profilo positivo.

L'insieme di queste considerazioni fornisce una nuova prospettiva sui risultati degli algoritmi nell'analisi della Digital VoC nel campo della gestione della qualità [15].

1.3.6.4 Key Attributes VoC Map (KA-VoC Map)

La combinazione di MTP ed MRP può porre le basi per la definizione di una nuova tassonomia di determinanti di prodotti, servizi o sistemi prodotto-servizio, che sia dinamica ed in grado di seguire come evolvono le percezioni dei clienti nel tempo [15].

Con l'avvento delle nuove tecnologie, è stato necessario introdurre uno strumento che permettesse di comprendere come analizzare e gestire la soddisfazione del cliente [15].

Lo strumento utilizzato in tal caso si chiama *"Key Attributes VoC Map"* (KA-VoC Map). Dove gli input sono i risultati degli algoritmi di topic modelling e l'output è una mappa strutturata che categorizza gli attributi chiave su due dimensioni: il modo e la misura in cui un attributo chiave viene discusso [15]. Gli attributi sono classificati come segue [15]:

- Obstacles, ovvero attributi molto discussi (MTP elevato) e fonte di insoddisfazione (profilo MRP negativo). Sono le fonti primarie di insoddisfazione, essendo i principali argomenti dei reclami dei clienti.
- Frictions, ovvero attributi poco discussi (MTP basso) e fonte di insoddisfazione (profilo MRP negativo). Rappresentano problemi minori, non sono ampiamente discussi, ma generano principalmente insoddisfazione del cliente.
- Indifferent, ovvero attributi poco discussi (MTP basso) che sono neutrali rispetto alla soddisfazione del cliente (profilo MRP neutro). Essendo scarsamente discussi, sono classificati come non rilevanti perché non hanno un'influenza chiara e definita sulla soddisfazione.
- Sleeping beauties, ovvero attributi neutri rispetto alla soddisfazione del cliente (profilo MRP neutro), ma molto discussi (MTP elevato). Spesso rappresentano dimensioni considerate essenziali e, pertanto, non possono impressionare positivamente o negativamente il cliente. Essendo molto dibattute, possono essere considerate essenziali per la soddisfazione del cliente.

- Promises, ovvero attributi poco discussi (MTP basso) che generano soddisfazione del cliente (profilo MRP positivo). Queste dimensioni rappresentano vantaggi minori o attributi emergenti forniti dall'oggetto analizzato.
- Delights, ovvero attributi molto discussi (MTP elevato) che generano soddisfazione (profilo MRP positivo). I clienti riconoscono un valore a questi attributi, che sono le principali fonti di soddisfazione.

La Tabella 13 raffigura la KA-VoC Map.

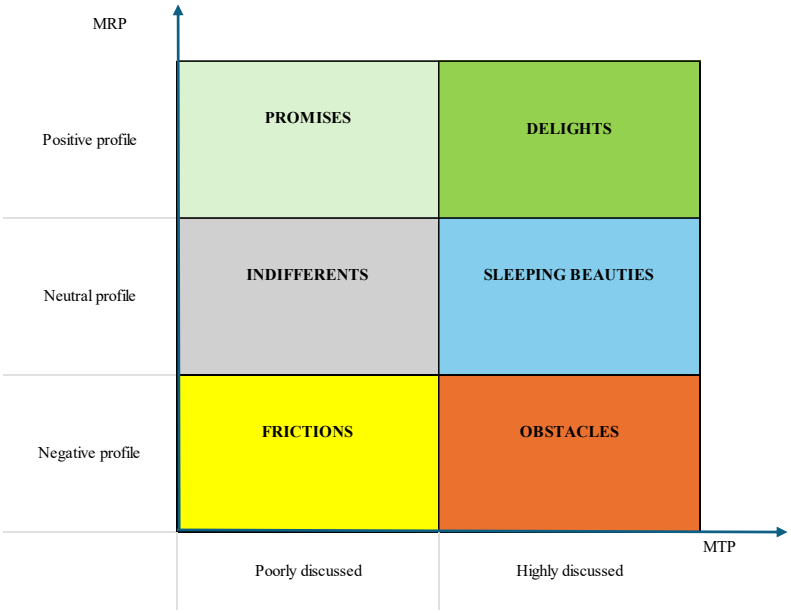

Tabella 13 KA-VoC Map (Adattato e tradotto[15])

La KA-VoC Map distingue tra attributi "scarsamente discussi" e "altamente discussi" in base all'MTP [15]. La soglia è convenzionalmente impostata su $\frac{1}{n}$, dove n è il numero di argomenti identificati. La Tabella 13 mostra le azioni che possono essere intraprese per gestire gli KA identificati [15].

	Azioni del management
Obstacles	Rappresentano barriere al raggiungimento della piena soddisfazione del cliente, e sono necessarie azioni radicali per rimuoverli; per questo i processi e le caratteristiche del prodotto devono essere modificati.
Frictions	Rappresentano fonti di insoddisfazione, ma il loro livello di discussione è inferiore rispetto agli ostacoli. Le frictions sono fonti secondarie di insoddisfazione e l'approccio più appropriato è quello di migliorare gradualmente le prestazioni per soddisfare le aspettative dei clienti.

Indifferent	Non rappresentano né una fonte di insoddisfazione che di soddisfazione nel cliente, per questo l'opzione migliore è ignorarli.
Sleeping beauty	Sebbene non siano una fonte né di soddisfazione né di insoddisfazione del cliente sono molto discussi e per questo bisognerebbe monitorarli, in quanto è possibile che si spostino rapidamente nella categoria obstacles.
Promises	Rappresentano fonti di soddisfazione secondarie per i clienti, ma sono ottimi candidati per trasformarsi in delights, per questo sono attributi che devono essere preservati e migliorati.
Delights	Tali attributi sono le principali fonti di soddisfazione espresse dai clienti attraverso la Digital VoC e, per questo motivo, dovrebbero essere i pilastri della proposta di valore del prodotto o servizio in analisi. La strategia migliore è continuare a investire nel miglioramento di tali attributi.

Tabella 14 Possibili azioni del management sulle azioni da intraprendere in base alla classificazione della KA VoC Map (Adattato e tradotto [15])

È importante sottolineare che l'appartenenza di un attributo a una categoria riflette l'effetto della determinante sulla percezione del cliente [15]. Per questo motivo, prodotti o servizi simili possono avere classificazioni diverse [15]. Nell'analisi aziendale pratica, la KA-VoC Map può essere utilizzata come strumento di benchmarking per analizzare gli attributi chiave di prodotti o servizi simili [15].

2 Capitolo 2: modelli di apprendimento supervisionato

L'applicazione degli algoritmi di topic modelling consente di estrarre gli argomenti latenti discussi all'interno delle recensioni che sono ricondotti alle determinanti di qualità di un prodotto o servizio necessarie per monitorare la qualità e migliorare la soddisfazione del cliente [1].

L'obiettivo del lavoro è stato scegliere un modello che prendesse come input una serie temporale, che nel nostro caso è rappresentata dall'IMTP cioè da quanto viene discusso mediamente un topic in un mese, e condurre analisi predittive sui valori futuri.

Per questo sono state esplorate diverse tecniche in grado di condurre *predictive analytics* in maniera accurata. I modelli di apprendimento supervisionato sono una sottocategoria dell'intelligenza artificiale e dell'apprendimento automatico che aiutano le organizzazioni a trovare una soluzione a diversi problemi reali su larga scala [16]. Nel seguente elaborato si applicano tali modelli per svolgere analisi predittive sui dati [16]. I modelli ricevono in input una serie di dati, opportunamente divisi in un set di "training" per addestrarli al fine di produrre l'output desiderato ed un set di "test" per validare il modello e condurre previsioni accurate [16].

L'apprendimento supervisionato risolve sia problemi di classificazione che di regressione come segue [16]:

- la classificazione utilizza un algoritmo in grado di suddividere il set di dati in categorie specifiche,
- la regressione permette di comprendere la relazione tra variabili dipendenti e indipendenti.

Di seguito sono riportate brevi descrizioni di alcuni dei metodi di apprendimento più comunemente utilizzati, implementati attraverso l'uso di programmi quali, ad esempio R o Python [16].

2.1 Regressione lineare

Per identificare la relazione tra una variabile dipendente ed una indipendente viene solitamente utilizzata la regressione lineare [16]. Si parla di regressione lineare semplice se esiste una sola variabile indipendente e una dipendente, se il numero di variabili aumenta viene definita multipla [16]. La regressione lineare è in grado di catturare pattern e tendenze lineari nei dati,

al crescere del numero di variabili riesce a adattarsi a pattern non lineari, ma non è adatta a catturare andamenti complessi [16]. Nei casi in cui il problema da risolvere sia di classificazione binaria e non un classico problema di regressione, il modello scelto è la regressione logistica. In tale modello la variabile dipendente è categorica esempio “vero” o “falso” [16]. La scelta sul modello di regressione lineare o logistica ricadrà sulla specificità del problema e dalla capacità di condurre previsione accurate sui dati [16].

2.2 Albero decisionale

Gli alberi decisionali sono modelli di classificazione che suddividono i dati in diverse categorie in base a variabili specifiche [16]. In un albero, ogni ramo rappresenta una possibilità e la foglia il risultato della decisione [16]. Questa suddivisione viene ripetuta diverse volte partendo dai rami alle foglie, ovvero dall'alto verso il basso, fino a quando i dati sono stati classificati tutti o in parte in classi specifiche [16]. Nel caso di modellazioni complesse e al fine di ottenere risultati più accurati viene utilizzato l’algoritmo “*Random Forest*” per scopi di classificazione e regressione [16]. La "forest" fa riferimento a un insieme di alberi decisionali non correlati tra loro, che poi vengono uniti per creare previsioni sui dati più precise [17]. Anche in questo caso sarà necessario valutare il problema e in base all’obiettivo, capire se un albero decisionale o una “foresta di alberi” conducono risultati adeguati sui dati in analisi [16-17].

2.3 Rete neurale

Le reti neurali sono modelli predittivi supervisionati che risolvono sia problemi di classificazione (come, ad esempio, il riconoscere immagini) che di regressione (come, ad esempio, la previsione del prezzo di un bene) [16]. Le reti neurali vengono spesso impiegate quando le relazioni tra i dati sono complesse e non lineari [16]. Sono utilizzate in una vasta gamma di applicazioni e sono in grado di cogliere pattern e relazioni nascoste nei dati [16]. Uno svantaggio è che richiedono molti dati per apprendere tali relazioni e quindi hanno bisogno di dataset con molti record. Le reti neurali vengono paragonate ai neuroni del cervello umano [18]. I neuroni ricevono i segnali di input da altri neuroni e sono in grado di trasmetterli in avanti tramite connessioni, allo stesso modo i neuroni, nel contesto delle reti neurali, inviano segnali in avanti in base al loro input, attraverso una funzione detta di “attivazione” [18]. Definiamo il layer o strato come lo spazio in cui viene implementata una serie di funzioni [18]. Ogni layer è composto da un certo numero di neuroni ed al crescere del numero connessioni aumenta anche la complessità della rete [18].

La Figura 5 mostra la *forword pass* ovvero la trasmissione del segnale in avanti ai vari strati della rete [19].

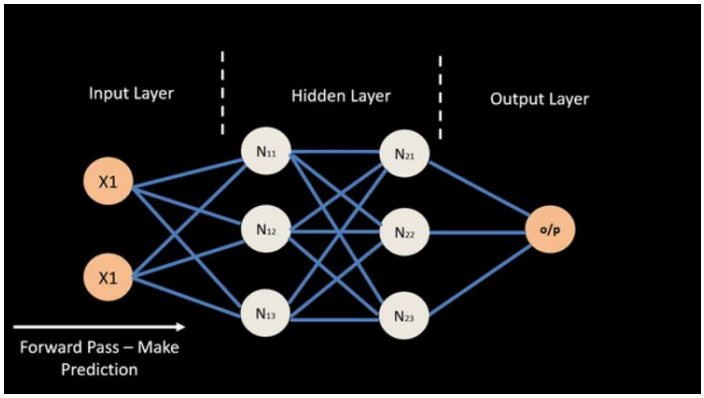

Figura 5 Propagazione del segnale "forward" in una rete neurale (Adattato [19])

Le reti neurali sono addestrate utilizzando un processo chiamato "*back-propagation*", in cui l'algoritmo regola i pesi delle connessioni tra i neuroni per ridurre al minimo l'errore tra l'output previsto e l'output effettivo [18]. In tal modo la rete viene addestrata ed i pesi delle connessioni tra i neuroni vengono variati fintanto che non si ottiene il minimo errore di previsione [18].

La Figura 6 raffigura il passaggio a ritroso e il calcolo dell'errore di previsione.

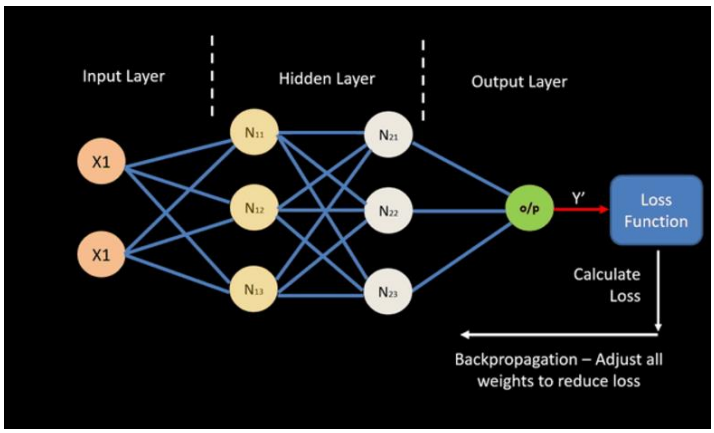

Figura 6 "Backpropagation" e calcolo della perdita in una rete neurale (Adattato [19])

Una rete neurali, deve essere in grado di ricevere gruppi di dati che rappresentano le osservazioni ed apprendere la relazione tra osservazioni e risposte corrette [19].

2.4 Validazione dei modelli di apprendimento supervisionato

Fondamentale non è solo stabilire il modello, ma soprattutto la sua validazione in base ai risultati ottenuti [20-21]. L'applicazione dei modelli di regressione e di classificazione utilizza metodologie differenti per la validazione [20-21].

Nei problemi di classificazione binaria si possono utilizzare delle metriche quali ad esempio l'accuratezza, che rappresenta il rapporto tra le istanze classificate correttamente ed il totale di istanze date in input al modello [20]. Nella pratica è possibile applicare l'approccio introdotto per la validazione dell'algoritmo STM, ma in questo caso, si considerano le istanze che sono state correttamente classificate o incorrettamente rispetto alla classificazione reale dei dati [20].

Nei problemi di regressione, invece, l'obiettivo è di riprodurre i dati della serie e per questo si valuta il discostamento della previsione reale da quella predetta e si utilizzano diverse metriche tra cui: il "*Mean Square Error" (MSE), "Root Mean Square Error" (RMSE)* ed il *"Mean Absolute Error" (MAE)* [21]. Di seguito il calcolo del MSE [21]:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2.1}$$

Dove:

i. y_i rappresenta la variabile reale,
ii. $\hat{y}_i$ la variabile predetta, ed
iii. *n* il numero di osservazioni presenti della variabile.

In sintesi, valuta lo scostamento medio quadratico della serie predetta rispetto ai valori reali [21]. Tale metrica valuta quanto il modello riesce a riprodurre l'andamento della serie ed un valore basso fa sì che il modello sia accurato [21]. L'MSE, però, è molto sensibile agli outlier ed in alcuni casi può essere utilizzata la metrica MAE [21]:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2.2}$$

Il vantaggio di tale metrica è che non considera l'elevamento a potenza di valori lontani dalle previsioni, ma solo il valore assoluto degli scostamenti dei valori reali rispetto ai valori predetti [21]. La radice quadrata del *MSE* viene definita come la deviazione standard dei residui [21]:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (2.3)$$

L'*RMSE* si utilizzerà nelle analisi successive in quanto ci restituisce lo scarto o errore medio delle predizioni rispetto il modello iniziale [21].

In sintesi, la scelta del metodo di validazione dipenderà dall'obiettivo delle analisi da condurre sui dati [20-21].

2.5 Scelta del modello

I modelli di apprendimento supervisionato possono essere applicati a diversi scopi in azienda, come ad esempio, per individuare, isolare e classificare oggetti da video o immagini [16]. Tali metodi sono comunemente utilizzati per condurre analisi predittive con lo scopo di anticipare determinati risultati in base a una variabile di input, aiutando i leader aziendali ad indirizzare ed intraprendere le giuste strategie per l'organizzazione [16].

L'analisi si è concentrata sullo studio predittivo di serie storiche, dove le osservazioni nel tempo sono tra loro dipendenti [22]. Oltre a valutare la previsione siamo interessati alla scomposizione della serie nelle sue componenti che sono [22]:

i. la tendenza, valutata in crescente, decrescente o stazionaria;
ii. la stagionalità, comportamento che si ripete ad intervalli regolari e
iii. il residuo solitamente ha un andamento casuale, stabile nel tempo, ma può essere influenzato da fattori esterni e contenere delle perturbazioni.

Partendo dalla valutazione preliminare della serie, analizzando la tendenza e la presenza di un eventuale stagionalità più o meno forte, si è in grado di scegliere il modello più adatto per l'analisi dei dati [22].

Per analizzare le serie storiche sono state esplorate due tecniche:

i. *"Seasonal Auto-Regressive Integrated Moving Average" (SARIMA)* che tiene conto della componente di autocorrelazione e della stagionalità [22-23];
ii. *"Long Short-Term Memory" (LSTM),* rete neurale ricorrente che consente di analizzare lunghe sequenze ordinate di dati ed è in grado di accumulare informazioni passate per apprendere le relazioni sottostanti [18].

2.5.1 Seasonal Auto-Regressive Integrated Moving Average (SARIMA)

Le serie storiche analizzate mostrano stagionalità e tendenza, quindi, è stato scelto per l'analisi il modello "*Seasonal Auto-Regressive Integrated Moving Average" (SARIMA)* [23-32]. Tale modello differisce dai comuni modelli di regressione in quanto questi ultimi non considerano che vi siano dipendenze tra dati passati della serie [22-23]. Per tal motivo, sono stati utilizzati modelli auto regressivi che tenessero conto dell'autocorrelazione e della stagionalità tramite l'incorporazione dei valori precedenti nella serie [22-23]. È necessario verificare la stazionarietà della serie prima di applicare modelli regressivi, perché, risultati spuri, possono essere generati da dati non stazionari [33]. La parola "stazionarietà" si riferisce al fatto che autocorrelazione, media e varianza sono costanti in una serie temporale stazionaria [33]. I parametri del modello si distinguono in componenti stagionali e non stagionali come mostrato [23]:

$$SARIMA = (p, d, q)\ \ (P, D, Q)_m \tag{2.4}$$

dove "*m*" è il periodo della stagionalità, ovvero il numero di osservazioni per anno.

Nello studio in esame tale valore è pari a 12.

Gli altri parametri del modello sono [23]:

- *p* "*ordine della componente autoregressiva*" *(AR(p))* che cattura il grado di autocorrelazione della variabile di interesse rispetto i valori passati della serie;
- *d* "*numero di differenziazioni*"; ovvero la differenza tra osservazioni successive della variabile di interesse necessaria per rendere la serie stazionaria;
- *q* "*ordine della componente a media mobile*" *(MA(q))* componente rappresentativa della media mobile che incorpora gli errori delle componenti passate nella serie, in quanto quest'ultimi vengono utilizzati come predittori nel modello che è simile a quello di regressione.

Allo stesso tempo i termini della componente stagionale, indicati con la lettera maiuscola, identificano i medesimi parametri [23]. Nelle analisi si vedrà che l'applicazione di tale metodo ha diversi limiti che non permettono di catturare pattern molto complessi ed in tal caso si valuterà la rete neurale per superarli.

2.5.2 Analisi predittiva con SARIMA su Python

Per l'applicazione del modello SARIMA, è stato necessario istallare le seguenti librerie:

- Pandas ci consente di analizzare i big data e di trarre conclusioni basate su teorie statistiche ed è in grado di ripulire set di dati disordinati e renderli leggibili e pertinenti [24];
- il pacchetto principale di Python per l'informatica scientifica è NumPy, è una libreria che ha una varietà di operazioni rapide su array [25];
- Matplotlib che consente di ottenere grafici partendo dai dati della serie [26];
- Statsmodel.tsa.seasonal in particolare la funzione "seasonal_decompose" che prevede la decomposizione della serie temporale [27];
- Statmodels.tsa.stattools in particolare la funzione "adfuller" necessaria per implementare il test di verifica di stazionarietà sui dati della serie prima di applicare il modello [28];
- Pmdarima in particolare "auto_arima" specifico per trovare i migliori parametri del modello [29];
- Statsmodels.tsa.arima.model in particolare "ARIMA" necessario per la fase di addestramento del modello [30];
- Sklearn.metrics la funzione "mean_square_error" [31].

La Figura 7 mostra le diverse fasi del modello.

Figura 7 Diagramma di flusso del modello SARIMA

2.5.2.1 Decomposizione della serie

La serie è stata decomposta nelle sue componenti: tendenza, stagionalità e residuo [23]. Per ciascun topic è stato possibile valutare il trend/tendenza e la presenza di una stagionalità più o meno forte nei dati [23]. Inoltre, è stato valutato il residuo ottenuto sottraendo dalla serie tendenza e stagionalità [23].

2.5.2.2 Test di stazionarità

I dati vengono pre-elaborarati di modo che sia verificata la condizione di stazionarietà per applicare il metodo SARIMA [33].

Pertanto, viene applicato ai dati il test "*Aumentated Dickey-Fuller*"(ADF). Tale test ha l'obiettivo di valutare la presenza di radici unitarie nei dati di una serie temporale [33]. La radice unitaria è una caratteristica dei dati che prevede che la media e la varianza non siano stabili nel tempo, ma che aumentano e questo potrebbe portare a risultati e conclusioni incorrette [33]. Il test ADF è un test di ipotesi dove [33]:

i. l'ipotesi nulla presuppone che la serie non sia stazionaria verificando la presenza di una radice unitaria;
ii. l'ipotesi alternativa suggerisce la stazionarietà.

Nel caso in cui sia verificata l'ipotesi nulla sulla serie vengono applicate alcune metodologie per rendere la serie stazionaria [33]. Nel caso studio in analisi è stata applicata la differenziazione che prevede di calcolare la differenza tra osservazioni successive nella serie in modo da stabilizzarla [23-33].

2.5.2.3 Suddivisione del dataset 80-20%

La fase di addestramento del modello viene preceduta dalla suddivisione opportuna del dataset [32-34]. Il dataset viene suddiviso in una percentuale variabile in "training" e "test" di modo che il modello sia addestrato su dati noti e siano valutate le prestazioni sui dati non visti. In questo modo, possiamo valutare quanto bene i modelli generalizzano su nuove istanze [32-34].

Nelle analisi, poiché l'ordine temporale dei dati è fondamentale la suddivisione del dataset è stata fatta scegliendo uno specifico istante temporale [34].

Il rapporto tra il set di training e di test deve essere scelto in modo accurato, in quanto [34]:

i. un set di training ridotto può portare ad *"under-fitting"* sottostimare il modello e ad una valutazione inefficace,
ii. un set di training che comprende tutti o la maggior parte dei dati della serie può condurre ad "*over-fitting*", imparare a memoria i dati del set e non riuscire a generalizzare efficacemente su dati non visti.

Solitamente il rapporto scelto è di allocare il 70-80% dei dati per il training e il restante 20-30% per il test [34]. Per valutare la robustezza dei modelli oltre alla semplice suddivisione del

dataset in 80-20%, spesso, viene utilizzato un altro metodo detto "*cross-validation*" che consiste nel suddividere non una singola volta, ma ripetutamente il dataset ed i valori che si ottengono sono rappresentati da una media dei risultati [34].

Nelle analisi sui due casi studio si implementerà la suddivisione 80-20% per il caso studio Disneyland e 60-40% per il caso studio Ryanair. La scelta della suddivisione ridotta nel secondo caso si valuterà nella fase applicativa (si veda Sezione 4.2).

2.5.2.4 Addestramento

I parametri del modello vengono selezionati in modo automatico tramite l'applicazione della funzione "*auto-arima*" che è in grado di impostare i parametri che meglio si adattano alla serie in esame [30]. Scelti i parametri il modello viene addestrato.

2.5.2.5 Previsione e calcolo RMSE

Si eseguono le previsioni e viene calcolato il "*Root Mean Square Error"* (RMSE) [31]. Infine, vengono rappresentati graficamente i valori effettivi e predetti [27].

2.5.3 Recurrent Neural Network (RNN)

Le reti neurali introdotte in precedenza (si veda Sezione 2.3) trattano i dati forniti in input come una sequenza di osservazioni indipendenti tra di loro; tuttavia, poiché i dati delle serie storiche sono intrinsecamente ordinati, la gestione richiederà l'introduzione di modifiche rispetto a quest'ultime [18].

La Figura 8 mostra un esempio di come opera un particolare tipo di rete neurale, detta "*Recurrent Neural Network"* (RNN) [35].

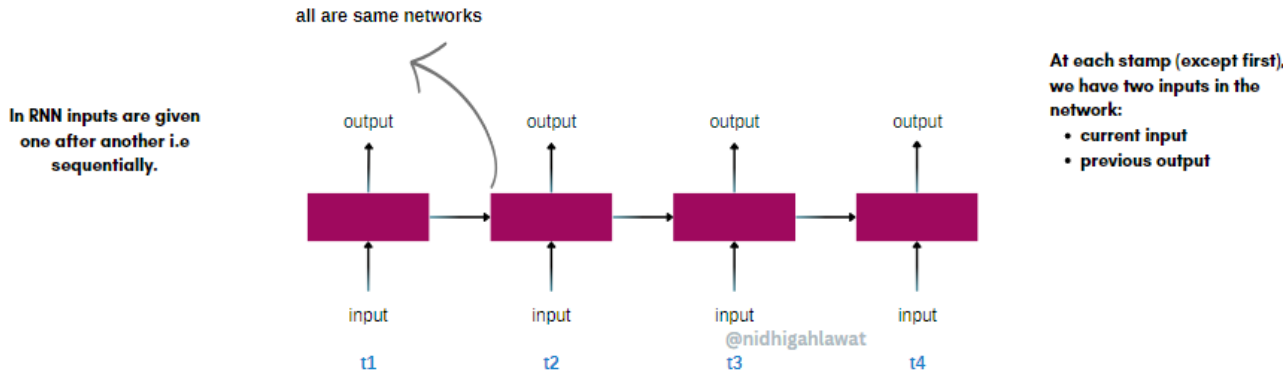

Figura 8 Schema di funzionamento di una rete neurale ricorrente (Adattato[[35])

Nella Figura 8, a ogni time step viene dato un input e viene prodotto un output, e tale output viene fornito per il time step successivo [18-35]. L'output di un time step viene prodotto utilizzando l'output del passo precedente con l'input del time step corrente [18-35]. I dati vengono fatti passare nella rete un elemento della serie alla volta di modo che la serie riesca a catturare le relazioni tra quest'ultimi [18]. Le RNN sono in grado di mantenere informazioni nel tempo su ciò che hanno visto nei time step precedenti mentre, vengono fatti passare i nuovi elementi della sequenza [18].

Scopo del lavoro è di costruire un modello in grado di comprendere le relazioni tra le sequenze fornite nella fase di addestramento e la fase temporale successiva [18]. Un ottimo candidato per la costruzione del modello nel caso di serie storiche è il "*Long Short-Term Memory"* (LSTM) che è un tipo di RNN progettata per gestire sequenze di dati e catturare le dipendenze nel tempo [18]. Rispetto alle semplici RNN, quest'ultime, hanno una cella di memoria e "gate" ovvero delle porte in grado di selezionare il flusso di informazioni e di aggiornarle in maniera efficace [18].

2.5.4 Analisi predittiva con LSTM su Python

L'applicazione del codice LSTM ha richiesto oltre alle librerie precedentemente installe, quali, Pandas, NumPy e Matplotlib, le librerie:

- Sklearn.preprocessing in particolare la funzione "MinMaxScalar" per la normalizzazione dei valori di input [36];
- Keras.model in particolare la funzione "Sequential" che permette di costruire una rete neurale con dei layer sequenziali [37];
- Keras.layers che definisce gli strati della rete in particolare, LSTM e Dense [38];
- Sklearn.model_selection in particolare la funzione "train_test_split" che prevede la suddivisione del dataset in training e test [39];
- Sklearn.metrics in partcolare le funzioni per il calcolo del MSE RMSE MAE [31].

La Figura 9 raffigura il diagramma di flusso del modello.

Figura 9 Diagramma di flusso del modello LSTM

2.5.4.1 Creazione delle sequenze

I dati prima di essere forniti al modello devono essere pre-elaborati per ottenere previsioni più accurate, per questo sono stati standardizzati in modo da avere media e varianza nulla [36-40].

Un algoritmo di apprendimento supervisionato richiede che i dati siano forniti come una raccolta di campioni, in cui ognuno ha una componente di input e una componente di output; quindi, la serie deve essere trasformata in campioni di input e output [37-40].

I modelli LSTM prendono in input sequenze di dati, e per questo è stato necessario definire il valore della "window", ovvero della finestra temporale data in input all'algoritmo [40-41]. Sequenze più lunghe catturano maggiormente le caratteristiche della serie ma risultano più complesse dal punto di vista computazionale [40-41].

Nel caso in esame la lunghezza della finestra temporale è stata variata manualmente e poi stabilita sul valore 10. Il modello prende in input 10 valori e predice il valore successivo e così via.

2.5.4.2 Suddivisione del dataset 80-20%

Il set di training viene utilizzato per addestrare il modello, ossia in questo caso per aggiornare iterativamente i pesi delle connessioni degli strati della rete [34]. Il set test, invece, viene utilizzato per stimare le prestazioni del modello [34].

Nelle analisi sui due casi studio si implementerà la suddivisione 80-20% per il caso studio Disneyland e 60-40% per il caso studio Ryanair. La scelta della suddivisione ridotta nel secondo caso si valuterà nella fase applicativa (si veda Sezione 4.2).

2.5.4.3 Creazione del modello

Le sequenze di input dell'LSTM, una volta create, vengono rimodellate come un vettore di tre componenti [40]:

i. il batch ovvero il numero di campioni,
ii. l'intervallo temporale che rappresenta il punto di osservazione del campione e
iii. il numero di caratteristiche della serie.

Nel caso in esame il numero di caratteristiche della serie è pari ad uno in quanto la previsione viene svolta su un singolo valore nel tempo, cioè sul valore dell'IMTP di un singolo topic.

Il modello è stato generato specificando il numero di strati della rete che dipende dalla complessità delle relazioni da apprendere nei dati e dalle previsioni da effettuare [40-42].

Nel modello in esame, è stato scelto uno strato LSTM di modo da addestrare il modello ed apprendere le relazioni tra i dati ed uno strato Denso detto "totalmente connesso" in grado di modellare gli output del modello che nel nostro caso è pari ad uno [41-42].

Durante la fase di ottimizzazione, obiettivo, è stato ridurre al minimo la perdita rappresentata dal "*Mean Square Error*" (MSE), ovvero dall'errore quadratico medio [31].

2.5.4.4 Addestramento

Nella fase di addestramento vengono stabiliti il numero di parametri, in particolare, il numero di epoche [41]. L'epoca corrisponde al numero di passaggi *forword* e *backword* nella rete con conseguente calcolo della perdita e quindi, della valutazione del modello (si veda Sezione 2.3) [18-19]. Facendo variare il numero di epoche, viene valutata la perdita andando ad impostare il numero ottimale che corrisponde al mimino del valore di MSE [18-19].

2.5.4.5 Previsioni e calcolo RMSE

Il modello è valutato mediante il calcolo del RMSE [31]. Se si divide questa metrica per il valore medio dell'obiettivo si ottiene una misura di quanto una previsione, è in media, lontana dal suo valore reale [18]. I modelli di previsione dovrebbero essere continuamente monitorati e aggiornati man mano che nuovi dati diventano disponibili, quindi è necessario valutare nel tempo e apportare modifiche per garantirne l'efficacia [18].

2.6 Applicazioni di predictive analytics nel customer service

Le tecniche di topic modelling sono ampiamente utilizzate per catturate informazioni rilevanti da set di dati e comprendere la soddisfazione o insoddisfazione dei clienti nei riguardi di un prodotto o servizio [1]. Tuttavia, le tecniche predittive applicate nel lavoro in esame risultano inesplorate o poco esplorate nel monitoraggio della qualità nel customer service. Attualmente le RNN ed LSTM, implementate nei casi studio (si veda Capitolo 3 e 4), sono utilizzate per indagare il sentiment all'interno di documenti testuali auto rilasciati dai clienti [43]. La valutazione del sentiment dei clienti permette la classificazione dei documenti mediante l'applicazione del modello, mentre lo studio in esame utilizzerà LSTM per risolvere problemi di regressione su serie storiche [43]. Il metodo SARIMA risulta anch'esso inesplorato per analisi nel customer service. Alcuni studi mostrano un'analisi parziale delle serie storiche, in

quanto partendo dai risultati degli algoritmi di topic modelling considerano solo la valutazione della tendenza della serie nel tempo e non conducono analisi più profonde sui dati [3].

I metodi scelti vengono comunemente utilizzati per svolgere analisi predittive su serie storiche in diversi campi, come ad esempio uno studio che tratta della previsione della velocità del vento nel campo delle energie rinnovabili [44]. Tali metodi, SARIMA ed LSTM vengono messi a confronto e valutati i risultati in termini di accuratezza tramite il calcolo delle metriche di valutazione viste precedentemente (si veda Sezione 2.4) [44].

Il lavoro svolto può essere considerato pioniere nell'analisi del customer service, sebbene tali metodologie vengano utilizzate da numerose aziende per prevedere dati finanziari, eventi metereologici e molto altro, mancano applicazioni che prendano come input i risultati ottenuti dai modelli di topic modelling, in particolare STM, per svolgere tali previsioni sui dati [43-44].

Nel capitolo seguente si vedrà come le tecniche utilizzate hanno un forte impatto nella gestione della qualità aziendale.

3 Capitolo 3: analisi delle recensioni di Disneyland

Il primo campione di recensioni analizzato è relativo al parco divertimenti Disneyland e presenta informazioni che riguardano la valutazione complessiva dell'esperienza a partire da 03- 2010 fino a 05- 2019.

I metadati associati al campione sono i seguenti:

- Rating: valutazione complessiva dell'esperienza che va da 1 a 5;
- Review_ID: codice identificativo dell'utente;
- Year_Month: data della recensione cliente;
- Reviewer_Location: paese di provenienza del cliente;
- Review_Text: testo della recensione;
- Branch: parco divertimenti Disneyland visitato nello specifico HongKong, Parigi o California.

Gli obiettivi dell'analisi sono:

- valutare le determinanti di qualità applicando STM alla Digital VoC e
- condurre analisi predittive sull'Interval Mean Topical Prevalence (IMTP) per ciascun topic mediante i metodi SARIMA ed LSTM precedentemente introdotti (si veda Sezione 2.5).

3.1 Applicazione STM

Il corpus testuale è stato ripulito da tutte quelle recensioni che non presentano una data, in quanto, l'analisi si concentra sulla valutazione dell'Interval Mean Topical Prevalence (IMTP) e questi valori potevano distorcere la stima (si veda Sezione 1.3.5).

Implementando lo script (si veda Allegato 1) nell'ambiente R, mediante la funzione di pre-processamento il testo delle recensioni è stato ulteriormente semplificato rimuovendo ad esempio, le parole che presentano una bassa frequenza e le "stopword" ovvero quelle non necessarie per l'individuazione dei topic.

Il testo delle recensioni presentava un numero abbastanza elevato di caratteri; quindi, nonostante l'applicazione del pre-processamento il corpus non si è ridotto in modo significativo (da 42 mila a circa 40 mila recensioni).

Il metodo scelto per l'individuazione del numero di topic ottimale è la metrica Held-out likelihood (si veda Sezione 1.3.2).

Dalla Figura 10, si è scelto un valore di *k* pari a 13 in quanto si può considerare che l'andamento della metrica sia quasi stazionario o comunque non subisca variazioni significativamente rilevanti a partire da tale valore.

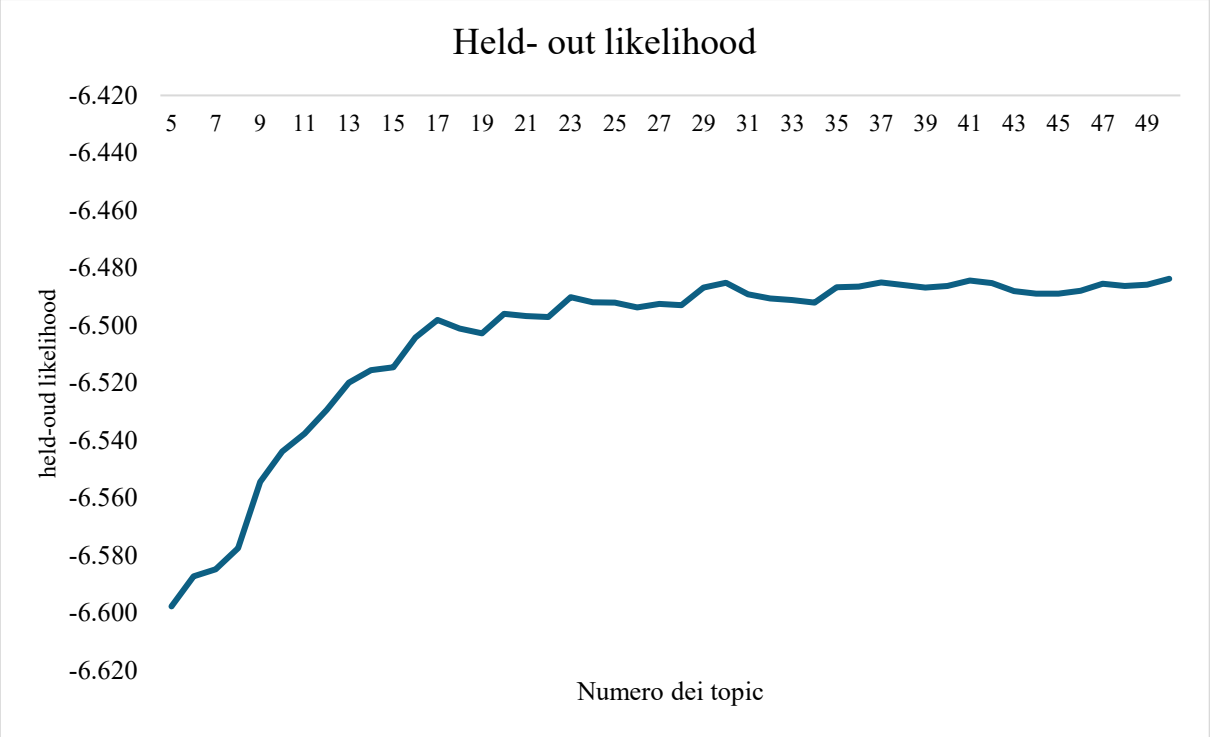

Figura 10 Held-out likelihood caso studio Disneyland

Il risultato di STM per Disneyland ha permesso di individuare la topical content e la topical prevalence, in particolare partendo dalla topical content, si è data un'etichetta a ciascun topic tenendo conto delle parole che avevano la massima probabilità (vedi Allegato 2 e Tabella 15).

La Tabella 15 presenta una descrizione dettagliata dei topic e delle parole chiave identificate dall'algoritmo.

Etichette topic	Parole chiave	Descrizione
1) Incontri con i personaggi	character, photo, princess, image, mouse	Il topic tratta delle interazioni con i personaggi in particolare, con quelli più iconici.
2) Confronto tra parchi	park, different, compare, area, walk	Tratta del confronto dell'esperienza vissuta in diversi parchi, nello specifico delle dimensioni e sulla disposizione.
3) Folla e visite stagionali	crowd, season, vacation, travel, summer	Tale topic è collegato ai periodi in cui c'è più folla all'interno del parco e come questa impatta sull'esperienza dei clienti.
4) Logistica di prenotazione	book, check, train, arrive, entrance	Il topic tratta della pianificazione della visita partendo dalla prenotazione all'arrivo al parco.

5) Esperienza di soggiorno e hotel	hotel, stay, excellent, room, breakfast	Tratta dell'esperienza dei visitatori che soggiornano all'interno del parco e la qualità dei servizi che vengono offerti.
6) Delusione per i prezzi e le giostre	close, disappoint, money, price, carousel	Tale topic riguarda l'esperienza dei visitatori circa i prezzi all'interno del parco e dei biglietti acquistati ed ha un'accezione negativa.
7) Personale amichevole e disponibile	magic, staff, friend, member, help	Tratta della disponibilità e cordialità del personale all'interno del parco nel risolvere problematiche.
8) FastPass e pianificazione del viaggio	pass, fast, trip, plan, FastPass	Tratta dell'esperienza di utilizzo di sistema salta coda come, ad esempio, il Fast Pass che permette di accedere più rapidamente alle attrazioni.
9) Esperienza per famiglie	child, place, carousel, family, children	Il topic tratta dell'esperienza complessiva delle famiglie e, in particolar modo, della valutazione della soddisfazione dei bambini.
10) Giostre popolari a tema	mountain, space, carousel, star, pirate	Tratta dell'esperienza su giostre note come, ad esempio, Space Mountain o i Pirati dei Caraibi.
11) Costi del cibo e ristorazione	food, expensive, restaurant, drink, eat	Tale topic tratta del confronto tra l'offerta di cibo e bevande all'interno del parco dal punto di vista dei prezzi e della qualità.
12) Spettacoli, sfilate e fuochi d'artificio	show, parade, fireworks, night, guard	Tale topic tratta dell'intrattenimento notturno come, ad esempio spettacoli, fuochi d'artificio o sfilate.
13) Halloween ed eventi speciali	Halloween, carousel, theme, party, event	Riguarda gli eventi stagionali, in particolare le feste a tema come ad esempio, Halloween, le decorazioni e le esperienze correlate.

Tabella 15 Labeling caso studio Disneyland

3.1.1 Validazione del modello

Considerando 100 recensioni estratte in maniera casuale il confronto tra l'assegnazione umana ed automatica ha condotto i risultati presenti in Tabella 16 (si veda Sezione 1.3.5):

		Assegnazione umana dei topic (true condition)	
		Presenza Ti	Assenza Ti
Assegnazione automatica dei topic	Presenza Ti	True positive (tp) 187	False positive (fp) 60
	Assenza Ti	False negative (fn) 26	True negative (tn) 1027

Tabella 16 Confronto tra assegnazione umana ed automatica caso studio Disneyland

Per valutare la bontà del modello sono state calcolate le metriche come mostrato in Tabella 17.

Indicatori	Valori	Valori target
Precision	0,76	>0,70
False omission	0,03	<0,05
False discovery rate	0,24	<0,05
Negative predictive value	0,98	>0,90
Accuracy	0,93	>0,95
Recall	0,88	>0,70
Fall-out	0,06	<0,05
Miss rate	0,12	<0,20
Specificity	0,94	>0,90
F score	0,81	>0,70

Tabella 17 Calcolo delle metriche di validazione caso studio Disneyland

Tutti gli indicatori, eccetto due, rientrano nei valori di soglia stabiliti, simbolo di un'accurata analisi che ha portato i risultati sperati. Tale valore è dovuto ad un numero abbastanza elevato di *false positive* posto al denominatore della metrica *false discovery rate* che ha condotto a tale risultato (si veda Sezione 1.3.5.1.4). Questo vuol dire che il modello sovrastima il valore dei veri positivi.

Dall'analisi risulta che le etichette scelte sono rappresentative dei topic, i parametri di input sono corretti ed è presente pertinenza e omogeneità nel database di riferimento.

3.1.2 Analisi dei risultati

La matrice topical prevalence ha permesso di sviluppare analisi quantitative e qualitative per il monitoraggio della qualità e per comprendere la soddisfazione o insoddisfazione del cliente [1].

Sui risultati STM sono state svolte le seguenti analisi:

(i) calcolo del Mean Topic Proportion (MTP) (si veda Sezione 1.3.6.1);
(ii) calcolo del Mean Rating Proportion (MRP) (si veda Sezione 1.3.6.3);
(iii) Key Attributes VoC Map (KA-VoC Map) (si veda Sezione 1.3.6.4);
(iv) calcolo dell'Interval Mean Topical Prevalence (IMTP) (si veda Sezione 1.3.6.2).

3.1.2.1 Calcolo del Mean Topic Proportion (MTP)

È stato calcolato il valore dell'MTP, per individuare quanto viene mediamente discusso ciascun topic.

La Figura 11 misura il risultato ottenuto.

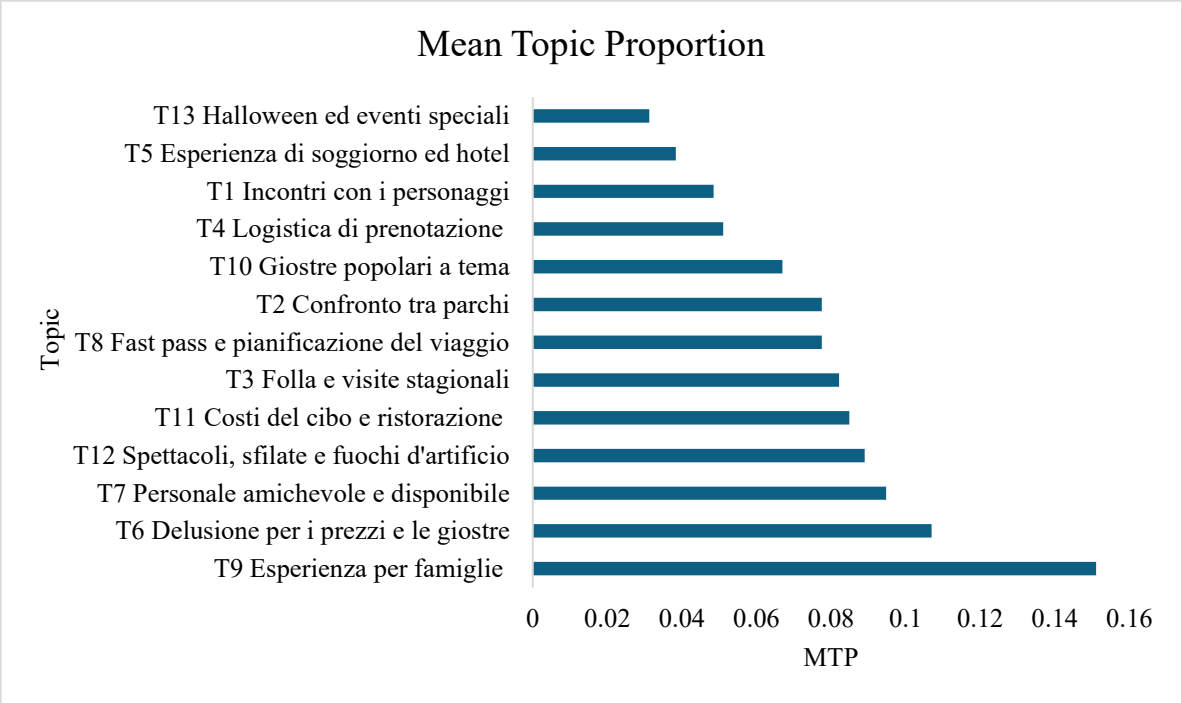

Figura 11 Analisi MTP caso studio Disneyland

Osservando la Figura 11 risulta che i topic maggiormente discussi, nell'ordine sono: *(T9)"Esperienza per famiglie", (T6)"Delusione per i prezzi e le giostre",(T7) "Personale amichevole e disponibile*".

3.1.2.2 Calcolo del Mean Rating Proportion (MRP)

Il rating associato ai documenti testuali ha permesso di calcolare il valore del MRP.

La Figura 12 mostra il profilo di MRP per ciascun topic individuato.

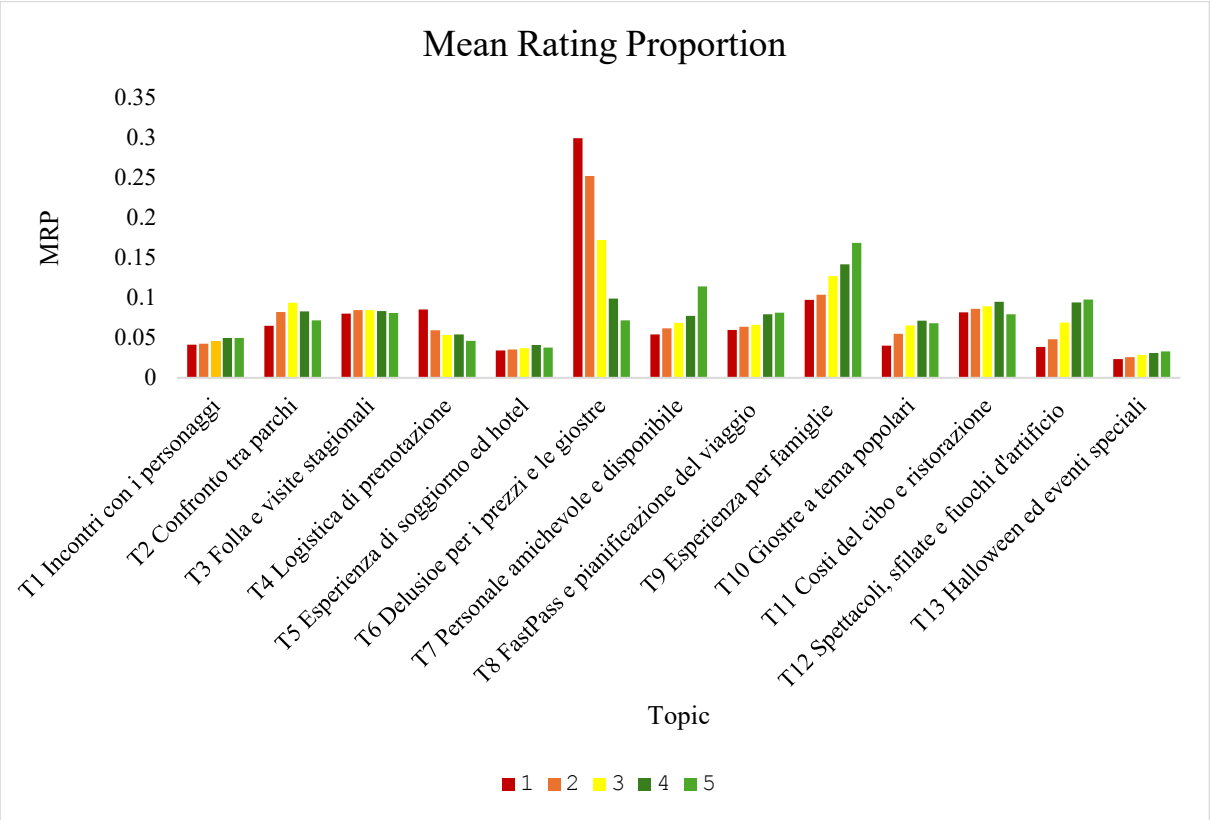

Figura 12 Analisi MRP caso studio Disneyland

I profili dei topic sono classificati come segue:

- profilo positivo: (*T7*) "*Personale amichevole e disponibile", (T8) "FastPass e pianificazione del viaggio", (T9) "Esperienza per famiglie", (T10) "Giostre popolari a tema", (T12) "Spettacoli, sfilate e fuochi d'artificio*";
- profilo neutro: *(T1) "Incontri con i personaggi", (T2) "Confronto tra parchi", (T3) "Folla e visite stagionali", (T5) "Esperienza di soggiorno e hotel",(T11) "Costo del cibo e ristorazione", (13) Halloween ed eventi speciali*;
- profilo negativo: *(T4) "Logistica di prenotazione", (T6)" Delusione per i prezzi e le giostre".*

3.1.2.3 Key Attributes VoC Map (KA-VoC Map)

Tramite i risultati ottenuti dal MTP e MRP è possibile costruire la KA-VoC Map.

Si considerano come:

1. poco discussi i topic con un $MTP < \frac{1}{n}$ (dove n è il numero di topic) e
2. molto discussi i topic con un $MTP \geq \frac{1}{n}$.

Nel dataset in esame il rapporto è pari a $\frac{1}{K} = 0{,}077$ e la successiva classificazione dei topic:

- molto discussi: *(T9) "Esperienza per famiglie", (T12) "Spettacoli, sfilate e fuochi d'artificio", (T6)" Delusione per i prezzi e le giostre", (T3) "Folla e visite stagionali*";
- poco discussi: *(T7) "Personale amichevole e disponibile", (T8) "FastPass e pianificazione del viaggio", (T10) "Giostre popolari a tema", (T4) "Logistica di prenotazione", (T1) "Incontri con i personaggi", (T2) "Confronto tra parchi", (T5) "Esperienza di soggiorno e hotel".*

La Tabella 18 mostra la KA-VoC Map.

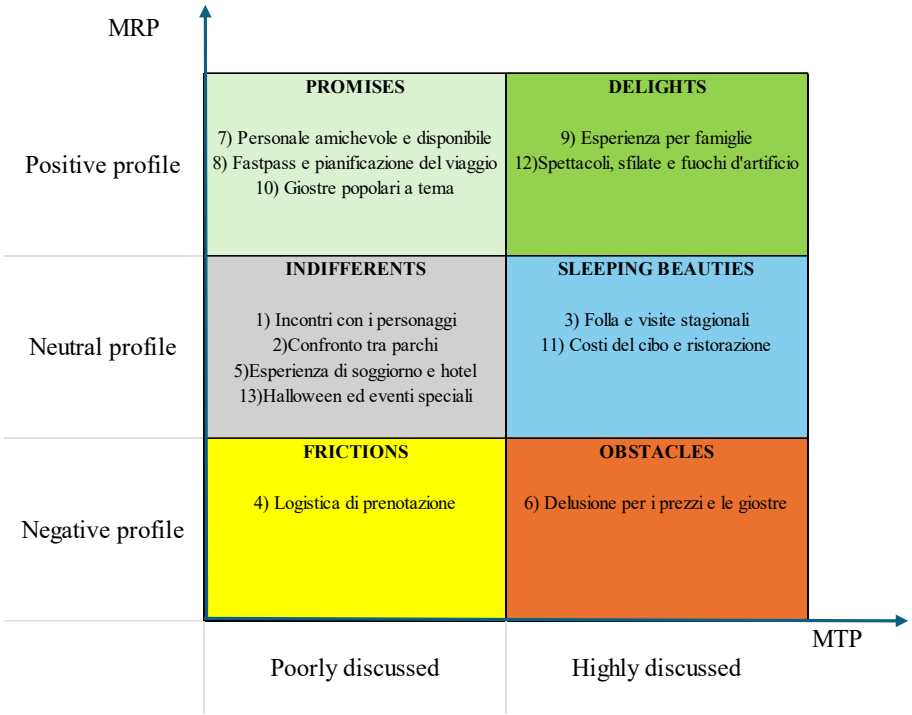

Tabella 18 KA-VoC Map caso studio Disneyland

Le categorie della KA-VoC Map sono:

- Obstacles*: 6) Delusione per i prezzi e le giostre*. Tale topic presenta un valore di MTP elevato e sono fonte di insoddisfazione per i clienti.
- Frictions: *4) Logistica di prenotazione*. Scarsamente discusso e fonte di insoddisfazione. Rappresenta problemi meno gravi, ma che possono comunque generare insoddisfazione nei clienti.

- Indifferents:*1) Incontri con i personaggi; 2) Confronto tra parchi ;5) Esperienza di soggiorno ed hotel;13) Halloween ed eventi speciali*. Sono topic scarsamente discussi che sono neutrali rispetto alla soddisfazione del cliente. Sono considerati non rilevanti poiché non hanno un'influenza chiara sulla soddisfazione o insoddisfazione.

- Sleeping Beauties: *3) Folla e visite stagionali; 11) Costi del cibo e ristorazione*. Sarebbero attributi molto discussi e neutri rispetto alla soddisfazione del cliente. Rappresentano dimensioni essenziali o caratteristiche di base che non hanno né un impatto positivo né negativo sulla soddisfazione dei clienti.

- Promises:*7) Personale amichevole e disponibile; 8) Fast pass e pianificazione del viaggio; 10) Giostre popolari a tema*. Sarebbero topic scarsamente discussi che generano soddisfazione nel cliente. Queste dimensioni rappresentano vantaggi minori o attributi emergenti del prodotto o servizio analizzato.

- Delights:*9) Esperienza per famiglie; 12) Spettacoli, sfilate e fuochi d'artificio*. Sono topic molto discussi e che generano un alto grado di soddisfazione. Questi topic sorprendono positivamente i clienti, superando le loro aspettative e aumentando significativamente la loro soddisfazione.

3.1.2.4 Calcolo dell'Interval Mean Topical Prevalence (IMTP)

È stato calcolato l'IMTP e sono stati classificati opportunamente i profili ottenuti.

La Figura 13 mostra l'IMTP relativo al topic con andamento crescente.

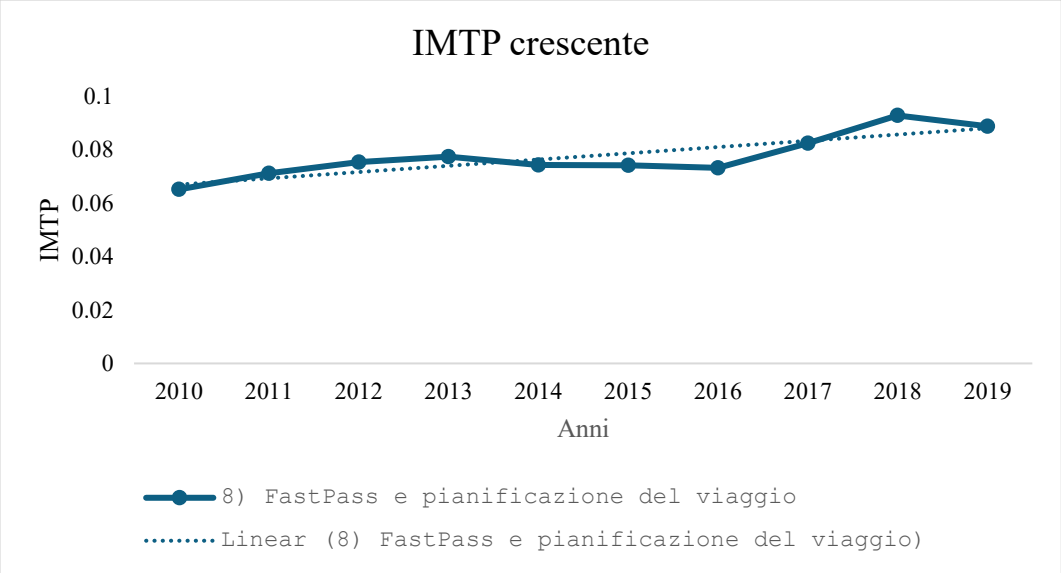

Figura 13 Analisi IMTP crescente caso studio Disneyland

La Figura 14 mostra l'IMTP del topic con andamento decrescente.

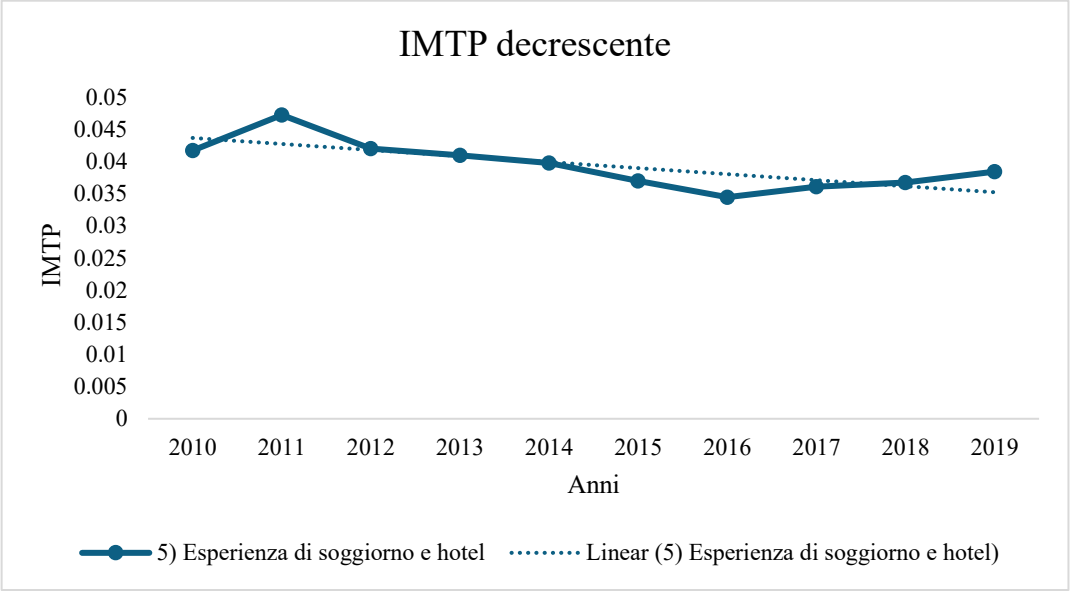

Figura 14 Analisi IMTP decrescente caso studio Disneyland

La Figura 15 mostra l'IMTP per i topic che presentano un andamento stazionario.

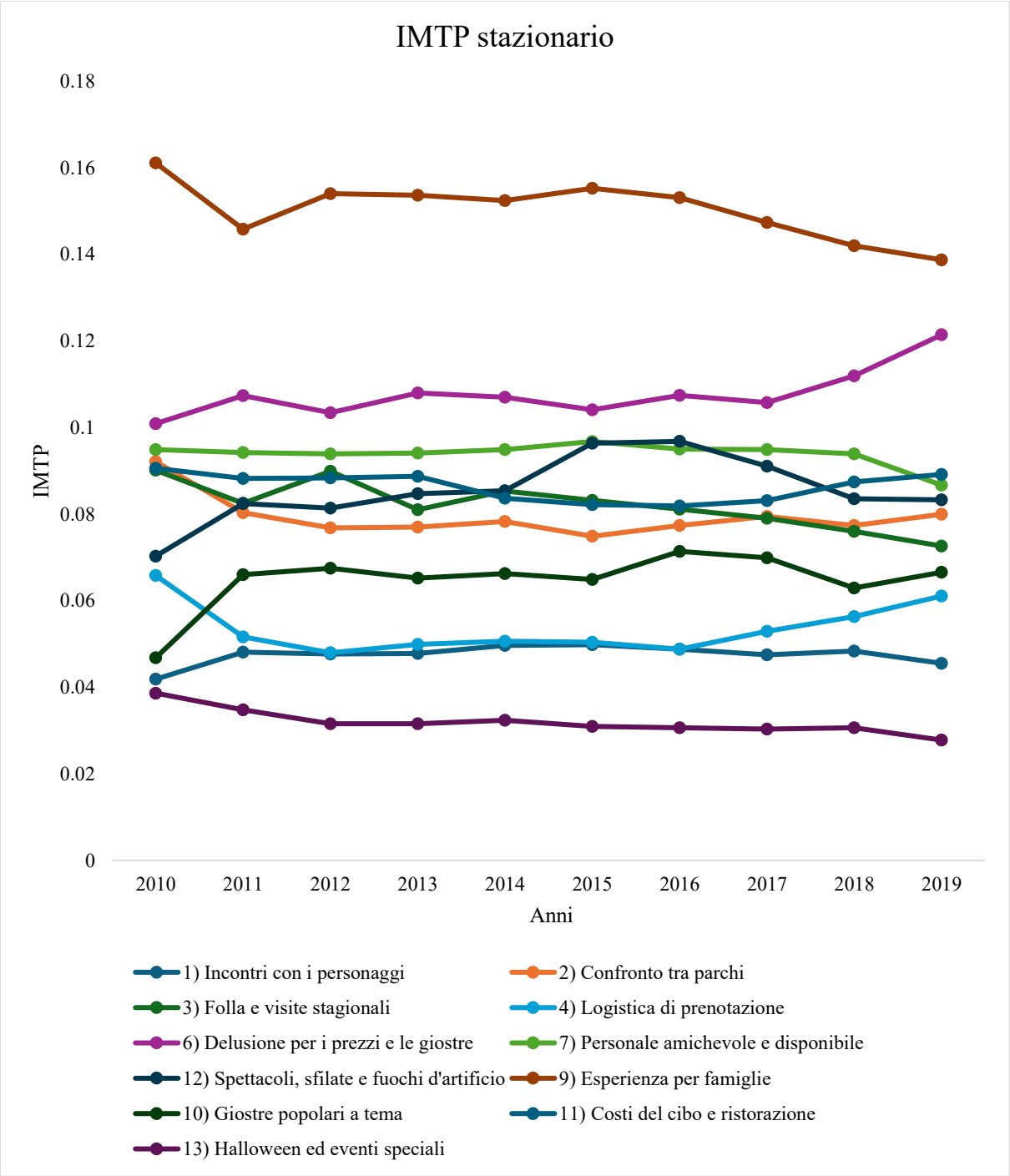

Figura 15 Analisi IMTP stazionario caso studio Disneyland

Partendo dal calcolo dell'IMTP si è analizzato l'andamento nel dettaglio di ciascun topic.

3.2 Analisi predittive sull'Interval Mean Topical Prevalence (IMTP)

Ottenuti i profili dell'andamento dell'IMTP e classificati opportunatamente, è possibile condurre analisi predittive su serie storiche.

L'implementazione di tale procedura richiede una valutazione delle predizioni mediante i metodi descritti precedentemente (si veda Sezione 2.5.2 e 2.5.4).

La Tabella 19 mostra una piccola parte del dataset campione del modello per le analisi predittive.

date	IMTP1	IMTP2	IMTP3	IMTP4	IMTP5	IMTP6	IMTP7	IMTP8	IMTP9	IMTP10	IMTP11	IMTP12	IMTP13	SOMMA
2011-1	0,07	0,07	0,07	0,10	0,07	0,11	0,08	0,07	0,13	0,04	0,09	0,07	0,02	1,00
2011-2	0,05	0,07	0,07	0,05	0,07	0,13	0,12	0,06	0,12	0,07	0,09	0,07	0,03	1,00
2011-3	0,06	0,07	0,07	0,06	0,07	0,11	0,08	0,05	0,15	0,07	0,10	0,09	0,03	1,00
2011-4	0,06	0,11	0,08	0,05	0,05	0,11	0,08	0,07	0,14	0,06	0,11	0,06	0,03	1,00
2011-5	0,05	0,08	0,07	0,07	0,06	0,11	0,08	0,06	0,15	0,06	0,09	0,08	0,03	1,00
2011-6	0,04	0,08	0,09	0,04	0,05	0,10	0,08	0,08	0,17	0,06	0,09	0,09	0,03	1,00
2011-7	0,04	0,08	0,08	0,05	0,05	0,13	0,08	0,08	0,14	0,07	0,09	0,08	0,03	1,00
2011-8	0,05	0,08	0,08	0,05	0,05	0,12	0,08	0,08	0,14	0,08	0,09	0,08	0,03	1,00
2011-9	0,05	0,08	0,08	0,06	0,05	0,10	0,09	0,07	0,16	0,07	0,09	0,07	0,03	1,00
2011-10	0,04	0,08	0,08	0,05	0,04	0,11	0,09	0,07	0,14	0,06	0,09	0,08	0,06	1,00
2011-11	0,05	0,08	0,08	0,05	0,04	0,09	0,11	0,06	0,16	0,06	0,09	0,10	0,03	1,00
2011-12	0,05	0,08	0,09	0,04	0,04	0,10	0,12	0,07	0,14	0,07	0,08	0,09	0,03	1,00
2012-1	0,05	0,07	0,10	0,04	0,03	0,09	0,10	0,07	0,16	0,07	0,09	0,09	0,03	1,00
2012-2	0,05	0,08	0,10	0,05	0,05	0,10	0,10	0,07	0,15	0,06	0,09	0,07	0,03	1,00
2012-3	0,05	0,08	0,09	0,05	0,05	0,10	0,09	0,07	0,15	0,07	0,09	0,07	0,03	1,00

Tabella 19 Esempio di input per le analisi predittive caso studio Disneyland

Per quanto riguarda il caso studio Disneyland, si è scelto di svolgere analisi predittive in base alla metrica IMTP per tre topic in esame:

- *8) "FastPass e pianificazione del viaggio"* che presenta un andamento dell'IMTP crescente;
- *5)" Esperienza di soggiorno e hotel"* che presenta un andamento dell'IMTP decrescente;
- *13)" Halloween ed eventi speciali* "che presenta un andamento dell'IMTP stazionario.

3.2.1 Topic 8 "Fast pass e pianificazione del viaggio"

La Figura 16 rappresenta il valore dell'IMTP per il *topic 8 "FastPass e pianificazione del viaggio".*

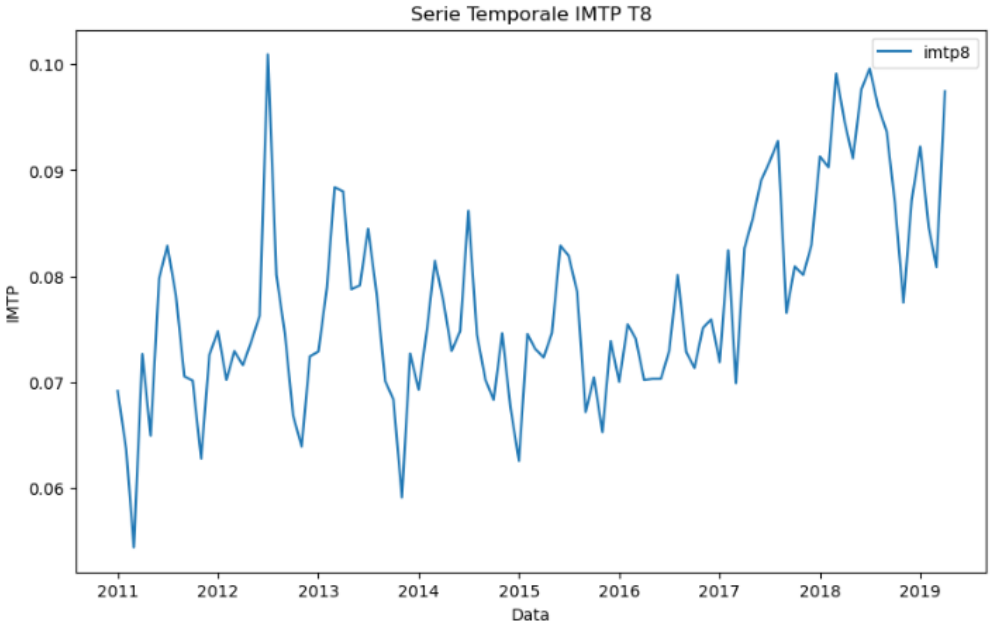

Figura 16 Rappresentazione grafica dell'IMTP del topic 8"Fast pass e pianificazione del viaggio"

La Figura 17 raffigura la decomposizione della serie temporale in: trend/tendenza, stagionalità e residuo.

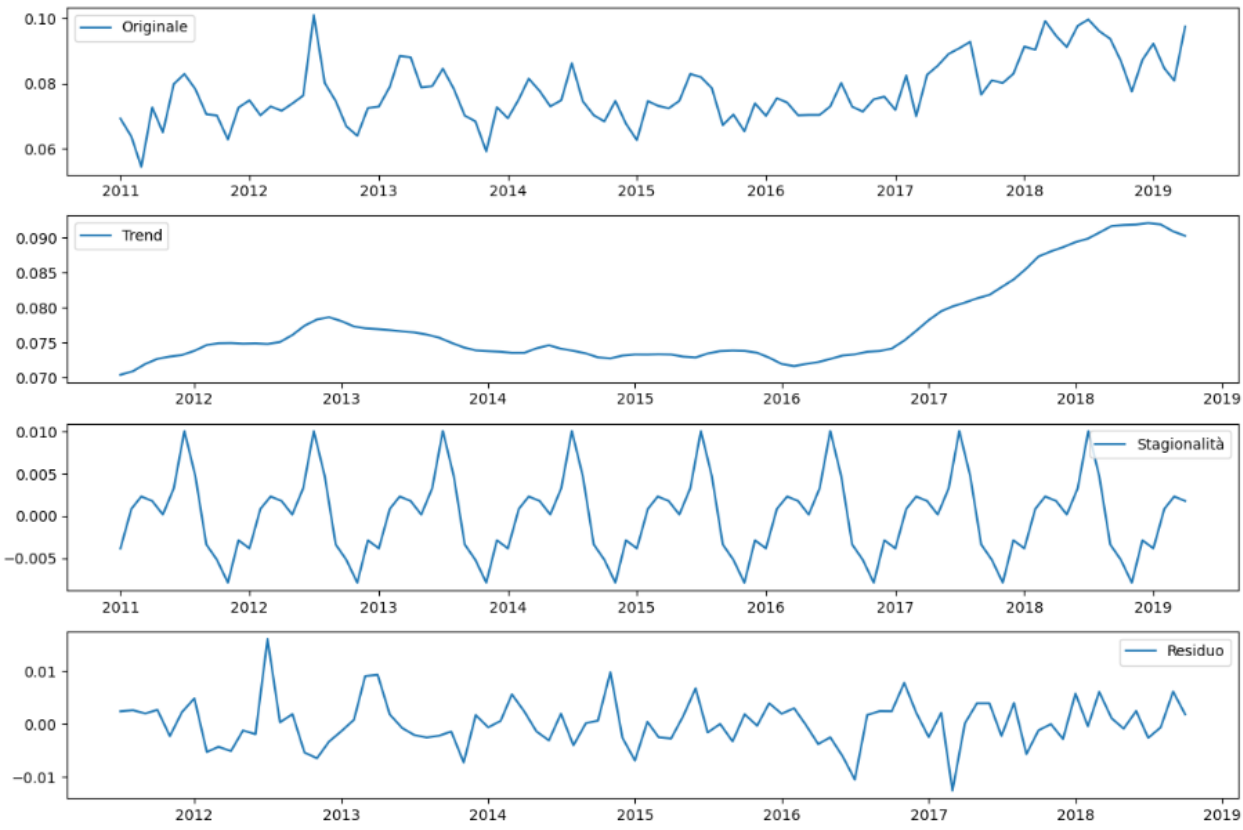

Figura 17 Decomposizione dell'IMTP del topic 8 "Fast pass e pianificazione del viaggio" in: 1. rappresentazione della serie originale, 2. rappresentazione del trend/tendenza della serie, 3. rappresentazione della stagionalità della serie, 4. rappresentazione del residuo della serie.

La Figura 17 mostra 4 diversi grafici:

1. la serie originale ha una certa variabilità con oscillazioni regolari;

2. la tendenza che presenta una crescita graduale;

3. le oscillazioni stagionali sono regolari e chiaramente definite;

4. i residui mostrano una lieve variabilità e ciò vuol dire che la tendenza e la stagionalità sono sufficienti per spiegare la maggior parte della serie.

3.2.1.1 Applicazione SARIMA

La Figura 18 mostra la previsione ottenuta mediante l'applicazione del metodo SARIMA (si veda Sezione 2.5.1 e 2.5.2 ed Allegato 4).

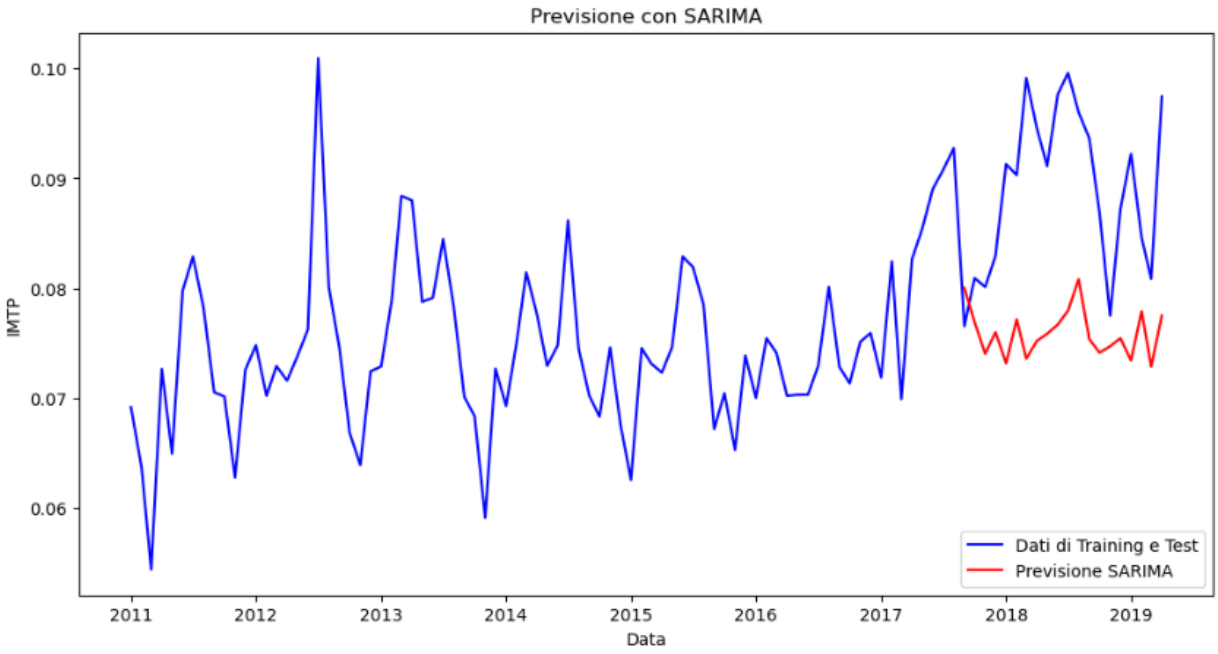

Figura 18 Applicazione di SARIMA all'IMTP del topic 8 "Fast pass e pianificazione del viaggio"

La linea blu rappresenta i dati reali utilizzati per addestrare e validare il modello, mentre la linea rossa rappresenta le previsioni.

Le previsioni seguono la tendenza generale e riproducono in parte la stagionalità della serie catturando in modo moderato le oscillazioni presenti nei dati storici, anche se queste ultime sono leggermente smorzate rispetto ai dati reali. Le previsioni ottenute risultano più basse rispetto la serie originale dei dati.

Il valore del *Root Mean Square Error (*RMSE*)* (si veda Sezione 2.4) in questo caso risulta essere pari a 0,015, ed indicherebbe un modello che riesce ad ottenere previsioni abbastanza accurate.

3.2.1.2 Applicazione LSTM

La Figura 19 mostra la previsione ottenuta mediante l'applicazione del metodo LSTM (si veda Sezione 2.5.3 e 2.5.4 ed Allegato 5).

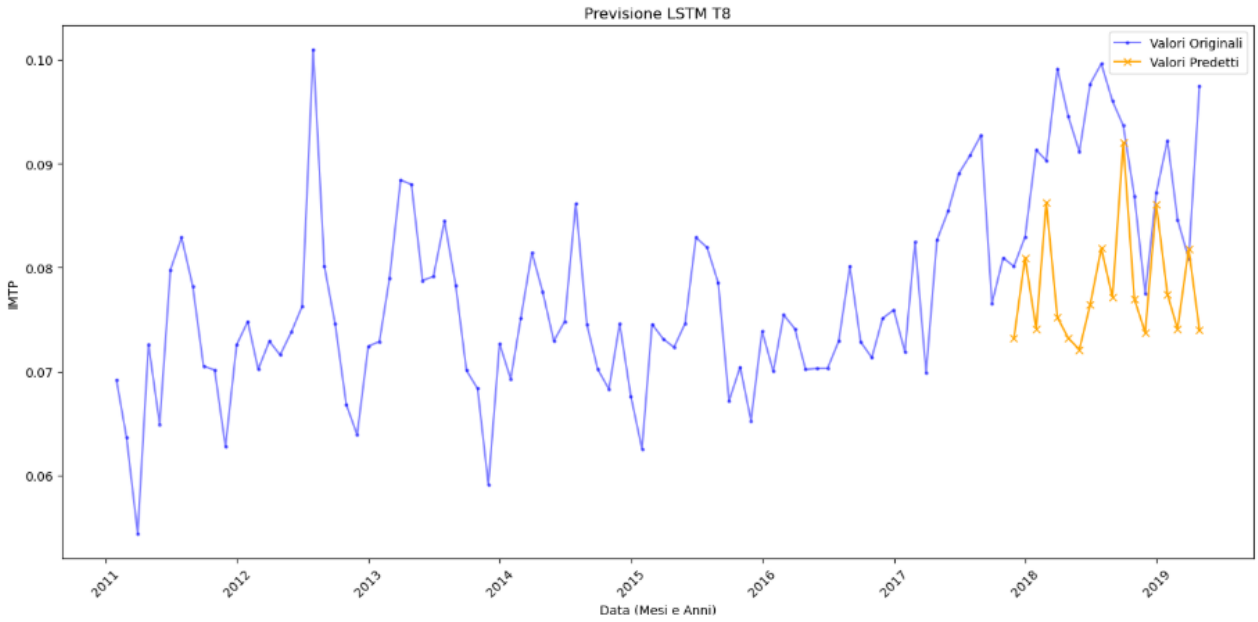

Figura 19 Applicazione di LSTM all'IMTP del topic 8 "Fast pass e pianificazione del viaggio"

La linea blu rappresenta i dati della serie, mentre la linea arancione rappresenta le previsioni generate dal modello LSTM.

Le previsioni seguono bene l'andamento generale della serie originale, considerando sia la tendenza della serie che le oscillazioni. Le oscillazioni previste sono evidenti, ma i picchi e le valli sono più basse rispetto ai valori predetti. Di seguito sono state calcolate le metriche per valutare l'accuratezza del modello (si veda Sezione 2.4):

Mean Absolute Error (MAE)= 0,0059

Mean Square Error (MSE)= 4,97

Root Mean Square Error (RMSE) = 0,007

Notiamo che l'RMSE è più piccolo rispetto all'applicazione di SARIMA, ed in questo caso, è stato calcolato anche il MAE, il cui valore è prossimo all'RMSE suggerendo che il modello non presenta elevati *outlier* che possono influenzare i risultati (si veda Sezione 2.4).

3.2.2 Topic 13 "Halloween ed eventi speciali"

La Figura 20 mostra la rappresentazione grafica della metrica IMTP per il *topic 13 "Halloween ed eventi speciali"*.

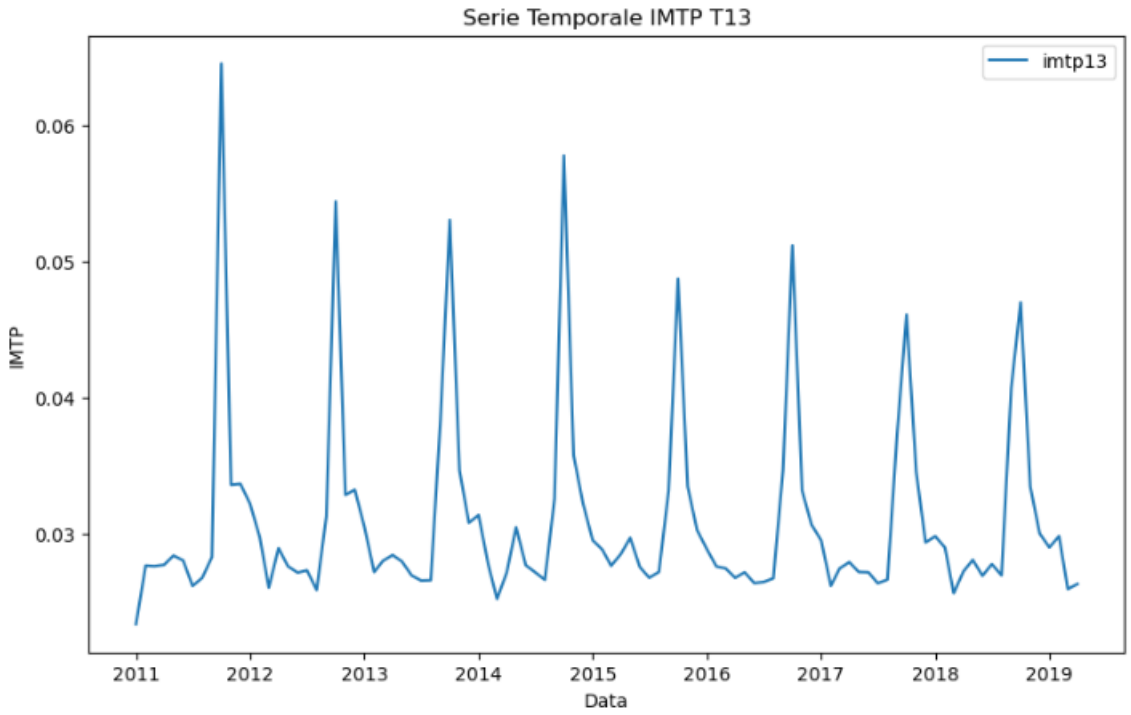

Figura 20 Rappresentazione grafica dell'IMTP del topic13 "Halloween ed eventi speciali"

La Figura 21 raffigura la decomposizione della serie in: trend/tendenza, stagionalità e residuo.

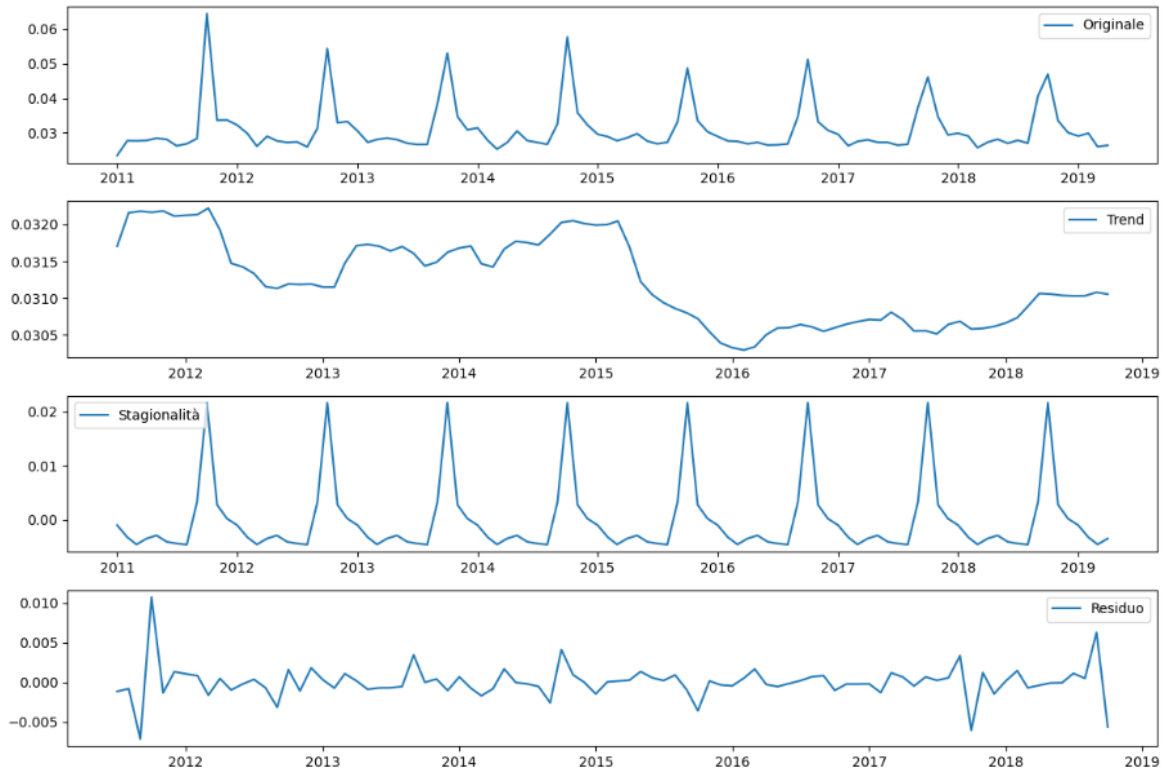

Figura 21 Decomposizione dell'IMTP del topic13 "Halloween ed eventi speciali" in: 1. rappresentazione della serie originale, 2. rappresentazione del trend/tendenza della serie, 3. rappresentazione della stagionalità della serie, 4. rappresentazione del residuo della serie.

La Figura 21 mostra 4 diversi grafici:

1.la serie originale ha un andamento con picchi regolari;

2. la tendenza mostra una fluttuazione iniziale seguita da una diminuzione graduale negli ultimi anni;

3.la stagionalità evidenzia picchi regolari e marcati, l'ampiezza delle oscillazioni è significativa rispetto ai valori della serie originale e questo suggerisce stagionalità forte;

4. i residui sono piuttosto piccoli, suggerendo che il modello cattura bene sia la tendenza sia la stagionalità.

Tali considerazioni sono ragionevoli in quanto ci si aspetta che il topic 13, poiché tratta di festività ricorrenti durante l'anno, abbia una discussione maggiore in alcuni mesi rispetto ad altri (esempio periodo natalizio, oppure Halloween).

3.2.2.1 Applicazione SARIMA

La Figura 22 mostra la previsione ottenuta mediante l'applicazione del metodo SARIMA (si veda Sezione 2.5.1 e 2.5.2 ed Allegato 4).

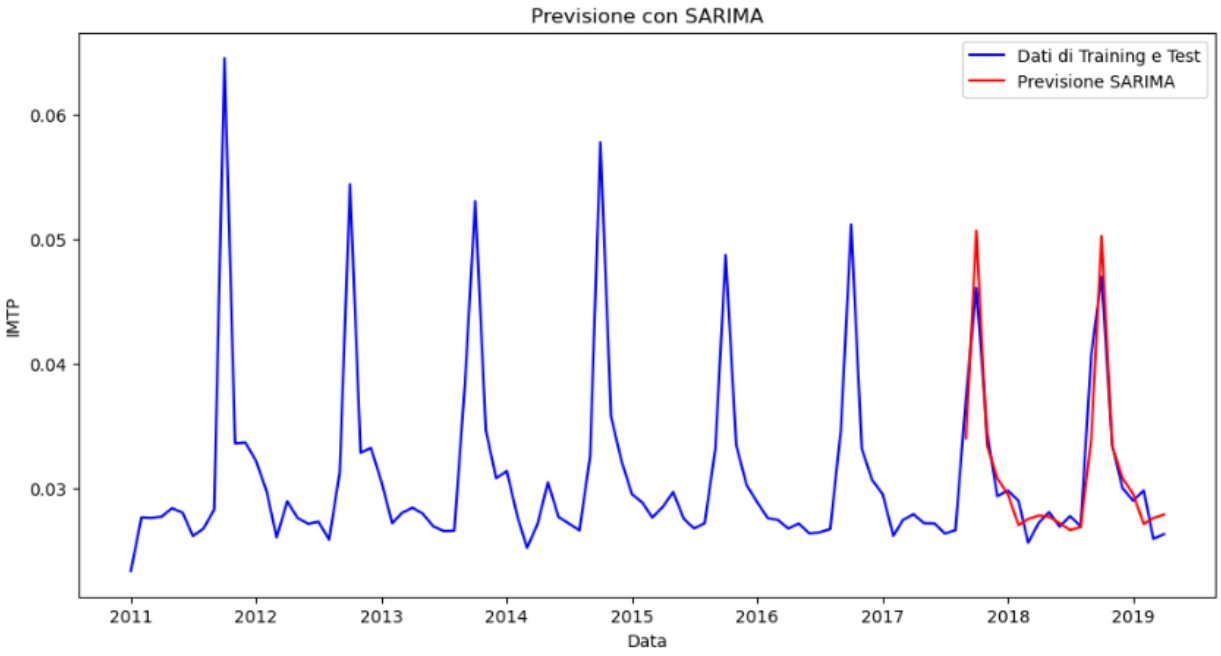

Figura 22 Applicazione di SARIMA all'IMTP del topic 13 "Halloween ed eventi speciali"

La linea blu rappresenta i dati storici utilizzati per il training e il test, mentre la linea rossa le previsioni effettuate dal modello.

Il modello cattura in modo eccellente la stagionalità dei dati, replicando i picchi e i cali in maniera regolare. La previsione è molto vicina ai valori originali, indicando che i parametri stagionali sono stati scelti in modo appropriato. Il valore dell'RMSE è basso in quanto il modello ci restituisce previsioni accurate:

Root Mean Square Error (RMSE) = 0,0024

Le discrepanze tra le previsioni e i dati reali sono minime e questo suggerisce un modello ben calibrato.

3.2.2.2 Applicazione LSTM

La Figura 23 mostra la previsione ottenuta mediante l'applicazione del metodo LSTM (si veda Sezione 2.5.3 e 2.5.4 ed Allegato 5).

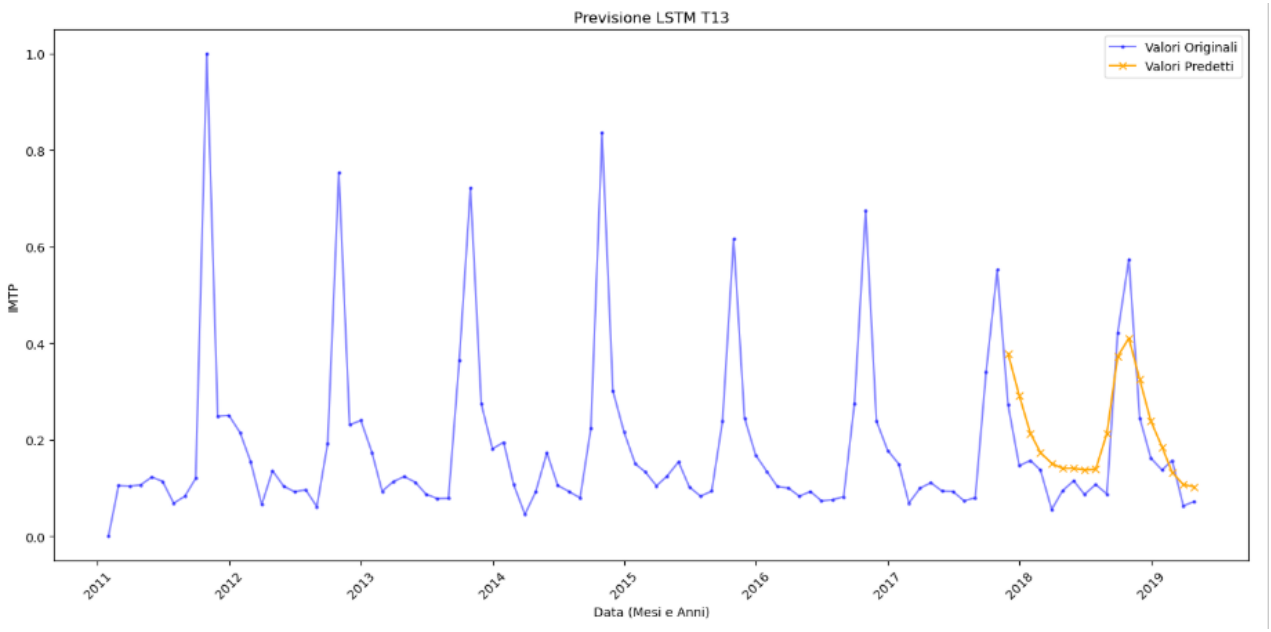

Figura 23 Applicazione di LSTM all'IMTP del topic 13 "Halloween ed eventi speciali"

La linea blu rappresenta i dati storici reali, mentre la linea arancione le previsioni effettuate.

Il modello riesce a catturare l'andamento generale e i pattern stagionali, ma presenta alcune discrepanze. I picchi sono approssimati e non hanno la stessa ampiezza dei dati reali. Di seguito il calcolo delle metriche di valutazione:

Mean Absolute Error (MAE)= 0,0028

Mean Square Error (MSE)= 1,062

Root Mean Square Error (RMSE)=0,0033

Le discrepanze tra i valori predetti e quelli reali sono maggiori rispetto il modello SARIMA, ed il valore del RMSE è più elevato rispetto al caso precedente.

Questo suggerisce che l'LSTM potrebbe non aver catturato completamente la complessità dei pattern stagionali della serie. Si potrebbe pensare di ottimizzare i parametri oppure di aumentare gli strati della rete neurale in quanto il pattern della serie risulta complesso.

3.2.3 Topic 5 "Esperienza di soggiorno e hotel"

La Figura 24 mostra l'andamento della metrica IMTP per il *topic 5 "Esperienza di soggiorno ed hotel"*.

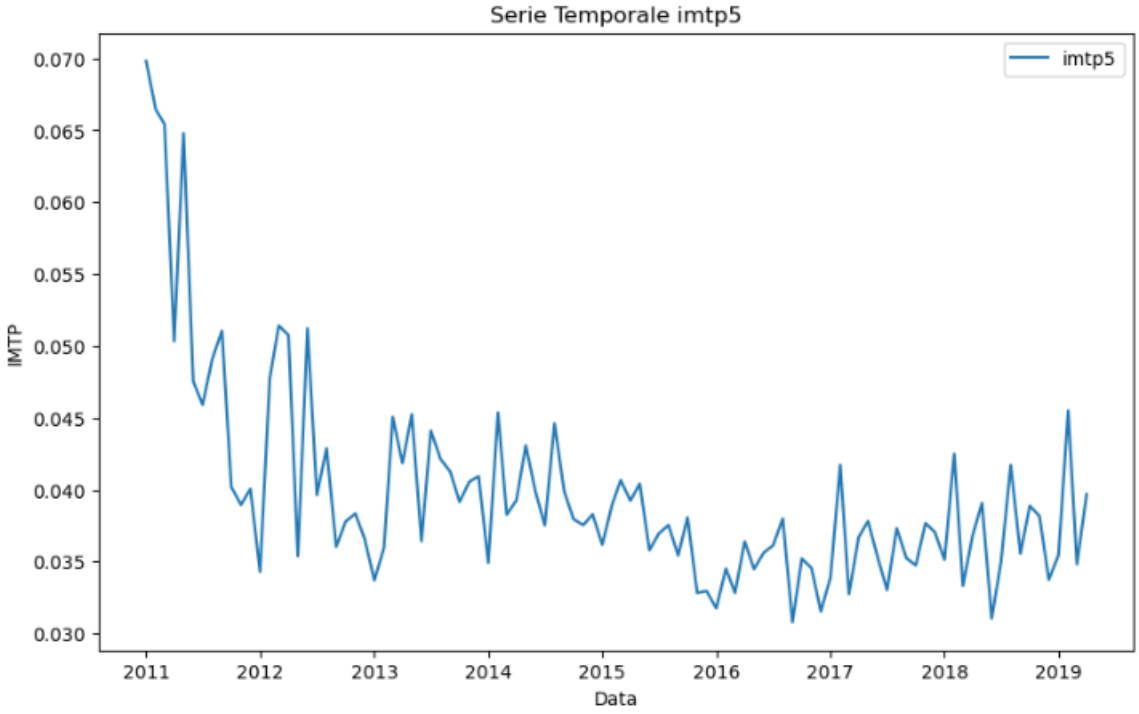

Figura 24 Rappresentazione grafica dell'IMTP del topic 5 "Esperienza di soggiorno ed hotel"

La Figura 25 mostra la decomposizione della serie in: trend/tendenza, stagionalità e residuo.

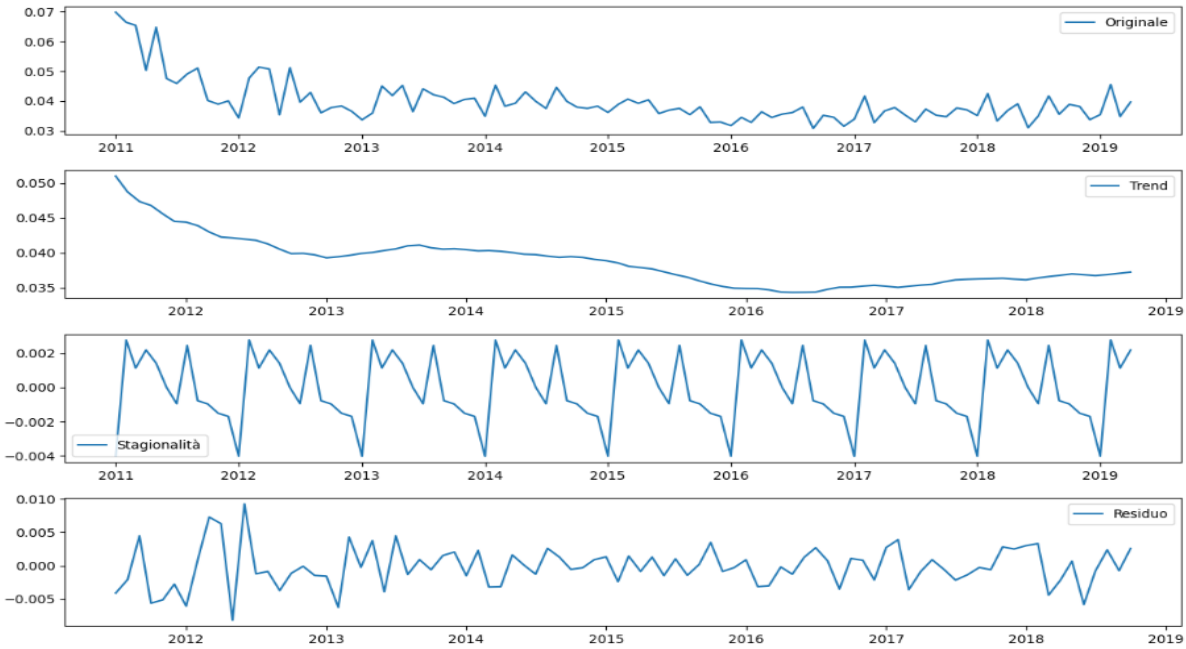

Figura 25 Decomposizione dell'IMTP del topic 5 "Esperienza di soggiorno ed hotel" in: 1. rappresentazione della serie originale, 2. rappresentazione del trend/tendenza della serie, 3. rappresentazione della stagionalità della serie, 4. rappresentazione del residuo della serie.

La Figura 25 mostra 4 diversi grafici:

1. la serie originale ha una diminuzione iniziale e poi un andamento stabile, con alcune fluttuazioni;

2. la tendenza risulta gradualmente decrescente nel tempo;

3. la componente stagionale presenta oscillazioni che si ripetono nel tempo, ma di piccola ampiezza rispetto ai valori originali della serie, segno di stagionalità debole;

4. i residui contengono una certa variabilità, segno che la componente stagionale non spiega completamente il modello.

3.2.3.1 Applicazione SARIMA

La Figura 26 mostra la previsione ottenuta mediante l'applicazione del metodo SARIMA (si veda Sezione 2.5.1 e 2.5.2 ed Allegato 4).

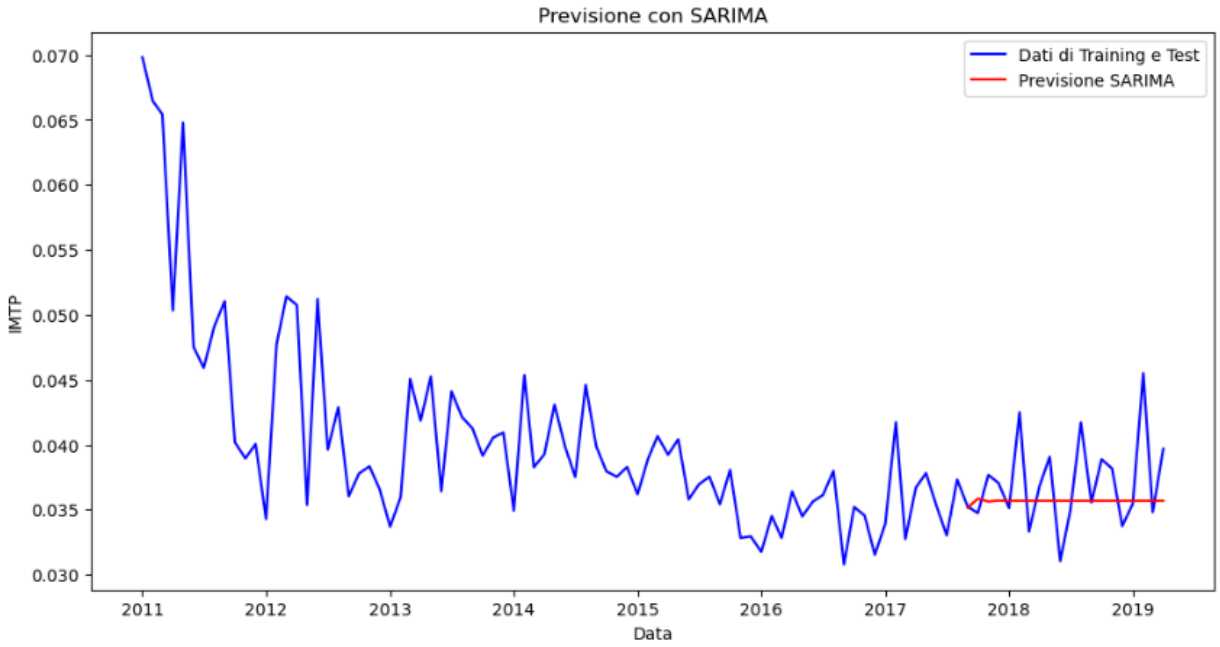

Figura 26 Applicazione di SARIMA all'IMTP del topic 5 "Esperienza di soggiorno ed hotel"

La linea blu rappresenta i dati storici reali, mentre la linea rossa le previsioni del modello. La previsione appare come una linea piatta, in quanto non cattura adeguatamente le variazioni presenti nella serie originale. Di seguito il calcolo della metrica RMSE:

Root Mean Square Error (RMSE)= 0,0036

Il valore dell'RMSE si discosta poco dai valori originali, sebbene il modello non colga il pattern sotteso ai dati della serie. In questo caso il modello SARIMA, non essendoci una chiara stagionalità, non cattura una tendenza ripetitiva nel tempo ed approssima la previsione ad una media dei valori.

3.2.3.2 Applicazione LSTM

La Figura 27 mostra la previsione ottenuta mediante l'applicazione del metodo LSTM (si veda Sezione 2.5.3 e 2.5.4 ed Allegato 5).

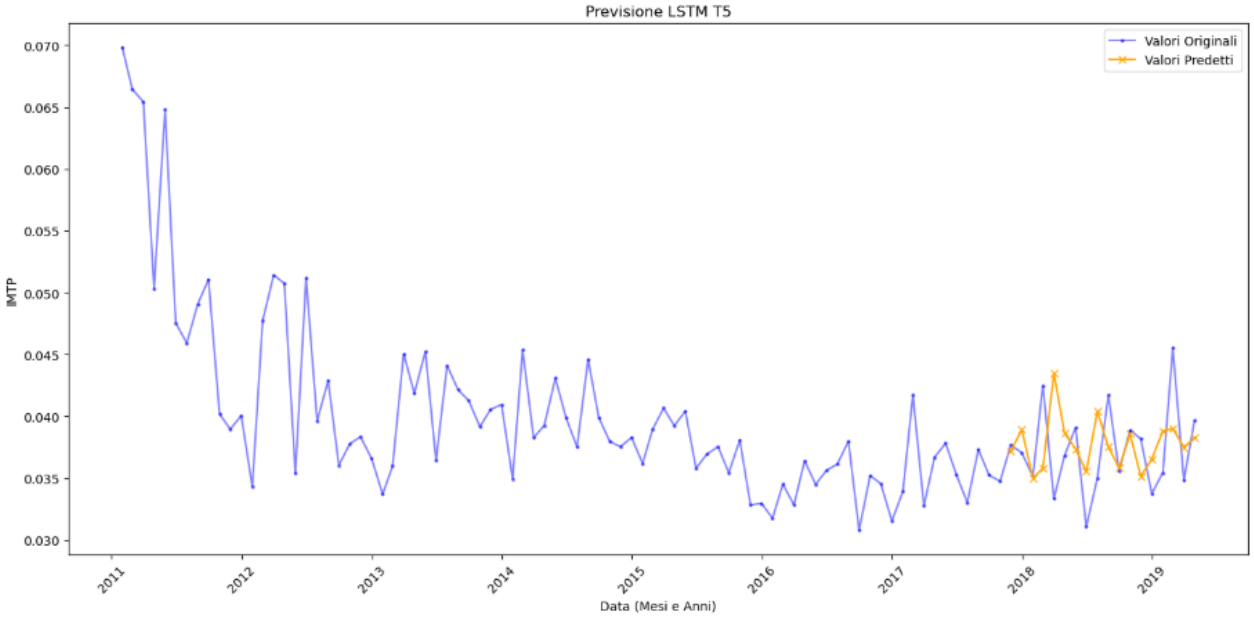

Figura 27 Applicazione di LSTM all'IMTP del topic 5 "Esperienza di soggiorno ed hotel"

La linea blu rappresenta i dati storici reali, mentre la linea arancione i valori previsti dal modello.

Le previsioni seguono abbastanza il comportamento generale dei dati reali, catturandone la tendenza generale. Di seguito il calcolo delle metriche per valutare l'accuratezza del modello:

Mean Absolute Error (MAE)= 0,0037

Mean Square Error (MSE)= 2,23

Root Mean Square Error (RMSE)= 0,0047

Il valore dell'RMSE è maggiore rispetto il modello SARIMA, ma osservando il grafico vediamo come tale previsione cerca di adattarsi maggiormente al pattern dei dati, sebbene con alcune discrepanze nella profondità delle valli o altezza dei picchi.

4 Capitolo 4: analisi delle recensioni di Ryanair

Il secondo campione di recensioni analizzato è relativo alla compagnia aerea Ryanair e presenta informazioni che valutano l'esperienza complessiva dei clienti con una finestra temporale che va da 08-2012 a 02-2024.

I metadati associati al campione sono i seguenti:

- Overall Rating: valutazione dell'esperienza vissuta con punteggio che va da 1 a 10;
- Passenger Country: paese di provenienza dei clienti;
- Trip Verified: verifica dell'utilizzo del servizio offerto da parte del cliente;
- Comment Title & Comment: titolo e recensione scritta dall'utente;
- Aircraft: velivolo relativo alla tratta;
- Type of traveller: famiglia, coppia, single, lavoratore;
- Seat Type: tipologia di comfort scelto, economy o business;
- Origine & Destinazione della tratta;
- Date Flown: data del volo;
- Seat Comfort: rating sulla comodità del viaggio;
- Cabin Staff Service: rating sul servizio offerto dallo staff (hostess, stewart, pilota);
- Food & Beverage: rating sulla disponibilità e qualità su cibo e bevande;
- Ground Service: rating dei servizi offerti come assistenza invalidi, trasferimenti, bagagli, ect;
- Value For Money: rating relativo al rapporto qualità prezzo, influenzato dal soddisfacimento delle richieste del cliente;
- Recommended: opzione che indica se il cliente consiglia l'utilizzo del servizio o meno;
- Inflight Entertainment: rating relativo ai servizi di intrattenimento offerti come film, giochi, riviste, ect;
- Wi-fi & Connectivity: rating della qualità della connessione presente sul velivolo.

Gli obiettivi dell'analisi sono:

- valutare le determinanti di qualità applicando STM alla Digital VoC e
- condurre analisi predittive sull'Interval Mean Topical Prevalence (IMTP) per ciascun topic mediante i metodi SARIMA ed LSTM precedentemente introdotti (si veda Sezione 2.5).

4.1 Applicazione STM

Implementando lo script (si veda Allegato 1) nell'ambiente R, mediante la funzione di pre-processamento il testo delle recensioni è stato semplificato rimuovendo, ad esempio, le parole che presentavano una bassa frequenza e le "stopword" ovvero quelle non necessarie per l'individuazione dei topic.

Applicando la funzione search-k, la Figura 28 rappresenta il valore della metrica Held-out likelihood scelta per definire il numero di topic ottimale (si veda Sezione 1.3.2).

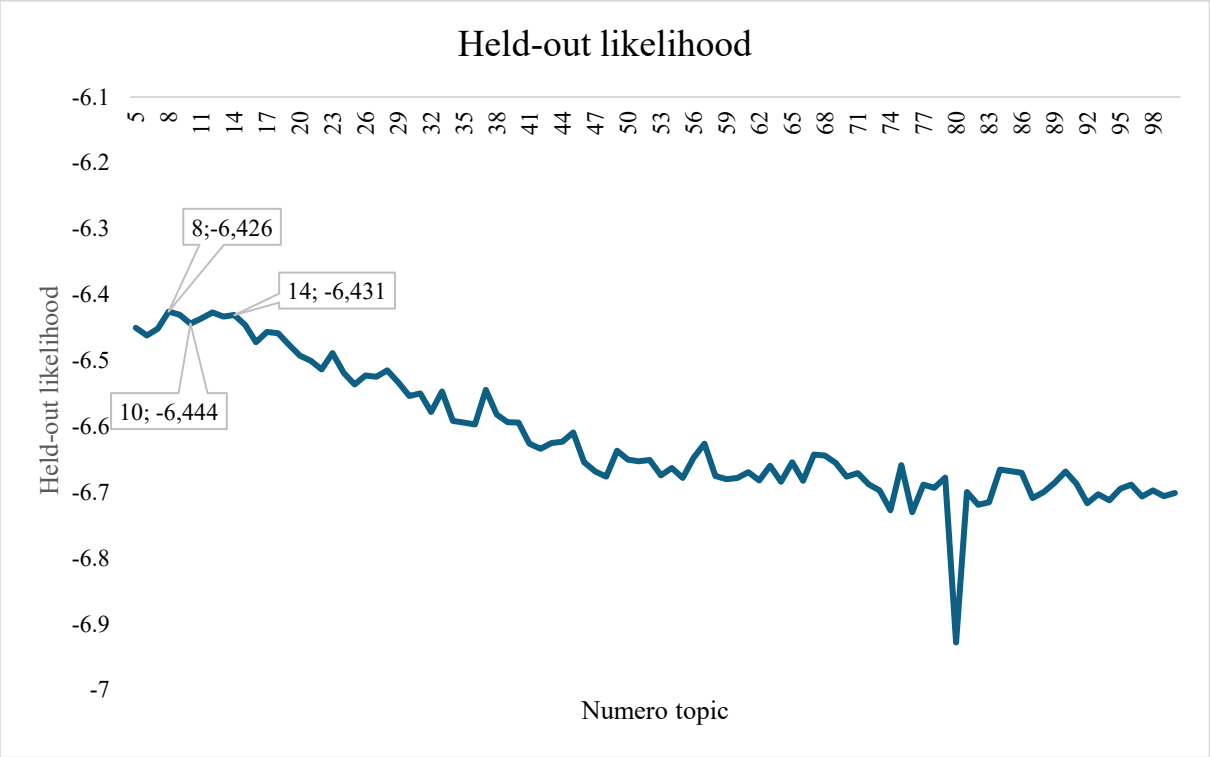

Figura 28 Andamento dell'Held-out likelihood caso studio Ryanair

L'andamento in questione è diverso da quello tipico riscontrato normalmente nella pratica. In particolare, i punti di massimo sottolineati corrispondono a numeri di topic compresi tra 8 e 14. A valle di una serie di passaggi esplorativi sul software R si è scelto un numero di topic pari a dieci.

Il risultato di STM per Ryanair ha permesso di individuare la topical content e la topical prevalence, in particolare partendo dalla topical content, si è data un'etichetta a ciascun topic tenendo conto delle parole che avevano la massima probabilità (si veda Allegato 3 e Tabella 20).

La Tabella 20 presenta una descrizione dettagliata dei topic e delle parole chiave identificate dall'algoritmo.

Etichette topic	Parole chiave	Descrizione
1. Attesa dell'imbarco	plane, flight, attend, wait, minut, board, water, staff, passeng, return	Il topic tratta dell'attesa dei passeggeri prima di salire a bordo, i quali spesso consumano cibo e bevande ed interagiscono con lo staff. A seconda del flusso dei passeggeri e dall'efficienza del personale vi sono variazioni dei tempi di attesa.
2. Prenotazione	flight, time, crew, friend, cabin, return, board, price, arriv, servic	Il topic tratta della fase di prenotazione che spesso avviene mesi prima di eventi o festività. I clienti cercano e confrontano su varie piattaforme il prezzo più vantaggioso.
3. Esperienza di volo	crew, passeng, cabin, staff, flight, help, member, assist, ask, aircraft	Il topic tratta dell'esperienza di volo, la quale spesso è influenzata dal prezzo pagato e dalla qualità del servizio offerto. I clienti valutano positivamente la puntualità e l'assenza di problematiche nel momento in cui usufruiscono del servizio.
4. Equipaggio	crew, passeng, cabin, staff, flight, help, member, assist, ask, aircraft	Il topic tratta della disponibilità dell'equipaggio nel rispondere alle esigenze dei passeggeri a bordo.
5. Gestione del bagaglio e tariffe aggiuntive	bag, luggag, pay, extra, hand, baggag, staff, paid, small, check	Il topic tratta della gestione della compagnia aerea dei bagagli nella fase di imbarco e della possibilità del pagamento di tariffe aggiuntive nel caso in cui il bagaglio non rispetti i canoni previsti.
6. Puntualità	flight, hour, delay, time, airport, arriv, stanst, late, fli, inform	Il topic tratta della comunicazione chiara e trasparente del personale, in particolare nel gestire ritardi o eventuali problematiche.
7. Imbarco e controlli di sicurezza	board, gate, prioriti, queue, wait, minut, secur, time, flight, long	Il topic tratta dell'organizzazione per i controlli e l'imbarco. È necessario garantire una buona organizzazione per ridurre le attese e garantire un servizio efficiente.
8. Check-in e pagamenti extra	check, pass, airport, onlin, pay, print, board, charg	Il topic tratta delle spese aggiuntive che i clienti possono sostenere nel momento in cui effettuano il check-in online, ad esempio scelta del posto o aggiunta di un bagaglio. Inoltre, tale topic tratta del servizio online della gestione della prenotazione e della stampa delle carte d'imbarco.
9. Servizio clienti	flight, custom, servic, refund, book, chang, tri, email, cancel, compani	Viene valutata la disponibilità dell'assistenza clienti e l'efficacia nella risoluzione delle problematiche quali ad esempio rimborsi o cancellazioni.
10. Assegnazione e comfort posti	seat, sit, flight, togeth, pay, plane, alloc, row, extra, next	Il topic tratta dell'assegnazione dei posti che può essere casuale oppure scelta dal passeggero, ed in questo caso, rappresenta un costo aggiuntivo.

Tabella 20 Labeling caso studio Ryanair

4.1.1 Validazione del modello

Considerando 100 recensioni estratte in maniera casuale il confronto tra l'assegnazione umana ed automatica ha condotto i risultati presenta in Tabella 21 (si veda Sezione 1.3.5):

		Assegnazione umana dei topic (true condition)	
		Presenza Ti	Assenza Ti
Assegnazione automatica dei topic	Presenza Ti	True positive (tp) 109	False positive (fp) 16
	Assenza Ti	False negative (fn) 39	True negative (tn) 836

Tabella 21 Confronto assegnazione manuale ed automatica caso studio Ryanair

Per valutare la bontà del modello sono state calcolate le metriche come mostrato in Tabella 22.

Indicatori	Valori	Valori target
Precision	0,87	>0,70
False omission	0,04	<0,05
False discovery rate	0,13	<0,05
Negative predictive value	0,96	>0,90
Accuracy	0,95	>0,95
Recall	0,74	>0,70
Fall-out	0,02	<0,05
Miss rate	0,26	<0,20
Specificity	0,98	>0,90
F score	0,8	>0,70

Tabella 22 Calcolo delle metriche di validazione caso studio Ryanair

Tutti gli indicatori, eccetto due, rientrano nei valori soglia stabiliti, questo è simbolo di un'accurata analisi che ha portato i risultati sperati. Gli indicatori *False discovery rate* e *Miss rate* si discostano rispettivamente di 0,08 e di 0,06. Entrambi sono dipendenti dal fattore *true positive*, presente al denominatore, che presumibilmente è stato sovrastimato (si veda Sezione1.3.5.1.4).

Dall'analisi risulta che le etichette scelte sono rappresentative dei topic, i parametri di input sono corretti ed è presente pertinenza e omogeneità nel database di riferimento.

4.1.2 Analisi dei risultati

La matrice topical prevalence ha permesso di sviluppare analisi quantitative e qualitative per il monitoraggio della qualità e per comprendere la soddisfazione o insoddisfazione del cliente [1].

Sui risultati STM sono state svolte le seguenti analisi:

(i) Calcolo del Mean Topic Proportion (MTP) (si veda Sezione 1.3.6.1);
(ii) Calcolo Mean Rating Proportion (MRP) (si veda Sezione 1.3.6.3);
(iii) Key Attributes VoC Map (KA-VoC Map) (si veda Sezione 1.3.6.4);
(iv) Calcolo Interval Mean Topical Prevalence (IMTP) (si veda Sezione 1.3.6.2).

4.1.2.1 Calcolo Mean Topic Proportion (MTP)

È stato calcolato il valore dell'MTP, per individuare quanto viene mediamente discusso ciascun topic.

La Figura 29 misura il risultato ottenuto.

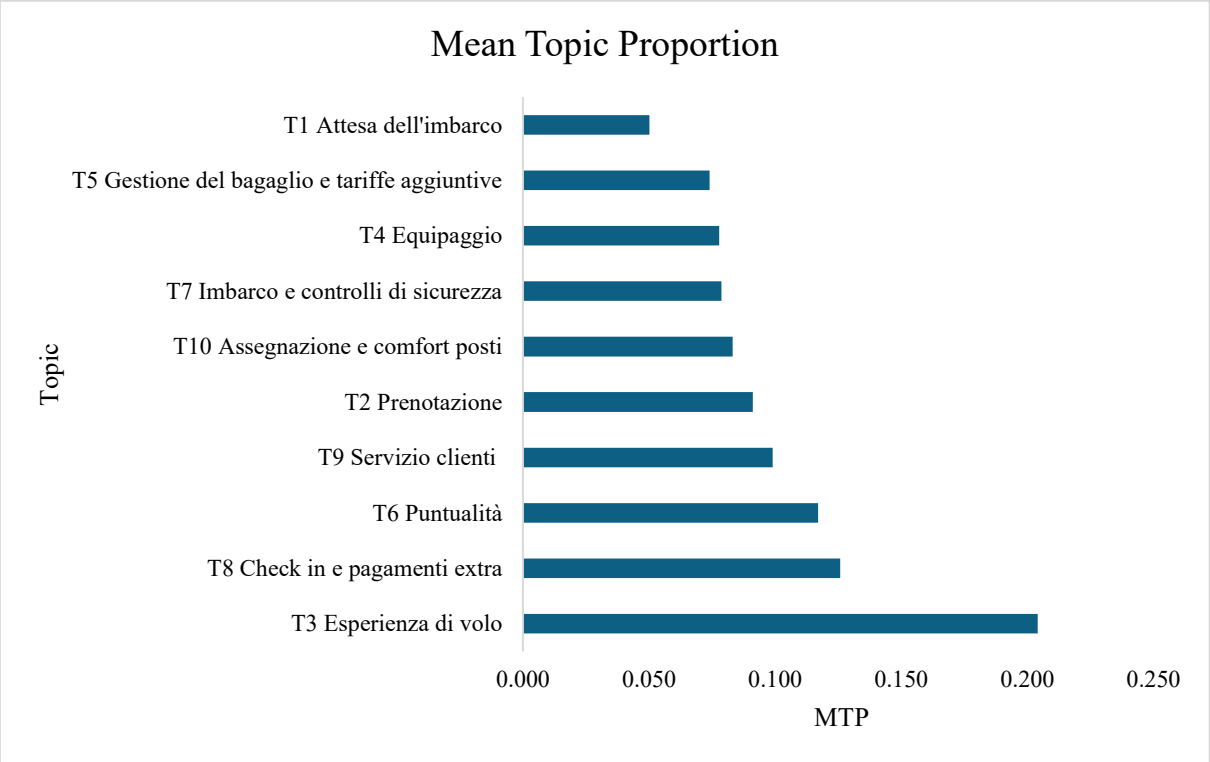

Figura 29 Analisi MTP caso studio Ryanair

Osservando la Figura 29 emerge che i topic maggiormente discussi, nell'ordine, sono: *(T3)"Esperienza di volo",(T8)" Check-in e pagamenti extra", (T6) "Puntualità".*

4.1.2.2 Calcolo Mean Rating Proportion (MRP)

Il rating associato ai documenti testuali ha permesso di calcolare il valore del MRP.

La Figura 30 mostra il profilo di MRP per ciascun topic individuato.

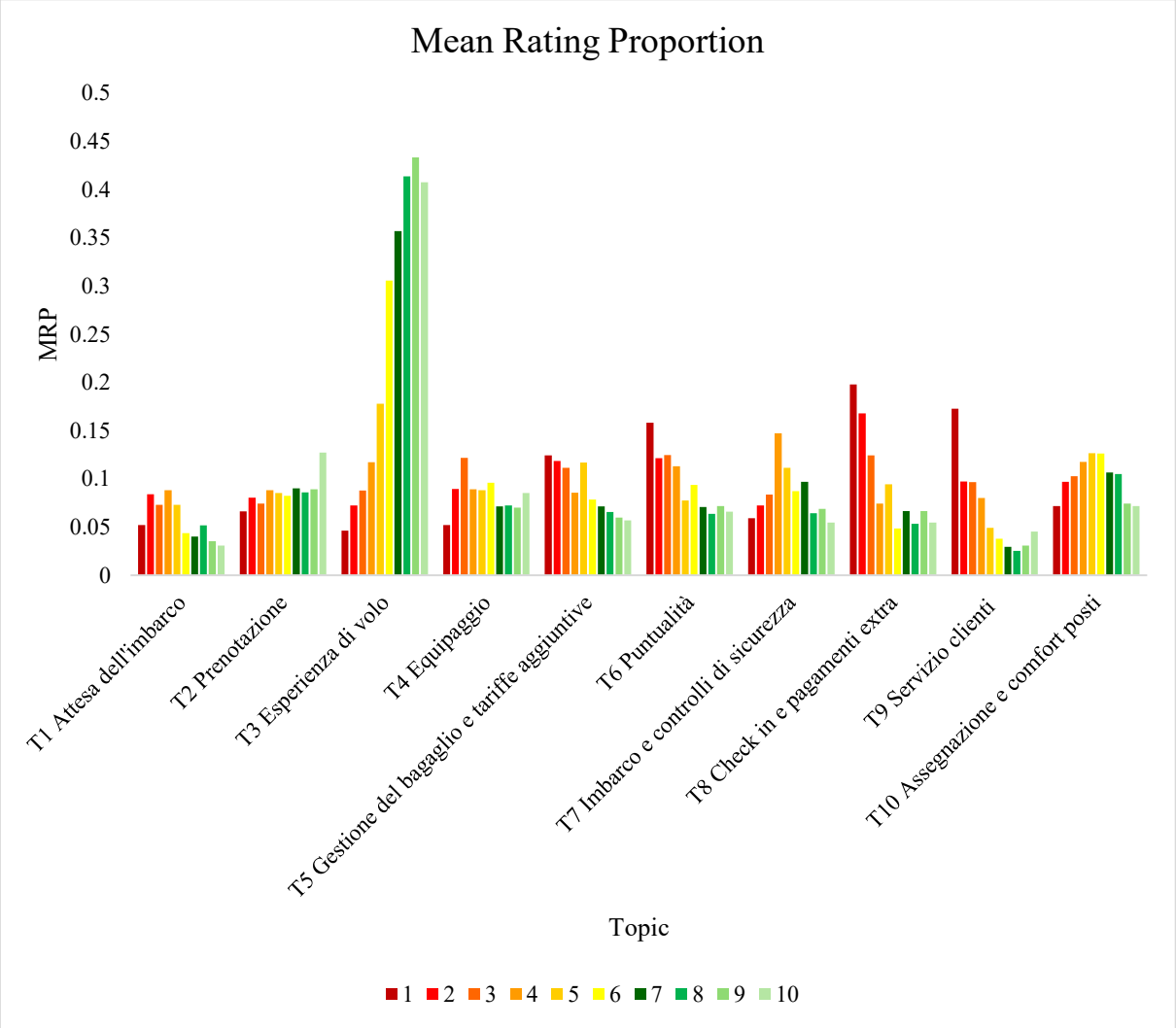

Figura 30 Analisi MRP caso studio Ryanair

In questo caso il rating è espresso su una scala a 10 livelli, ed è stato possibile classificare i profili come segue:

- profilo positivo: "*Esperienza di volo" (T3);*
- profilo negativo: "*Attesa dell'imbarco" (T1)," Gestione dei bagagli e tariffe aggiuntive" (T2)," Puntualità" (T6)," Check-in e pagamenti extra" (T8)," Servizio clienti" (T9);*
- profilo neutro: *"Prenotazione" (T2), "Equipaggio" (T4)," Imbarco e controlli di sicurezza" (T7)," Assegnazione e comfort posti" (T10).*

4.1.2.3 Key Attributes VoC Map (KA-VoC Map)

Tramite i risultati ottenuti dal MRP ed MTP si è costruita la Ka-VoC Map.

Sono stati considerati come:

1. poco discussi i topic con un $MTP < \frac{1}{n}$ (dove n è il numero di topic) e
2. molto discussi i topic con un $MTP \geq \frac{1}{n}$.

Nel caso in esame tale rapporto è pari a $\frac{1}{\mathrm{n}} = 0{,}1$ ed i topic sono stati suddivisi in:

- molto discussi: *3) Esperienza di volo, 8) Check-in e pagamenti extra,6) Puntualità*;
- poco discussi: *2) Prenotazione, 4) Equipaggio,7) Imbarco e controlli di sicurezza, 10) Assegnazione e comfort posti, 1) Attesa dell'imbarco, 5) Gestione dei bagagli e tariffe aggiuntive, 9) Servizio clienti*.

La Tabella 23 mostra la KA-VoC Map.

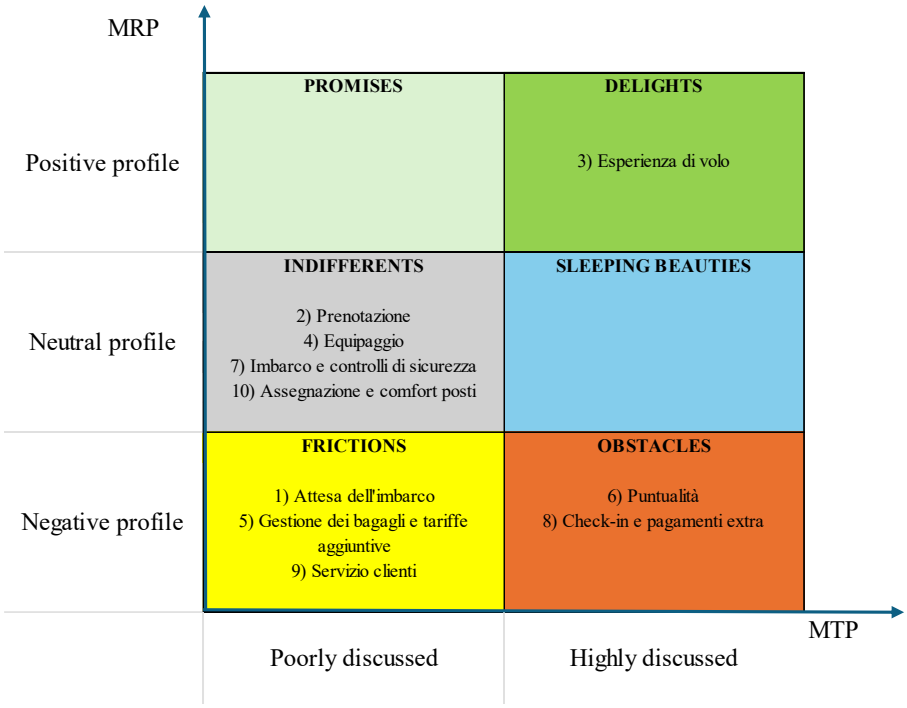

Tabella 23 KA VoC Map caso studio Ryanair

Le categorie della KA-VoC Map sono:

- Obstacles: *6) Puntualità, 8) Check-in e pagamenti extra*. Sono attributi molto discussi e fonte di insoddisfazione. Questi attributi rappresentano le principali fonti di critica e insoddisfazione da parte dei clienti.

- Frictions: *1) Attesa dell'imbarco, 5) Gestione dei bagagli e tariffe aggiuntive, 9) Servizio clienti*. Scarsamente discussi e fonte di insoddisfazione. Questi rappresentano problemi meno gravi, ma possono comunque generare insoddisfazione nei clienti.

- Indifferents: *2) Prenotazione, 4) Equipaggio, 7) Imbarco e controlli di sicurezza, 10) Assegnazione e comfort posti*. Sono attributi scarsamente discussi che sono neutrali rispetto alla soddisfazione del cliente. Sono considerati non rilevanti poiché non hanno un'influenza chiara sulla soddisfazione o insoddisfazione.

- Sleeping Beauties: nessun topic. Sarebbero attributi molto discussi e neutri rispetto alla soddisfazione del cliente. Rappresentano dimensioni essenziali o caratteristiche di base che non impressionano né positivamente né negativamente i clienti.

- Promises: nessun topic. Sarebbero attributi scarsamente discussi che generano soddisfazione del cliente. Queste dimensioni rappresentano vantaggi minori o attributi emergenti del prodotto o servizio analizzato.

- Delights: *3) Esperienza di volo*. Sono attributi molto discussi e che generano un alto grado di soddisfazione. Questi attributi sorprendono positivamente i clienti, superando le loro aspettative e aumentando significativamente la loro soddisfazione.

4.1.2.4 Calcolo Interval Mean Topical Prevalence (IMTP)

Il profilo dell'IMTP viene classificato nel modo seguente:

- decrescente: *2) Prenotazione, 3) Esperienza di volo, 10) Assegnazione e comfort posti*
- crescente: *6) Puntualità, 8) Check-in e pagamenti extra, 9) Servizio clienti*
- stazionario: *1) Attesa dell'imbarco, 4) Equipaggio, 5) Gestione dei bagagli e tariffe aggiuntive, 7) Imbarco e controlli di sicurezza.*

La Figura 31 mostra l'andamento dell'IMTP decrescente per i topic in esame.

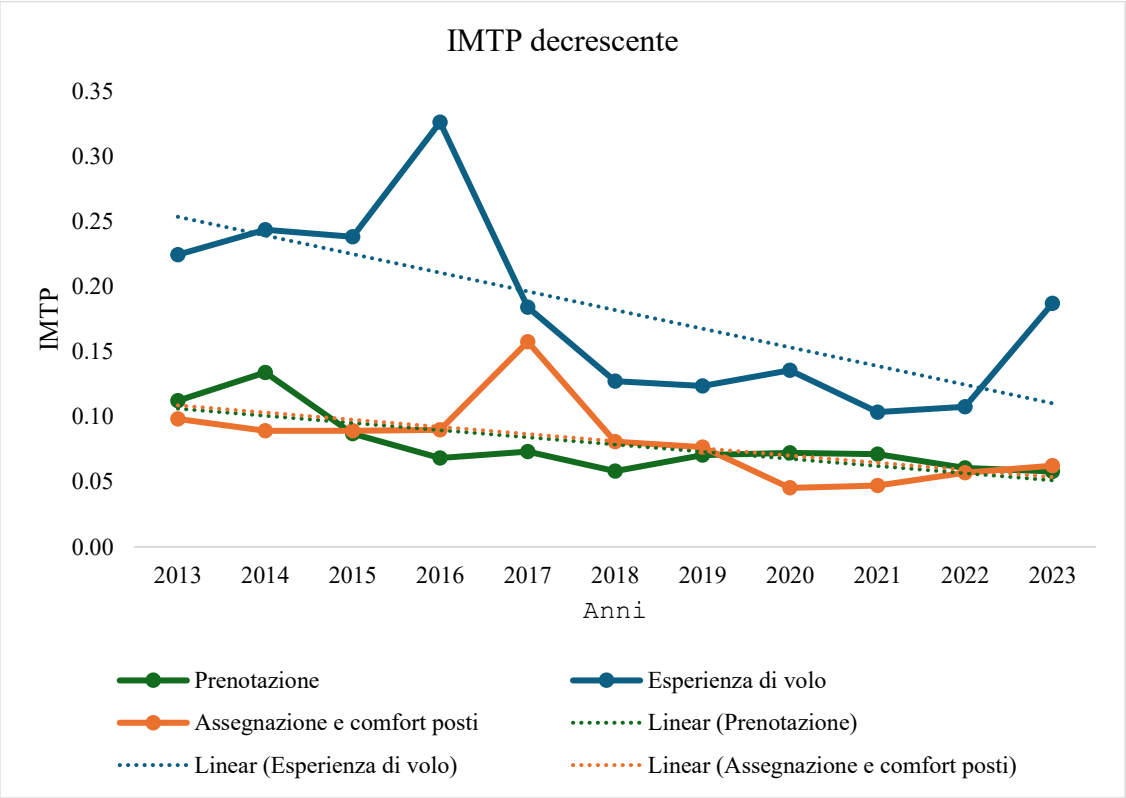

Figura 31 Analisi IMTP decrescente caso studio Ryanair

La Figura 32 mostra l'andamento dell'IMTP crescente per i topic in esame.

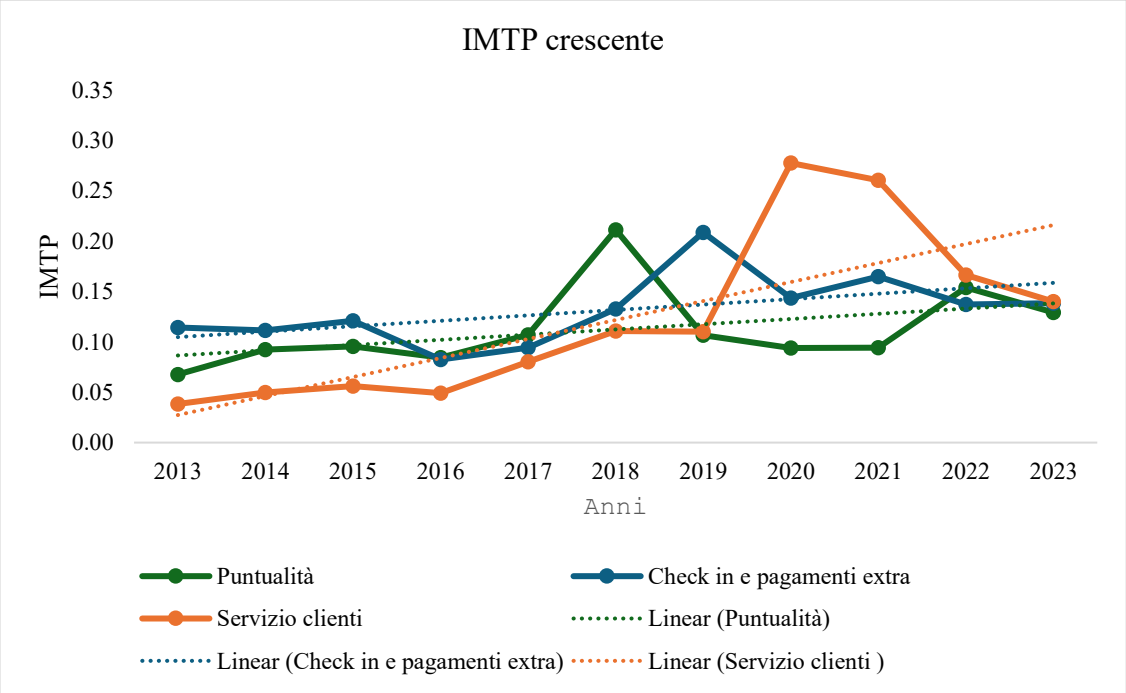

Figura 32 Analisi IMTP crescente caso studio Ryanair

La Figura 33 mostra l'andamento dell'IMTP stazionario per i topic in esame.

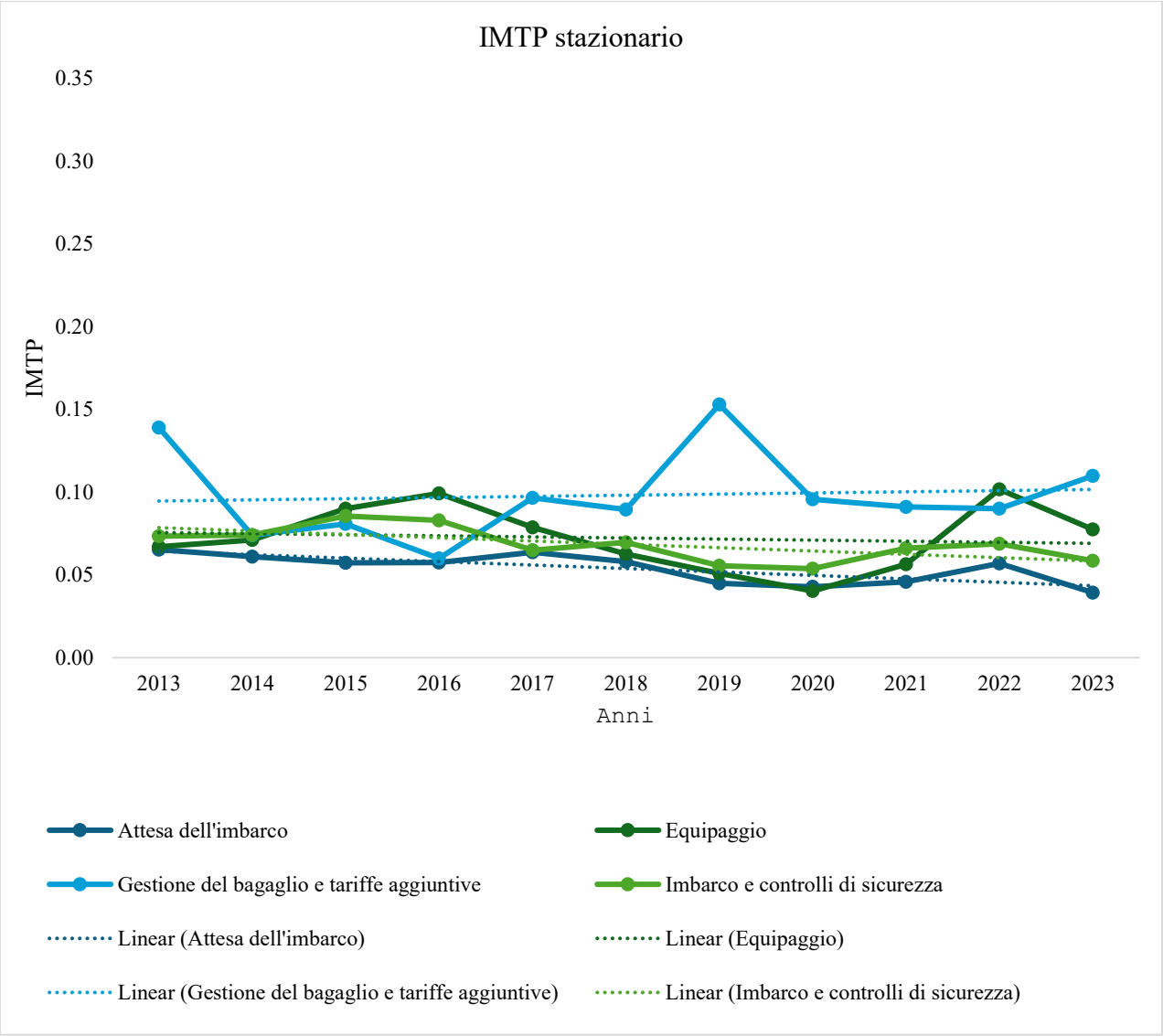

Figura 33 Analisi IMTP stazionario caso studio Ryanair

4.2 Analisi predittive sull'Interval Mean Topical Prevalence (IMTP)

Ottenuti i profili dell'andamento dell'IMTP e classificati opportunatamente, è possibile condurre analisi predittive su serie storiche.

L'implementazione di tale procedura richiede una valutazione delle predizioni mediante i metodi descritti precedentemente (si veda Sezione 2.5.2 e 2.5.4).

La Tabella 24 mostra una piccola parte del dataset campione del modello per le analisi predittive.

date	IMTP1	IMTP2	IMTP3	IMTP4	IMTP5	IMTP6	IMTP7	IMTP8	IMTP9	IMTP10	SOMMA
2013-08	0,27	0,12	0,28	0,02	0,10	0,02	0,06	0,02	0,01	0,10	1,00
2013-09	0,06	0,15	0,17	0,05	0,16	0,07	0,05	0,13	0,04	0,11	1,00
2013-10	0,05	0,10	0,22	0,09	0,11	0,08	0,11	0,11	0,05	0,07	1,00
2013-11	0,05	0,07	0,32	0,04	0,10	0,05	0,06	0,21	0,03	0,07	1,00
2013-12	0,05	0,13	0,32	0,07	0,10	0,07	0,09	0,04	0,02	0,11	1,00
2014-1	0,03	0,14	0,23	0,07	0,05	0,16	0,08	0,05	0,11	0,08	1,00
2014-2	0,03	0,13	0,26	0,07	0,07	0,06	0,05	0,11	0,03	0,18	1,00
2014-3	0,05	0,12	0,38	0,06	0,08	0,09	0,04	0,07	0,06	0,07	1,00
2014-4	0,04	0,16	0,20	0,06	0,10	0,09	0,05	0,15	0,06	0,11	1,00
2014-5	0,02	0,16	0,19	0,04	0,08	0,12	0,05	0,14	0,14	0,05	1,00
2014-6	0,05	0,14	0,19	0,04	0,09	0,08	0,08	0,19	0,05	0,09	1,00
2014-7	0,08	0,12	0,25	0,08	0,06	0,17	0,07	0,07	0,05	0,06	1,00
2014-8	0,08	0,13	0,16	0,08	0,09	0,09	0,12	0,13	0,03	0,08	1,00
2014-9	0,12	0,13	0,24	0,10	0,05	0,12	0,07	0,06	0,03	0,08	1,00
2014-10	0,06	0,13	0,29	0,11	0,07	0,06	0,08	0,10	0,02	0,08	1,00
2014-11	0,05	0,13	0,34	0,09	0,06	0,05	0,10	0,05	0,02	0,12	1,00
2014-12	0,02	0,12	0,34	0,03	0,04	0,15	0,07	0,11	0,02	0,09	1,00
2015-1	0,11	0,12	0,26	0,08	0,06	0,07	0,12	0,07	0,03	0,07	1,00

Tabella 24 Input dei modelli predittivi caso studio Ryanair

Per quanto riguarda il caso studio Ryanair, si è scelto di svolgere le analisi predittive in base alla metrica IMTP per tre topic in esame:

- *9) "Servizio clienti*" che presenta un andamento dell'IMTP crescente;
- *3) "Esperienza di volo"* che presenta un andamento dell'IMTP decrescente;
- *7) "Imbarco e controlli di sicurezza"* che presenta un andamento dell'IMTP stazionario.

In questo caso l'arco temporale va dal 2012 sino al 2024 e comprende il periodo del COVID-19. La scelta della suddivisione del dataset in training e test è stata fatta in base alla manifestazione della pandemia considerata un fattore esogeno.

Nella fase di addestramento dei modelli, si è scelto di ridurre il set di training in quanto, addestrare il modello su dati del periodo compreso tra il 2020-2021 poteva condurre a stime poco accurate. Il 60% del dataset è stato utilizzato per la fase di training, di modo da addestrare il modello sui dati prima dell'avvento della pandemia ed il restante 40% dei dati come test per valutare la capacità del modello di effettuare previsioni accurate.

4.2.1 Topic 9 "Servizio clienti"

La Figura 34 rappresenta il grafico dell'IMTP del *topic 9 "Servizio clienti*".

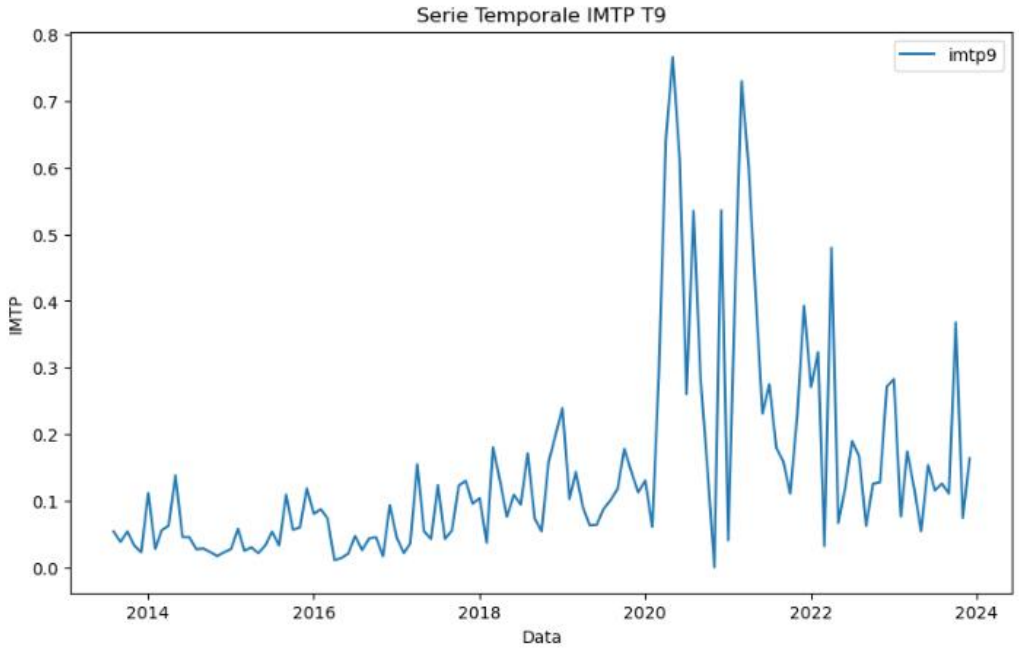

Figura 34 Rappresentazione grafica dell'IMTP del topic 9 "Servizio clienti"

I picchi nel grafico sono dovuti alla manifestazione del COVID-19 nel periodo 2020-2021.

Durante la pandemia, molte compagnie aeree hanno affrontato disagi operativi come cancellazioni di voli, cambiamenti di regolamenti e gestione dei rimborsi [45]. Questi eventi hanno aumentato notevolmente le interazioni tra i clienti e il servizio clienti, alimentando la discussione [45].

Dopo il 2021, con la ripresa dei voli e la stabilizzazione delle operazioni, il volume delle discussioni si è ridotto e ciò viene letto positivamente indicando una diminuzione delle interazioni legate a problematiche operative straordinarie.

La Figura 35 raffigura la decomposizione della serie temporale in: trend/tendenza, stagionalità e residuo.

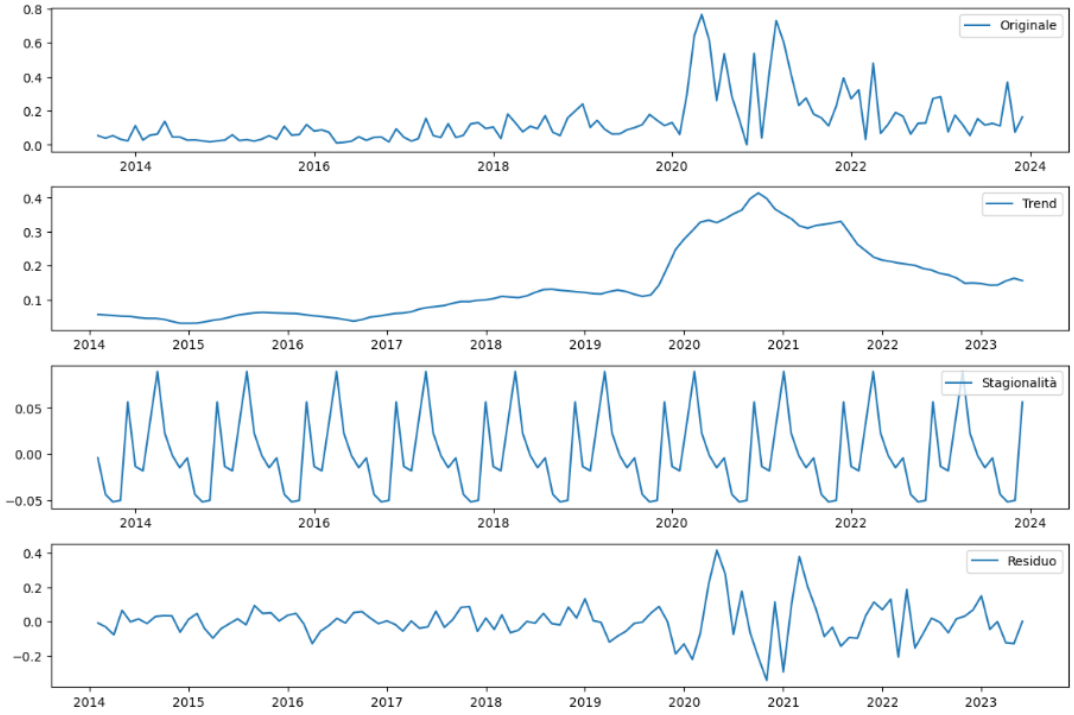

Figura 35 Decomposizione grafica dell'IMTP del topic 9 "Servizio clienti" in: 1. rappresentazione della serie originale, 2. rappresentazione del trend/tendenza della serie, 3. rappresentazione della stagionalità della serie, 4. rappresentazione del residuo della serie.

La Figura 35 mostra 4 diversi grafici:

1. la serie ha una significativa variabilità, con picchi evidenti intorno al 2020-2021;

2. la tendenza è gradualmente crescente fino al 2021 poi decresce;

3. la stagionalità presenta oscillazioni ben definite e regolari, e risulta moderata se confrontata con il valore della serie originale;

4. i residui presentano una notevole variabilità, specialmente nei periodi di picco, indicando che la stagionalità e la tendenza non spiegano completamente possibili eventi anomali che si manifestano in questo arco temporale.

4.2.1.1 Applicazione SARIMA

La Figura 36 mostra la previsione ottenuta mediante l'applicazione del metodo SARIMA (si veda Sezione 2.5.1 e 2.5.2 ed Allegato 4).

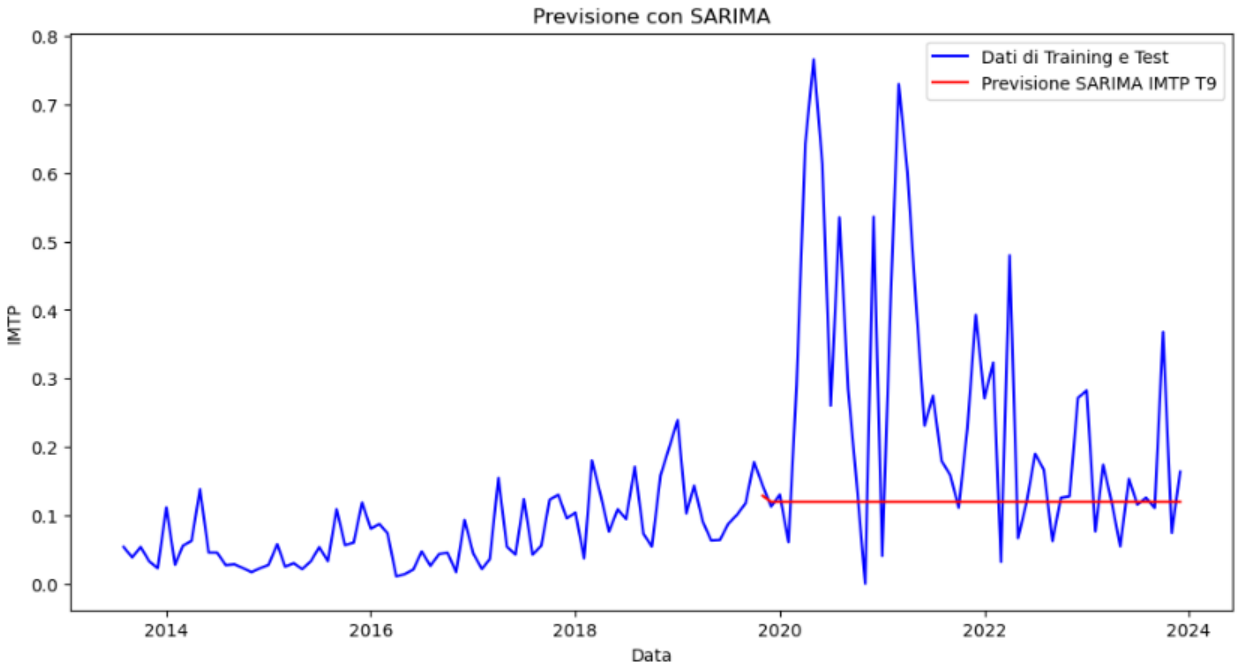

Figura 36 Applicazione di SARIMA all'IMTP del topic 9 "Servizio clienti"

La linea blu rappresenta i dati reali, mentre la linea rossa rappresenta le previsioni del modello SARIMA.

La previsione è piatta, indicando che il modello non è stato in grado di prevedere l'andamento dei dati della serie, in quanto non riesce a catturare una stagionalità forte nei dati. Il risultato ottenuto, inoltre, è dovuto alla presenza di un pattern nei dati complesso, che il modello SARIMA non riesce a catturare. Di seguito il calcolo dell'errore di previsione:

Root Mean Square Error (RMSE) = 0,23

La metrica è decisamente più alta rispetto ai valori riscontrati nel caso studio precedente (si veda Capitolo 3), segno che il modello non riesce a condurre previsioni accurate sui dati.

4.2.1.2 Applicazione LSTM

La Figura 37 mostra la previsione ottenuta mediante l'applicazione del metodo LSTM (si veda Sezione 2.5.3 e 2.5.4 ed Allegato 5).

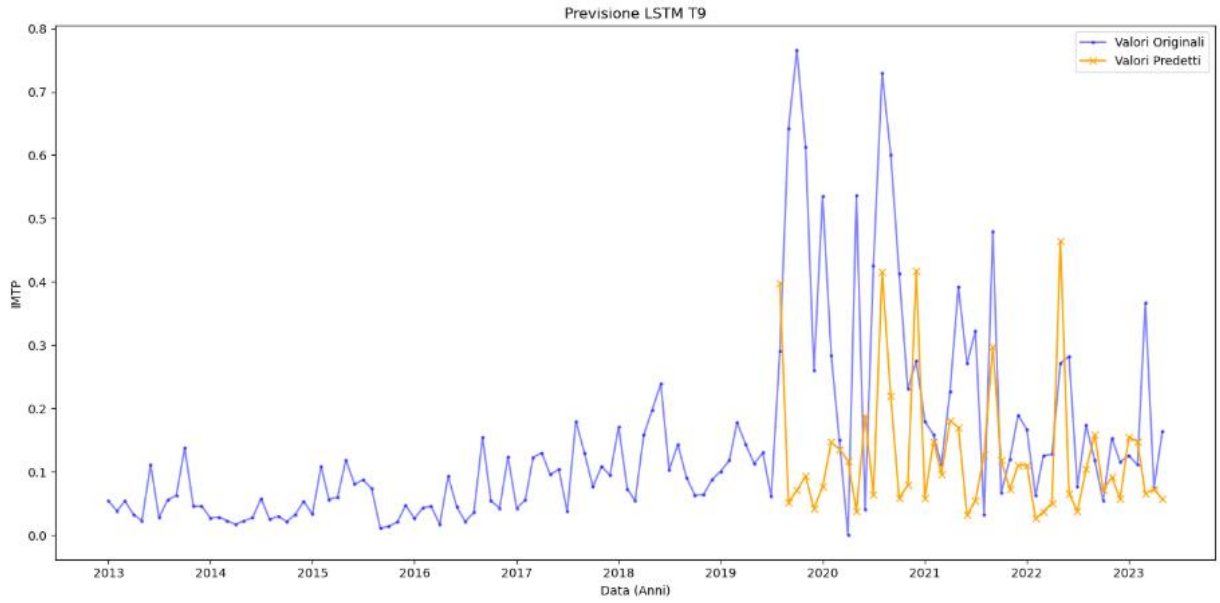

Figura 37 Applicazione di LSTM all'IMTP del topic 9 "Servizio clienti"

La linea blu rappresenta i dati storici, mentre la linea arancione rappresenta i valori previsti dal modello LSTM.

Il modello segue i picchi e le fluttuazioni dei dati reali e riesce a catturarne la tendenza generale, sia nella crescita iniziale che nella stabilizzazione successiva. Tenta di adattarsi ai picchi significativi, ma non riesce a replicarli con precisione, sottostimandone l'ampiezza. Di seguito sono state calcolate le metriche di accuratezza:

Mean Absolute Error (MAE)= 0,088

Mean Square Error (MSE)= 0,022

Root Mean Square Error (RMSE) = 0,148

In questo caso, il valore dell'RMSE è più basso rispetto il modello SARIMA. Se si confronta il valore dell'RMSE con il MAE si vede che questi differiscono; segno della presenza di *outliers* che potrebbero distorcere la stima.

Il modello SARIMA nel caso di pattern complessi e poco chiari non riesce a catturare l'andamento della serie, mentre le reti neurali ricorrenti riescono a studiare relazioni complesse sottese ai dati.

4.2.2 Topic 7 “Imbarco e controlli di sicurezza”

La Figura 38 mostra la rappresentazione dell’IMTP per il *topic 7 “Imbarco e controlli di sicurezza”*.

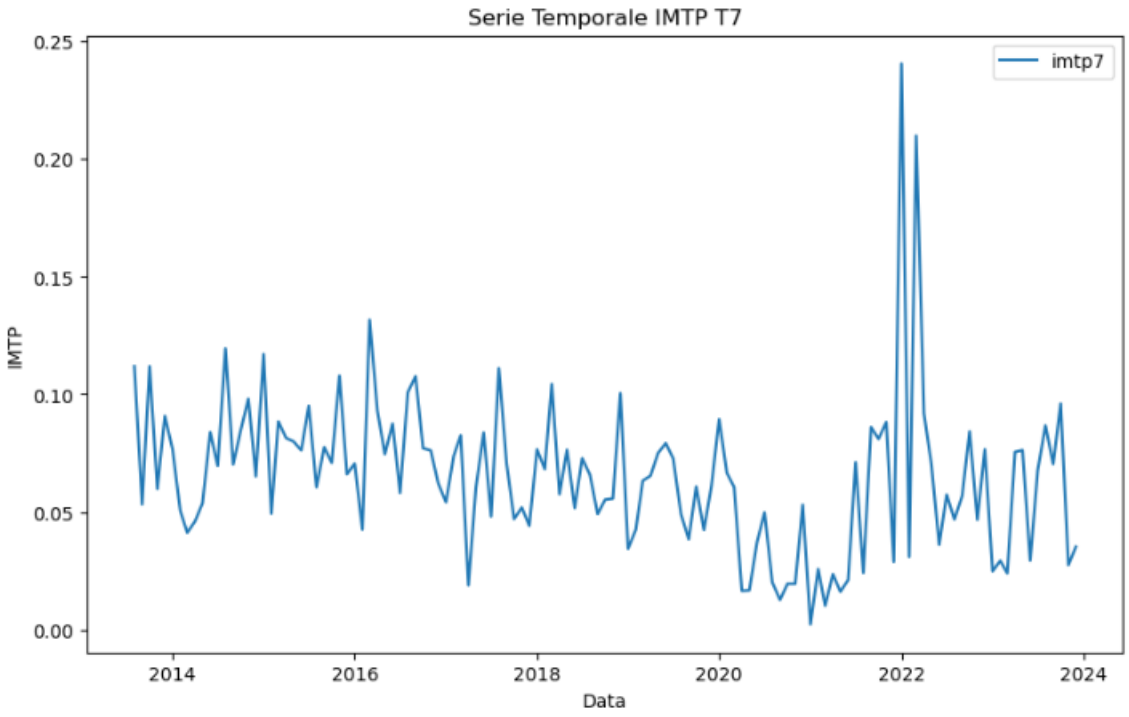

Figura 38 Rappresentazione grafica dell’IMTP del topic 7 “Imbarco e controlli di sicurezza”

Quando il volume di passeggeri aumenta, solitamente, i processi di imbarco e sicurezza diventano critici. Nel grafico si riscontrano oscillazioni stagionali e picchi, in particolare, nel 2022 si manifesta un picco, dovuto alle restrizioni e norme imposte a causa del COVID-19 [46].

La Figura 39 raffigura la decomposizione della serie temporale in: trend/tendenza, stagionalità e residuo.

Figura 39 Decomposizione dell'IMTP del topic 7 "Imbarco e controlli di sicurezza" in: 1. rappresentazione della serie originale, 2. rappresentazione del trend/tendenza della serie, 3. rappresentazione della stagionalità della serie, 4. rappresentazione del residuo della serie.

La Figura 39 mostra 4 diversi grafici:

1. la serie mostra oscillazioni regolari nel tempo, con un picco significativo nel 2022, il resto dei dati sembra seguire un pattern stabile;

2. la tendenza raggiunge un minimo all'inizio del 2021 e poi una rapida crescita nel 2022;

3. le oscillazioni stagionali sono evidenti, regolari sebbene risultino moderate rispetto i valori della serie originale;

4. i residui sono ridotti, eccetto per il 2022, dove mostrano una variabilità più elevata.

4.2.2.1 Applicazione SARIMA

La Figura 40 mostra la previsione ottenuta mediante l'applicazione del metodo SARIMA (si veda Sezione 2.5.1 e 2.5.2 ed Allegato 4).

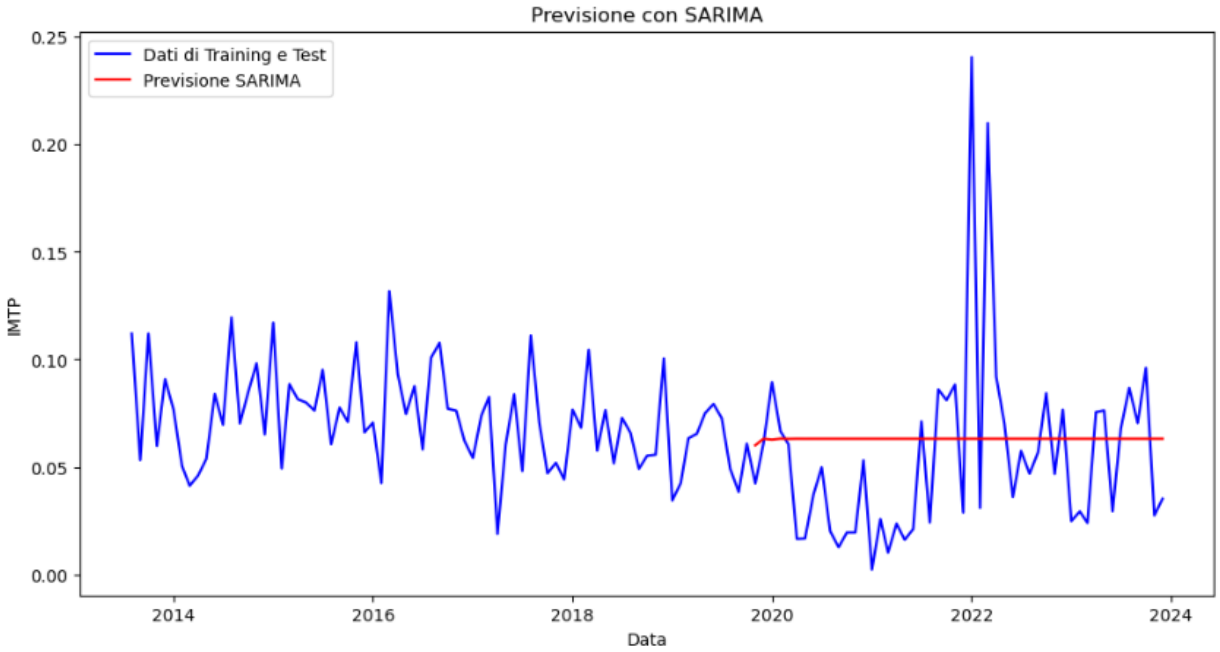

Figura 40 Applicazione di SARIMA all'IMTP del topic 7 "Imbarco e controlli di sicurezza"

La linea blu rappresenta i dati storici, mentre la linea rossa la previsione del modello, che risulta piuttosto piatta e non segue l'andamento e i picchi dei dati reali.

Questo suggerisce che il modello non è in grado di catturare né la variabilità né i picchi eccezionali osservati nel periodo post 2020. Di seguito il calcolo dell'errore di precisione del modello:

Root Mean Square Error (RMSE) = 0,044

Il modello SARIMA produce previsioni che non riflettono la variabilità dei dati originali, suggerendo un limite nella capacità di adattarsi a cambiamenti repentini o straordinari presenti nei dati. Inoltre, poiché la serie non ha una stagionalità forte prevede mediamente la tendenza sulla base dei dati del set di training.

4.2.2.2 Applicazione LSTM

La Figura 41 mostra la previsione ottenuta mediante l'applicazione del metodo LSTM (si veda Sezione 2.5.3 e 2.5.4 ed Allegato 5).

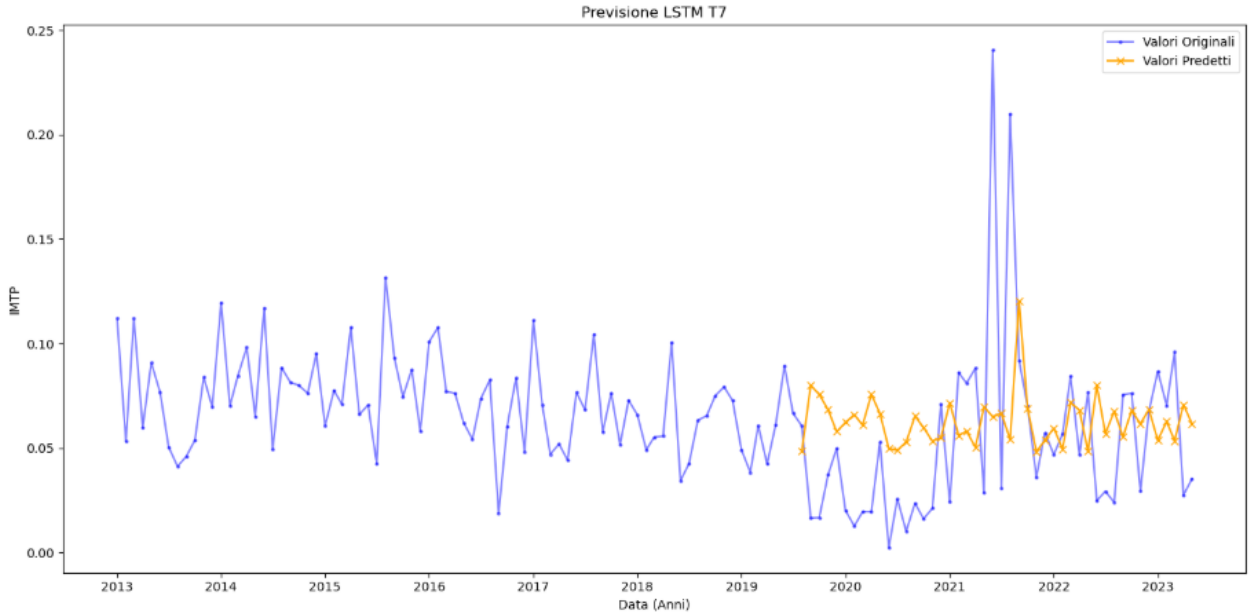

Figura 41 Applicazione di LSTM all'IMTP del topic 7 "Imbarco e controlli di sicurezza"

La linea blu rappresenta i dati reali, mentre la linea arancione rappresenta le previsioni.

Il modello cattura la tendenza generale dei dati, ma non riesce a rappresentare correttamente il picco eccezionale del 2021. Le previsioni sono più smorzate rispetto la variabilità reale, specialmente nei periodi di alta volatilità. Di seguito il calcolo dell'errore di precisione del modello:

Mean Absolute Error (MAE)= 0,022

Mean Square Error (MSE)= 0,00076

Root Mean Square Error (RMSE) = 0,028

L'errore risulta più basso rispetto SARIMA e la previsione cerca di adattarsi maggiormente al pattern dei dati. Nei periodi di alta variabilità, il modello mostra difficoltà nel catturare i comportamenti anomali.

4.2.3 Topic 3 "Esperienza di volo"

La Figura 42 mostra la rappresentazione dell'IMTP per il *topic 3 "Esperienza di volo"*.

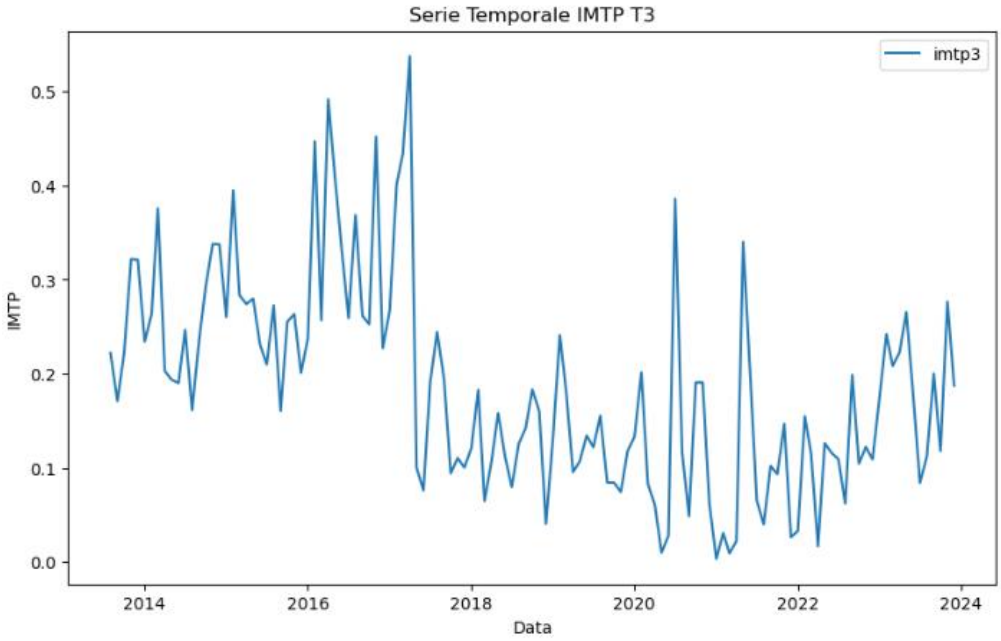

Figura 42 Rappresentazione grafica dell'IMTP del topic 3 "Esperienza di volo"

Nel 2017, si manifesta un picco nel grafico ed è collegato al fatto che Ryanair ha affrontato una significativa crisi operativa che ha compromesso l'esperienza di volo dei suoi clienti cancellando numerosi voli [47].

Durante la pandemia, le discussioni sull'esperienza di volo sono aumentate a causa di cambiamenti straordinari: misure sanitarie a bordo, riduzione dei servizi per ragioni di sicurezza [46]. Nel periodo post-pandemico, le discussioni si sono stabilizzate, ma continuano a riflettere il feedback dei passeggeri su aspetti ricorrenti.

La Figura 43 raffigura la decomposizione della serie temporale in: trend/tendenza, stagionalità e residuo.

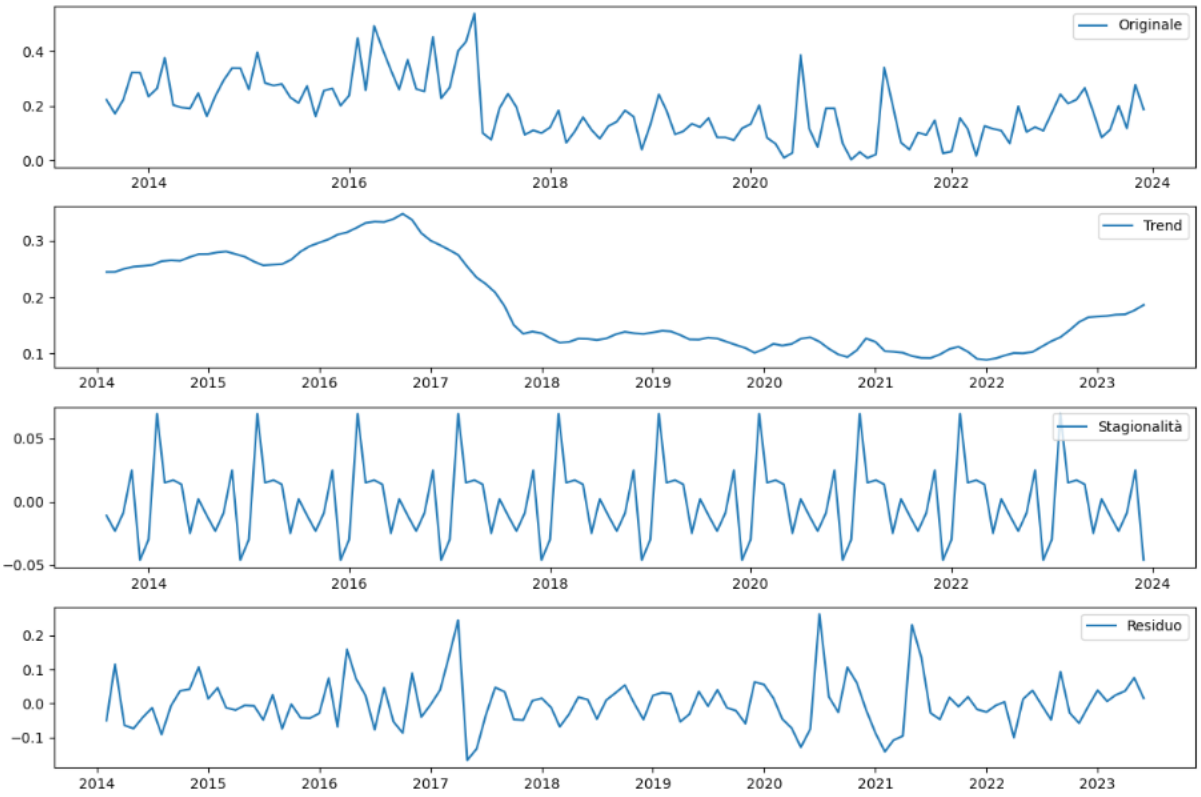

Figura 43 Decomposizione dell'IMTP del topic 3 "Esperienza di volo" in: 1. rappresentazione della serie originale, 2. rappresentazione del trend/tendenza della serie, 3. rappresentazione della stagionalità della serie, 4. rappresentazione del residuo della serie.

La Figura 43 mostra 4 diversi grafici:

1. la serie originale presenta ampie oscillazioni fino al 2016, dal 2017 i valori della serie si abbassano e dal 2020 presentano delle oscillazioni più irregolari;

2. la tendenza segue inizialmente un andamento crescente subendo una forte decrescita fino al 2018;

3. la stagionalità è ben definita e regolare durante tutto il periodo, con un'ampiezza moderata rispetto ai valori complessivi della serie;

4. i residui mostrano una certa variabilità che sembra correlata con eventi specifici non spiegati dalla tendenza e dalla stagionalità.

4.2.3.1 Applicazione SARIMA

La Figura 44 mostra la previsione ottenuta mediante l'applicazione del metodo SARIMA (si veda Sezione 2.5.1 e 2.5.2 ed Allegato 4).

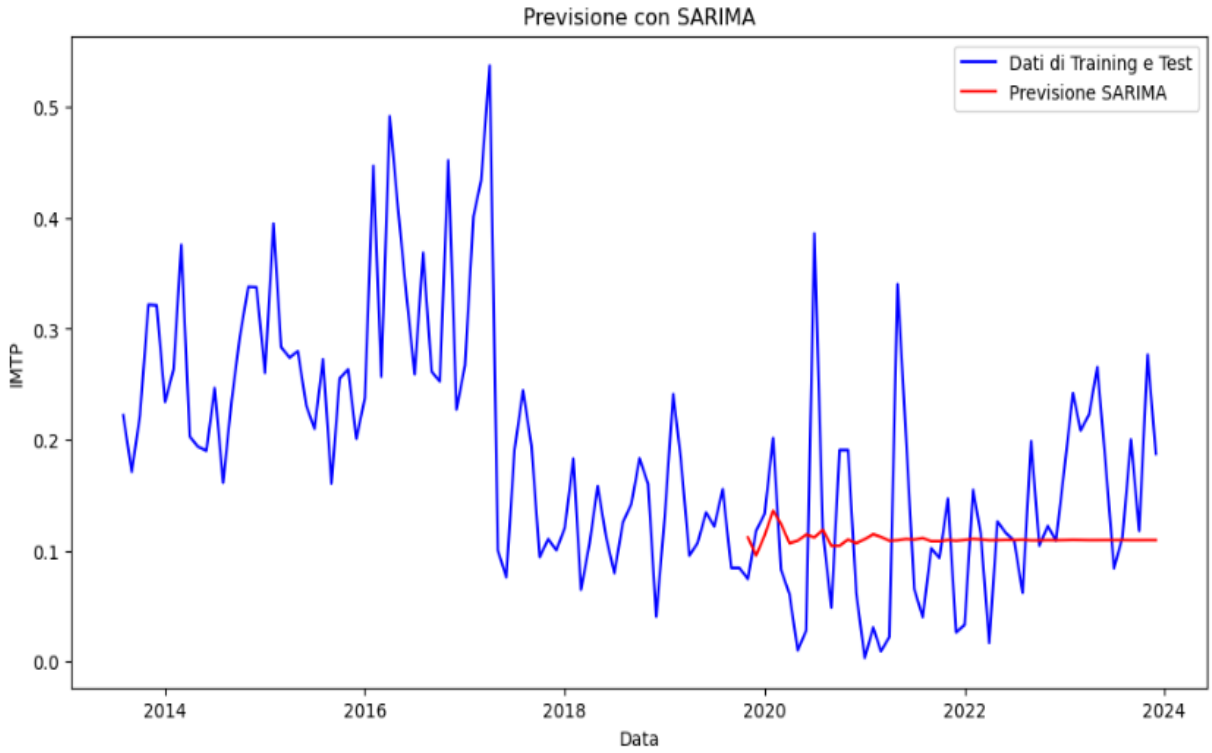

Figura 44 Applicazione di SARIMA all'IMTP del topic 3 "Esperienza di volo"

La linea blu mostra i dati storici, mentre la linea rossa rappresenta le previsioni del modello, che appaiono piatte e non catturano la variabilità della serie. Di seguito il calcolo dell'errore di accuratezza del modello:

Root Mean Square Error (RMSE) = 0,87

Il valore dell'RMSE in questo caso risulta essere molto elevato segno di un modello che non è in grado di replicare l'andamento della serie. La previsione sottolinea l'importanza di integrare il modello SARIMA con variabili esogene e approcci alternativi per catturare eventi straordinari e picchi di discussione.

4.2.3.2 Applicazione LSTM

La Figura 45 mostra la previsione ottenuta mediante l'applicazione del metodo LSTM (si veda Sezione 2.5.3 e 2.5.4 ed Allegato 5).

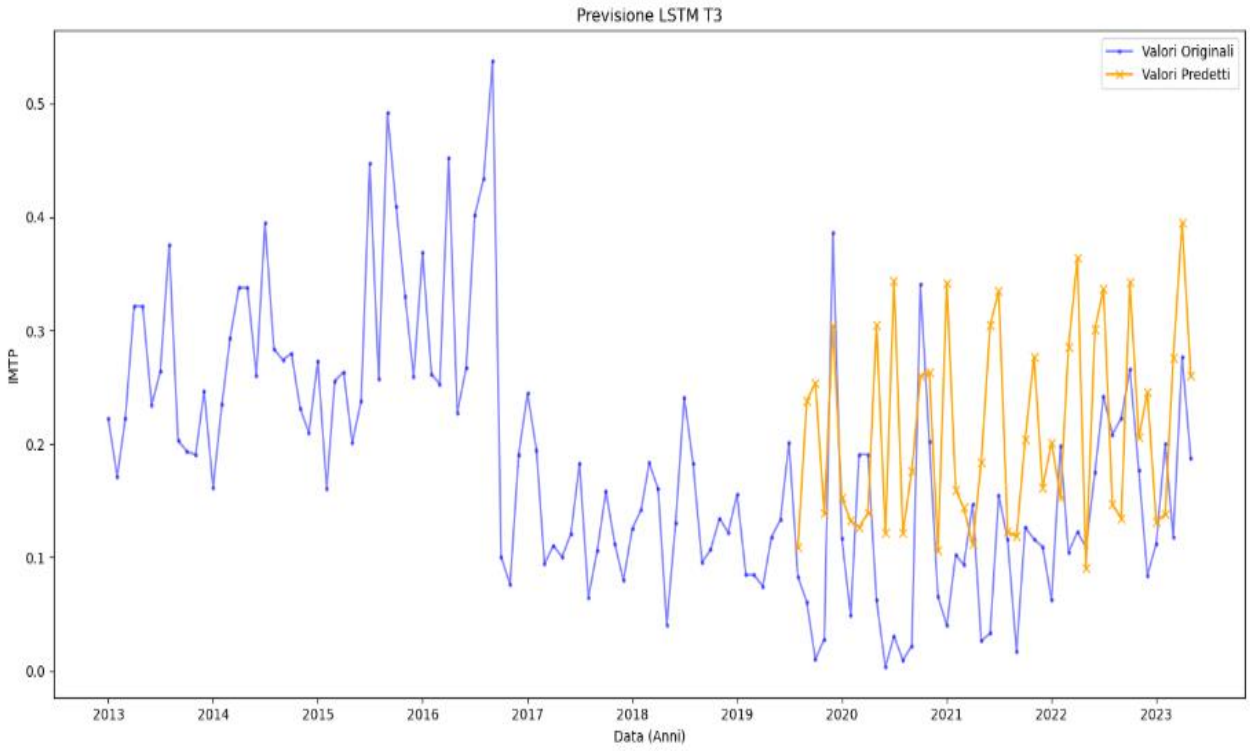

Figura 45 Applicazione di LSTM all'IMTP del topic 3 "Esperienza di volo"

La linea blu rappresenta i dati reali, mentre la linea arancione rappresenta le previsioni.

Il modello LSTM cattura bene la tendenza generale post 2020, seguendo i picchi e le oscillazioni. Tuttavia, alcune previsioni mostrano una sovrastima o sottostima rispetto i dati reali. Il modello LSTM cattura la tendenza generale e parte della variabilità dei dati, ma richiede ottimizzazioni per gestire meglio i picchi e ridurre il sovra adattamento. Di seguito il calcolo dell'errore di precisione del modello:

Mean Absolute Error (MAE)= 0,077

Mean Square Error (MSE)= 0,0106

Root Mean Square Error (RMSE) = 0,103

Anche in questo caso l'errore di previsione risulta maggiore rispetto il modello SARIMA, in quanto tende a replicare i picchi ma con una certa discrepanza rispetto ai valori originali della serie.

5 Capitolo 5: Discussione dei risultati e applicazioni future

5.1 Discussione dei risultati ottenuti

Lo studio in esame mira a trovare una metodologia per condurre analisi predittive su come varia l'andamento dell'IMTP dei topic individuati dal modello STM. Entrambi i metodi proposti hanno dei limiti nell'individuare la corretta previsione dei dati:

i. SARIMA: prova a catturare l'andamento della serie, ma nei casi in cui quest'ultima non presenta una evidente e chiara stagionalità o in caso di pattern complessi e poco chiari, non prevede correttamente l'andamento approssimandolo ad una linea piatta. In presenza di eventi anomali che disturbano l'andamento della serie, non prevede in modo accurato variazioni repentine.

ii. LSTM: riesce a catturare pattern più complessi e ad ottenere previsioni che seguono maggiormente i dati della serie originale. In caso di eventi anomali, riesce ad adattarsi, seppur con una certa incertezza, all'andamento della serie originale. Le reti neurali ricorrenti, però, necessitano di dataset di dimensioni elevate per essere accuratamente addestrate ed i parametri del modello devono essere impostati in modo opportuno.

Le Tabella 25 e 26 mostrano un confronto dei valori di RMSE ottenuti nelle applicazioni dei modelli nei due casi studio esaminati.

Caso studio Disneyland	RMSE SARIMA	RMSE LSTM
T8	0,015	0,007
T13	0,0024	0,0033
T5	0,0036	0,0047

Tabella 25 Confronto valori RMSE caso studio Disneyland

Caso studio Ryanair	RMSE SARIMA	RMSE LSTM
T9	0,23	0,148
T7	0,044	0,028
T3	0,87	0,103

Tabella 26 Confronto valori RMSE caso studio Ryanair

In presenza di stagionalità marcate e di serie stabile nel tempo, SARIMA risulta essere un ottimo candidato per il modello, un esempio nell'analisi del topic 13 nel caso Disneyland (si veda Sezione 3.2.2).

Nel caso studio Ryanair, invece, si è visto come il COVID-19 ha avuto un impatto significativo sul valore dell'IMTP e quindi eventi imprevedibili o poco prevedibili non riescono ad essere catturati con precisione dai modelli (si veda Sezione 4.2).

Analizzando e confrontando il valore dell'RMSE nei due casi, sebbene il modello LSTM si avvicina maggiormente al pattern della serie, vi sono casi in cui il valore dell'RMSE riscontrato risulta maggiore rispetto al modello SARIMA.

Il modello LSTM, tuttavia, risulta un ottimo candidato per la valutazione di tali serie storiche. I requisiti da rispettare per applicare il modello sono di impostare in modo opportuno i parametri: la finestra temporale, il numero di strati della rete ed il numero di epoche (si veda Sezione 2.5.4).

La procedura è iterativa e richiede di esplorare diverse soluzioni in modo da ottenere i parametri che meglio si adattano al modello in esame. Per monitorare la discussione dei topic nel tempo si può implementare la rete neurale ricorrente LSTM.

5.2 Proposte applicative per il management della qualità

Condotte le analisi predittive sui topic singolarmente (si veda Capitolo 3 e 4) e scelto il modello LSTM come miglior candidato per le analisi, sono state elaborate due possibili soluzioni che possono essere applicate in azienda per monitorare l'IMTP dei topic nel tempo.

Il reparto qualità a seconda delle esigenze potrà valutare l'applicazione del modello LSTM con i seguenti approcci:

1. applicazione LSTM all'IMTP di ciascun topic separatamente (analisi disgiunta), oppure
2. applicazione LSTM all'IMTP di tutti i topic contemporaneamente (analisi congiunta).

La prima proposta considera l'applicazione di LSTM ai topic singolarmente, come visto nei Capitoli 3 e 4. La seconda proposta, invece, prende in input la matrice che contiene i valori di IMTP per ciascun topic ed analizza e conduce analisi predittive per ognuno di essi (si veda Tabella 19 e 24).

A seconda delle analisi da condurre sui dati il management della qualità aziendale potrà valutare l'applicazione di un'analisi congiunta o disgiunta dell'IMTP. L'applicazione della prima o della seconda proposta ha vantaggi e svantaggi che si approfondiranno nelle Sezioni 5.2.1 e 5.2.2.

5.2.1 Soluzione 1: applicazione LSTM all'IMTP di ciascun topic separatamente

La prima soluzione prevede di studiare singolarmente l'IMTP dei topic estratti dal modello STM. La procedura applica il modello LSTM (si veda Allegato 5), per condurre analisi predittive sui dati della serie.

La Figura 46 mostra il diagramma di flusso del modello.

Figura 46 Diagramma di flusso del modello LSTM analisi disgiunta

Le fasi previste sono state approfondite nei capitoli precedenti (si veda Sezione 2.5.4): creazione delle sequenze, suddivisione del dataset 80-20%, addestramento, creazione del modello e previsioni.

Il rispetto del vincolo viene spiegato nel paragrafo 5.2.1.1.

5.2.1.1 Rispetto del vincolo

Le previsioni ottenute per il valore dell'IMTP di ciascun topic sono correlate tra loro, quindi, dopo aver applicato i modelli è necessario fare alcune considerazioni.

Per ogni periodo di campionamento *t-esimo*, la somma di *IMTPd,t* relativa a tutti i topic identificati è uguale a 1 (si veda Sezione 1.3.6.2):

$$\sum_{d=1}^{D} IMTP_{d,t} = 1 \qquad \forall\, t \in (1, \dots, T) \tag{5.1}$$

Dove:

- *d* è il numero di topic;
- *t* è il periodo di campionamento.

Se si analizza l'IMTP del topic 1, e mostra, ad esempio, un andamento crescente, necessariamente un altro topic estratto dal modello, mostrerà un andamento decrescente, questo affinché il vincolo sui dati venga rispettato.

Si possono presentare due differenti scenari:

i. se lo scarto tra le previsioni e i valori originali è trascurabile, il vincolo sulle previsioni è rispettato;
ii. se le previsioni ottenute su ciascun periodo si discostano da quelle reali, a causa ad esempio della presenza di un evento anomalo, non vi è il rispetto del vincolo, quindi si verificherà:

$$\sum_{d=1}^{D} \widehat{IMTP}_{d,t} \neq 1 \qquad \forall\, t \in (1, \dots, T) \tag{5.2}$$

dove $\widehat{IMTP}_{d,t}$ sono i valori predetti per ciascun *d-esimo* topic in uno specifico *t-esimo* istante temporale. Valutato sui periodi di campionamento nel set di test dei dati.

Nel secondo caso, si può applicare la seguente divisione, in modo che il vincolo valutato sulle previsioni venga rispettato:

$$\widehat{IMTP}_{d,t} = \frac{\widehat{IMTP}_{d,t}}{\sum_{d=1}^{D} \widehat{IMTP}_{d,t}} \quad \forall\, d = 1, \dots, D \tag{5.3}$$

Ottenuti i valori è possibile calcolare l'errore di previsione di ciascun topic e notare come questi hanno subito delle variazioni rispetto i valori calcolati dal modello.

Tale soluzione risulta essere lunga e laboriosa, ma analizza nel dettaglio ciascun topic e fornisce previsioni accurate ottimizzando i parametri per ognuno di essi.

5.2.2 Soluzione 2: applicazione LSTM all'IMTP di tutti i topic contemporaneamente

La seconda soluzione prende come input la matrice che ha come colonne la data ed il valore dell'IMTP per ciascun topic (si veda Tabella 19).

La Figura 47 mostra il diagramma di flusso del modello.

Figura 47 Diagramma di flusso soluzione LSTM analisi congiunta

Le fasi presentate nel diagramma sono le medesime analizzate per l'applicazione del modello LSTM ai due casi studio precedentemente analizzati.

In questo caso, si vedrà che il rispetto del vincolo è insito nel modello in quanto, la funzione "Softmax" applicata all'ultimo strato della rete, garantisce che la condizione sia verificata [52].

5.2.2.1 Soluzione 2 applicata al caso studio Disneyland su Python

Il modello LSTM, in questo caso è stato applicato al valore dell'IMTP di tutti i topic estratti dal modello STM (si veda Allegato 6).

Sono state istallate le seguenti librerie oltre quelle già istallate in precedenza:

- sklearn preprocessing in particolare "MinMaxScalar", che permette la normalizzazione dei dati [48];
- sklearn model selection in particolare la funzione "train_test_split" che permette di suddividere opportunatamente il dataset [49];
- tensorflow keras model in particolare "sequential" per creare le sequenze di input del modello [50];
- tensorflow keras layer in particolare gli strati LSTM e Dense e la funzione per l'attivazione dello strato denso "softmax" [51-52].

5.2.2.1.1 Creazione delle sequenze e suddivisione del dataset in 80-20%

Si è stabilita una finestra temporale pari a 12 ed i dati sono stati suddivisi opportunatamente in training e test in 80-20% (si veda Sezione 2.5.4.1 e 2.5.4.2).

5.2.2.1.2 Creazione del modello

La creazione del modello LSTM ha richiesto la definizione degli strati e per ognuno di essi è stato stabilito il numero di connessioni necessarie per rappresentare la serie in esame [40-42]. In particolare, due strati LSTM e uno strato Denso [42]. L'ultimo strato, detto Denso, ovvero totalmente connesso ha 13 neuroni che corrispondono al numero di output del modello e lo

strato è attivato dalla funzione “Softmax” la quale garantisce che il vincolo della correlazione tra i valori dell’IMTP dei topic venga rispettato nella previsione [52].

Il modello una volta impostato è stato addestrato considerando come perdita il “mean square error” (MSE) [18]. Il numero di epoche nella fase di addestramento consiste nel numero di volte in cui si ha il passaggio sia forword che backword nella rete ed il conseguente ricalcolo dei pesi delle connessioni dei neuroni e dell’errore di previsione [18-19]. Il numero di epoche è stabilito in modo da ottenere il minimo del MSE, cercando di evitare di impostare un valore troppo elevato causando over-fitting [18-19].

5.2.2.1.3 Previsioni e calcolo RMSE

Dopo la fase di addestramento e la valutazione delle previsioni è stato calcolato il valore del RMSE come mostrato nella Tabella 27.

Topic	RMSE
T1	0,0044
T2	0,0047
T3	0,0109
T4	0,0074
T5	0,0037
T6	0,0154
T7	0,0083
T8	0,0155
T9	0,0084
T10	0,007
T11	0,006
T12	0,0065
T13	0,0042

Tabella 27 Calcolo RMSE previsione con LSTM

Di seguito il calcolo della previsione dei topic applicato al caso studio Disneyland. I risultati ottenuti per i 13 topic vengono salvati in un dataframe che presenta la data nella prima colonna e le restanti il valore dell’IMTP per ciascun topic.

La Figura 48 mostra la previsione per il topic 1.

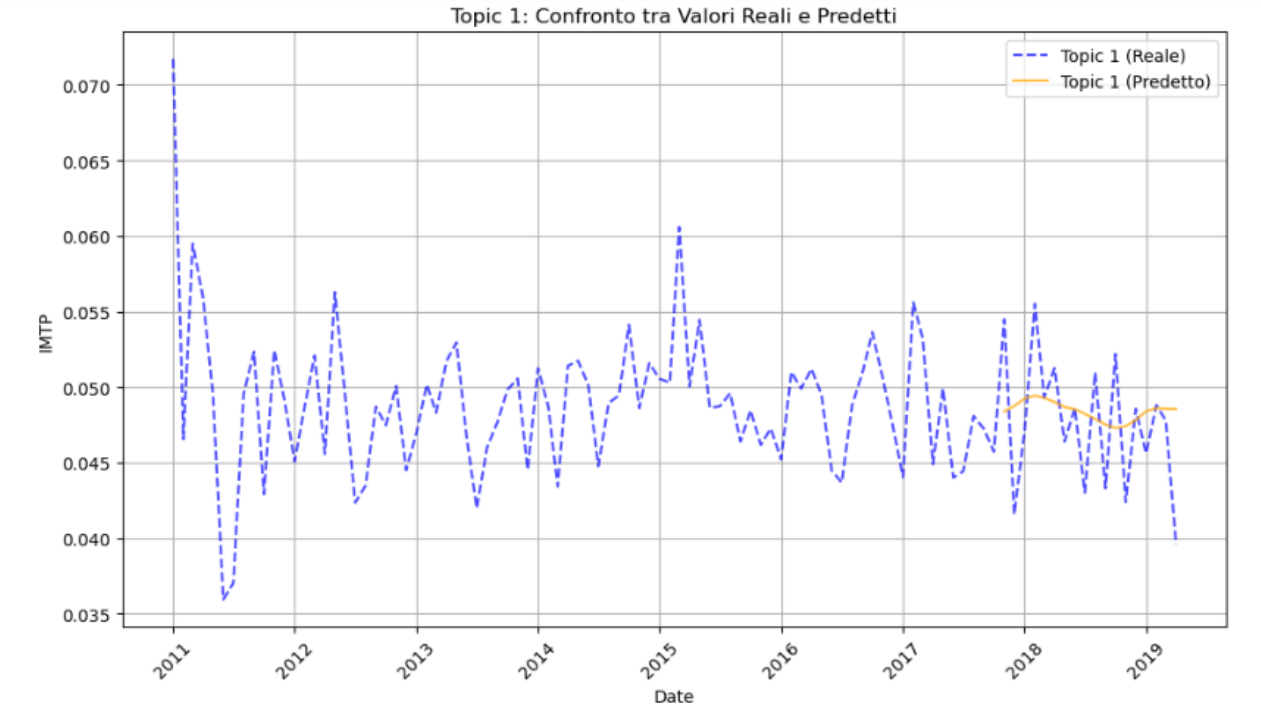

Figura 48 Previsione LSTM topic 1

La Figura 49 mostra la previsione per il topic 2.

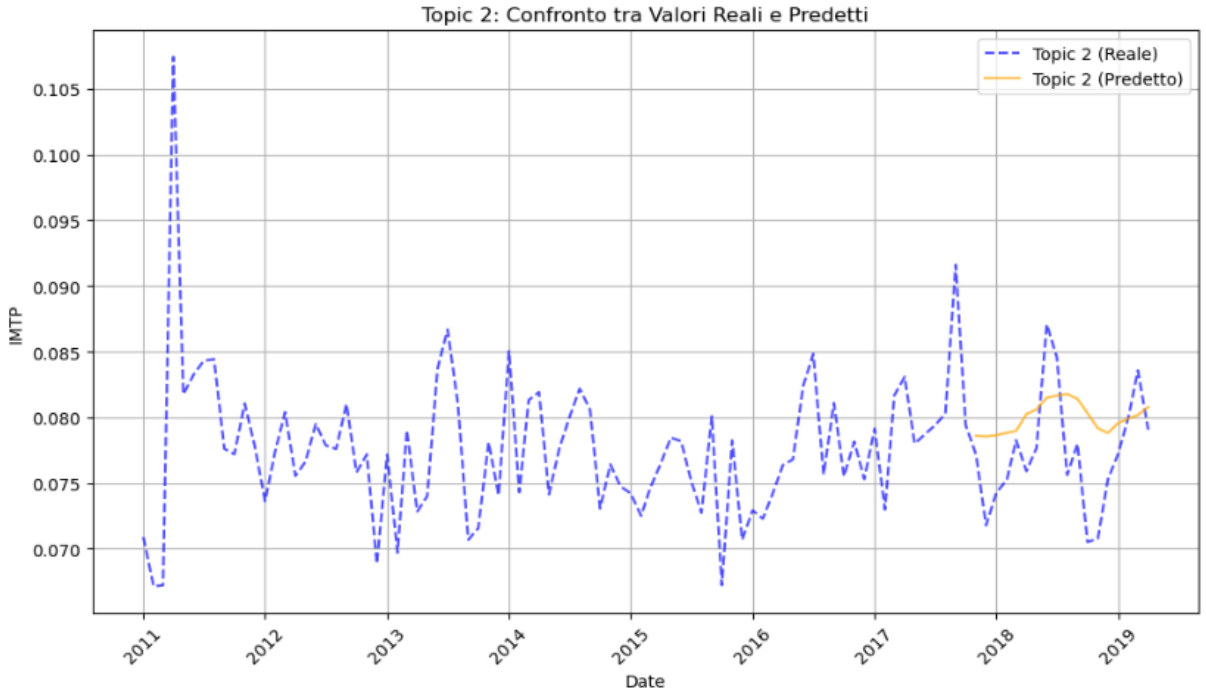

Figura 49 Previsione LSTM topic 2

La Figura 50 mostra la previsione per il topic 3.

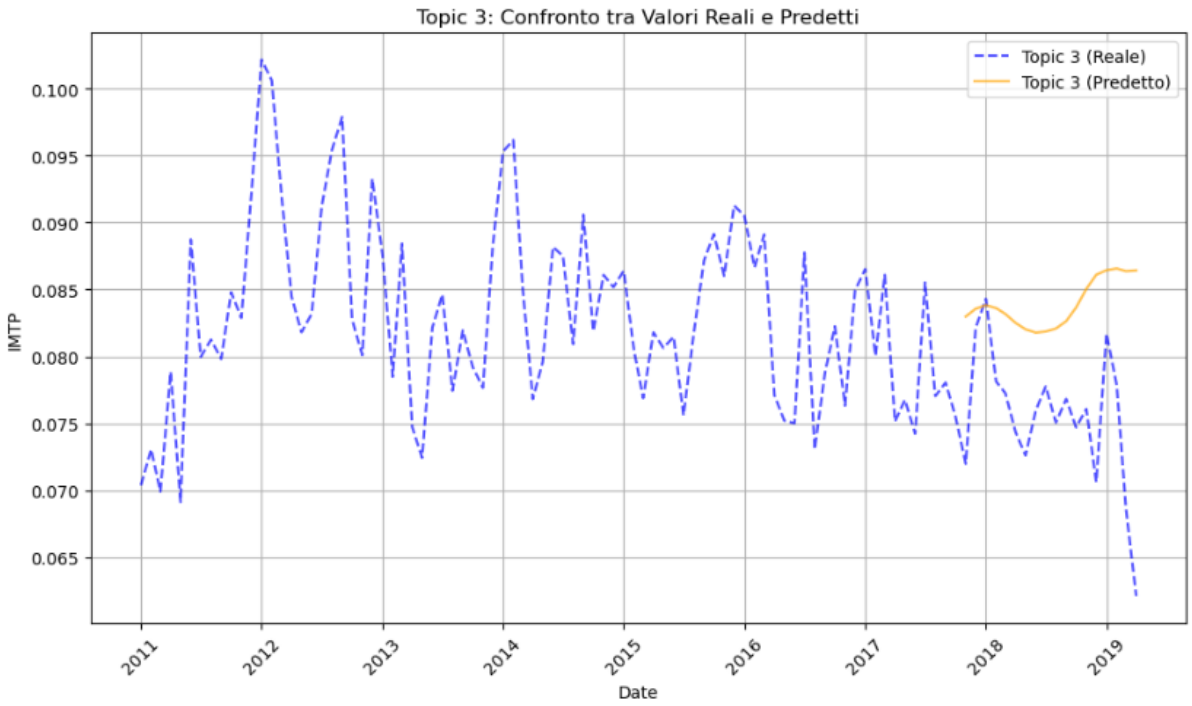

Figura 50 Previsione LSTM topic 3

La Figura 51 mostra la previsione per il topic 4.

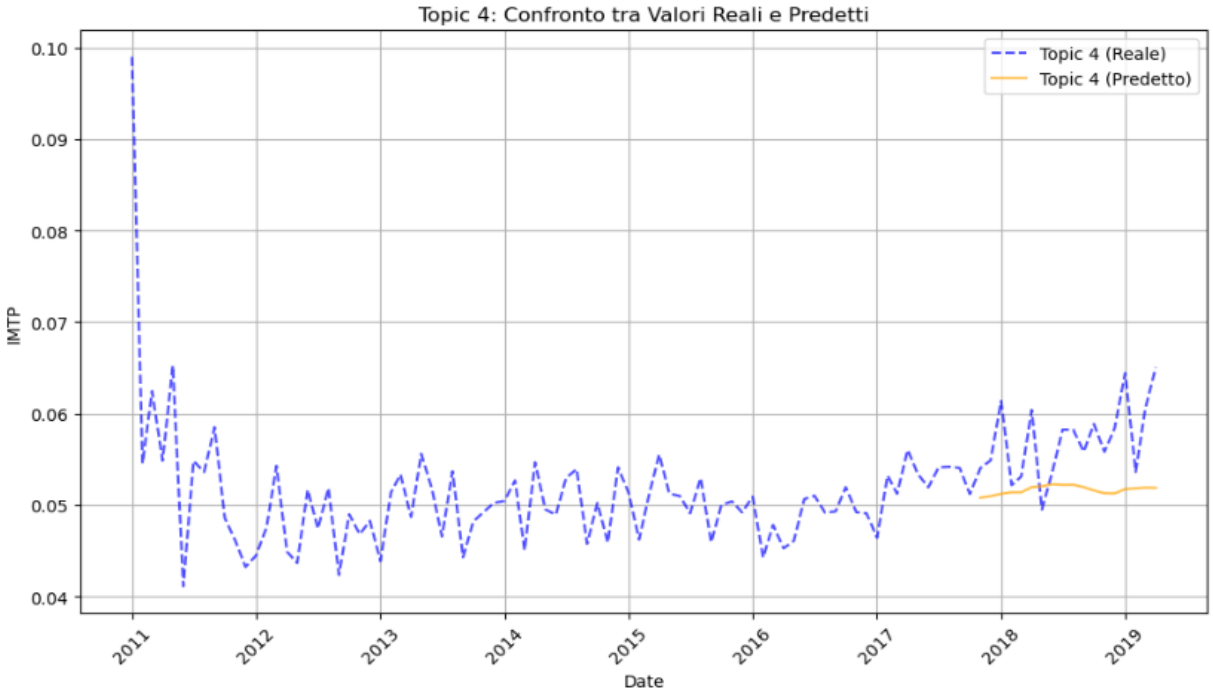

Figura 51 Previsione LSTM topic 4

La Figura 52 mostra la previsione per il topic 5.

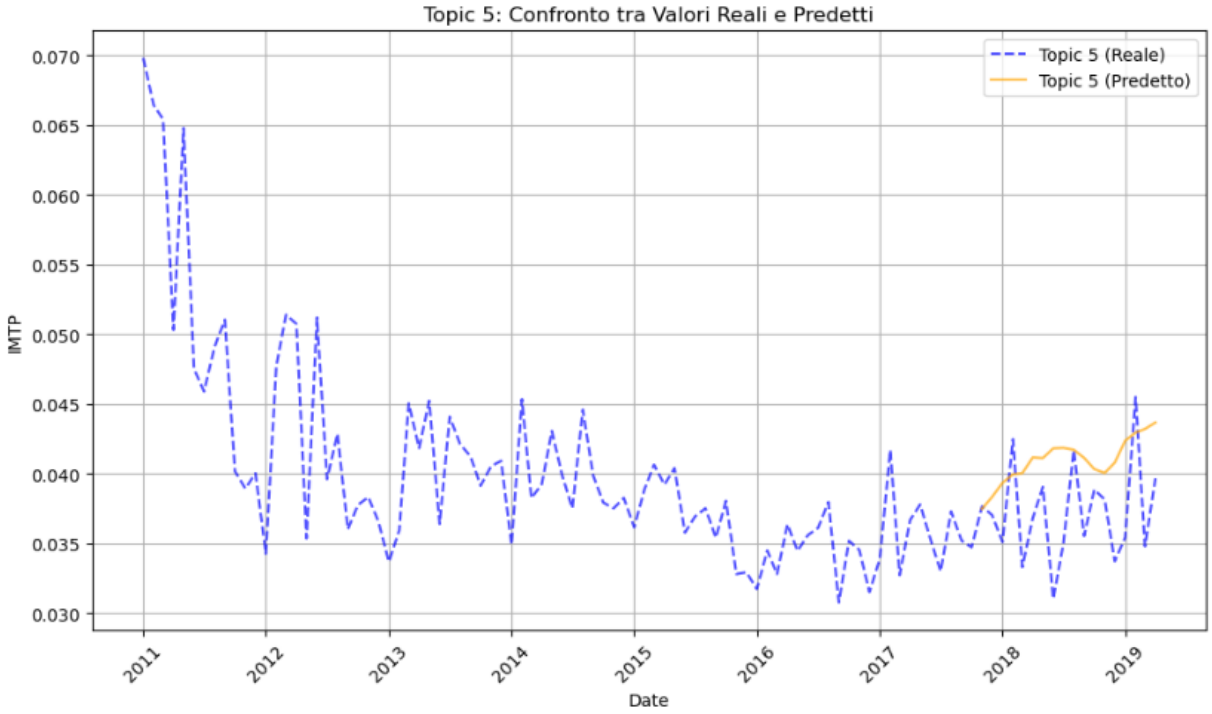

Figura 52 Previsione LSTM topic 5

La Figura 53 mostra la previsione per il topic 6.

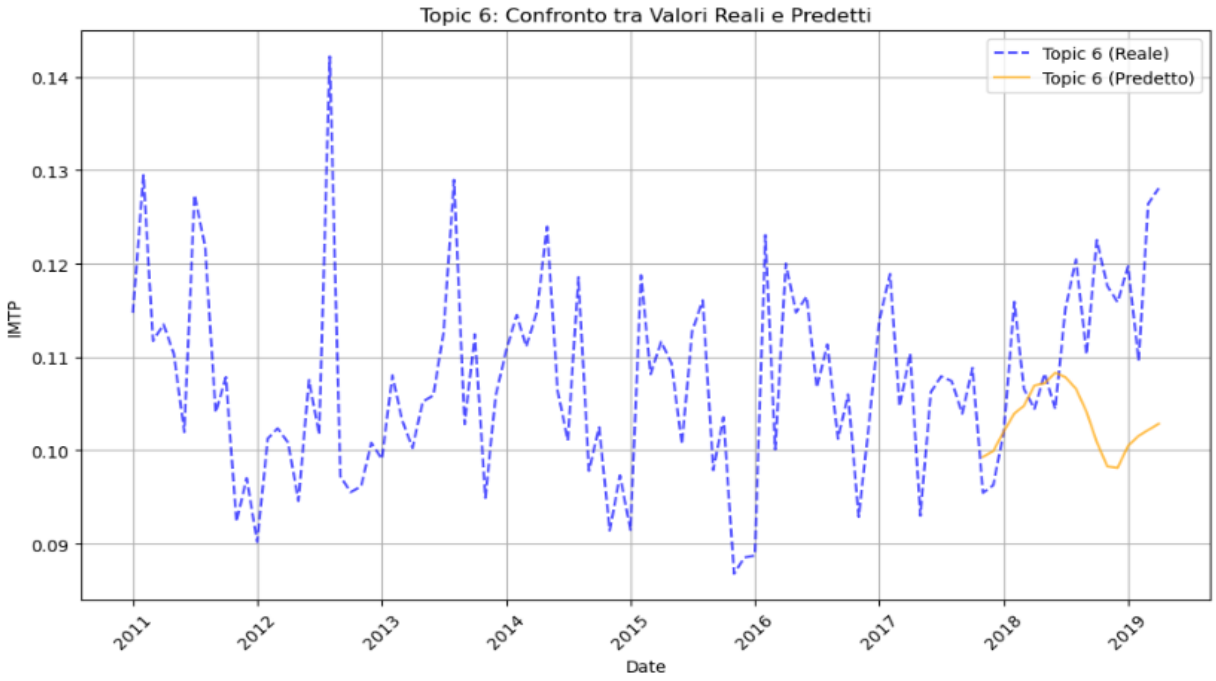

Figura 53 Previsione LSTM topic 6

La Figura 54 mostra la previsione per il topic 7.

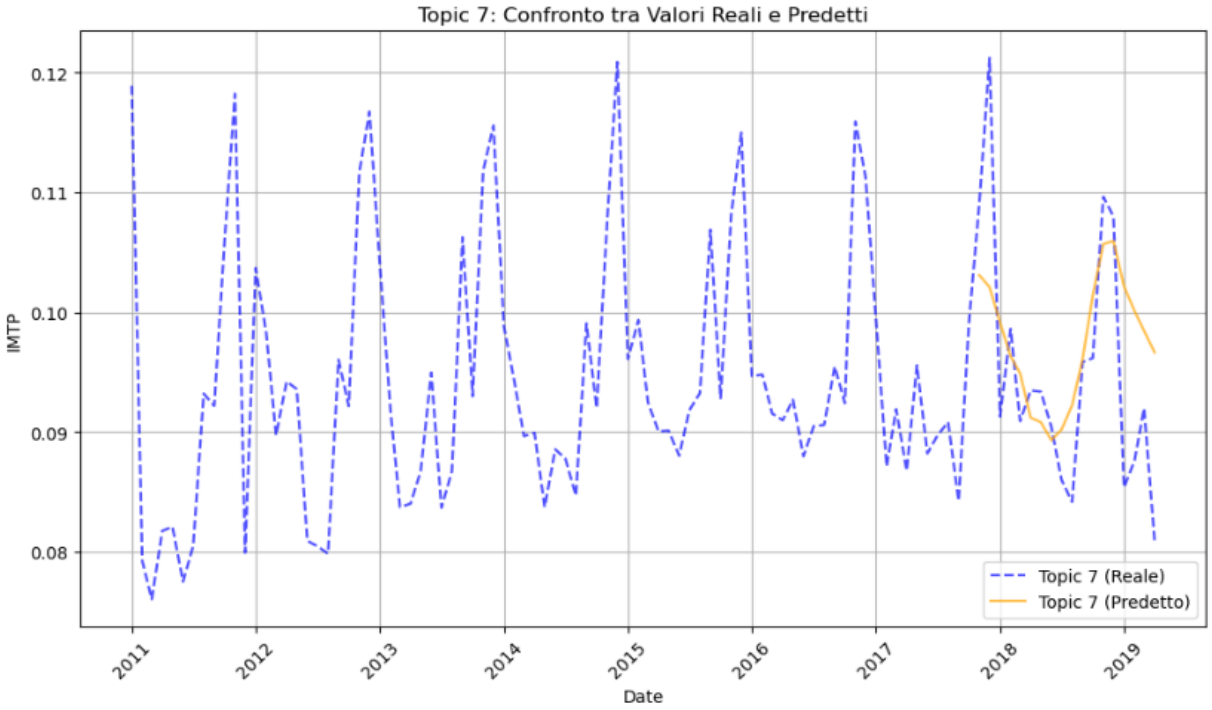

Figura 54 Previsione LSTM topic 7

La Figura 55 mostra la previsione per il topic 8.

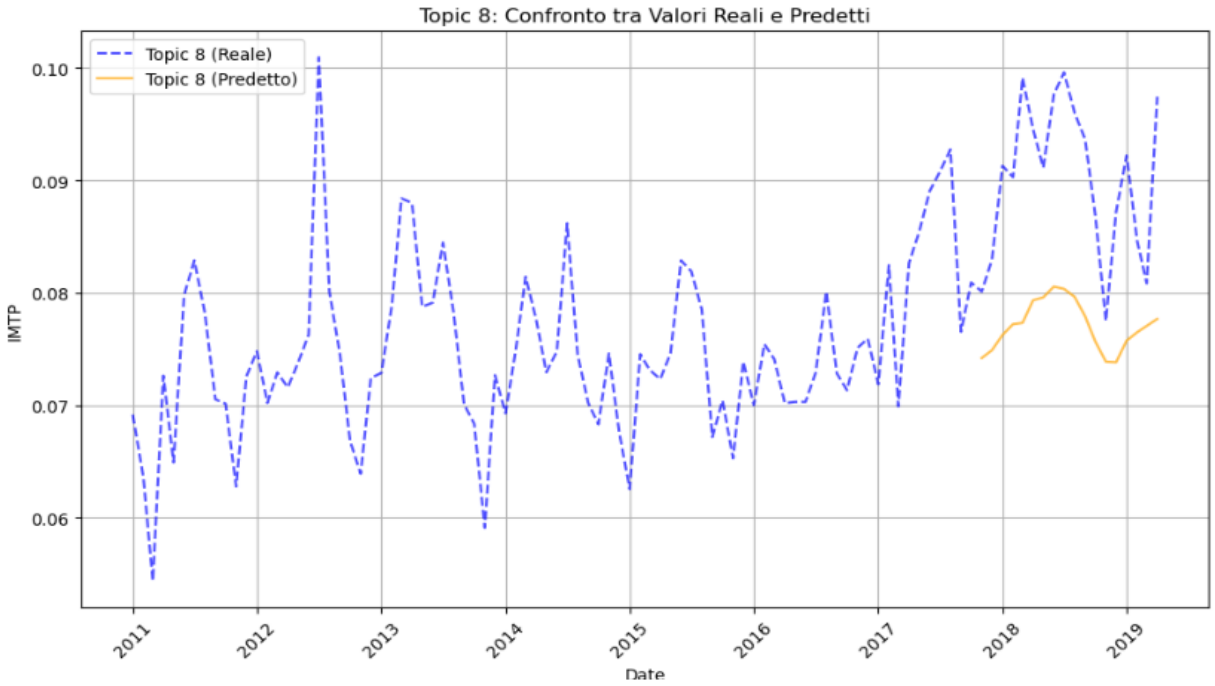

Figura 55 Previsione LSTM topic 8

La Figura 56 mostra la previsione per il topic 9.

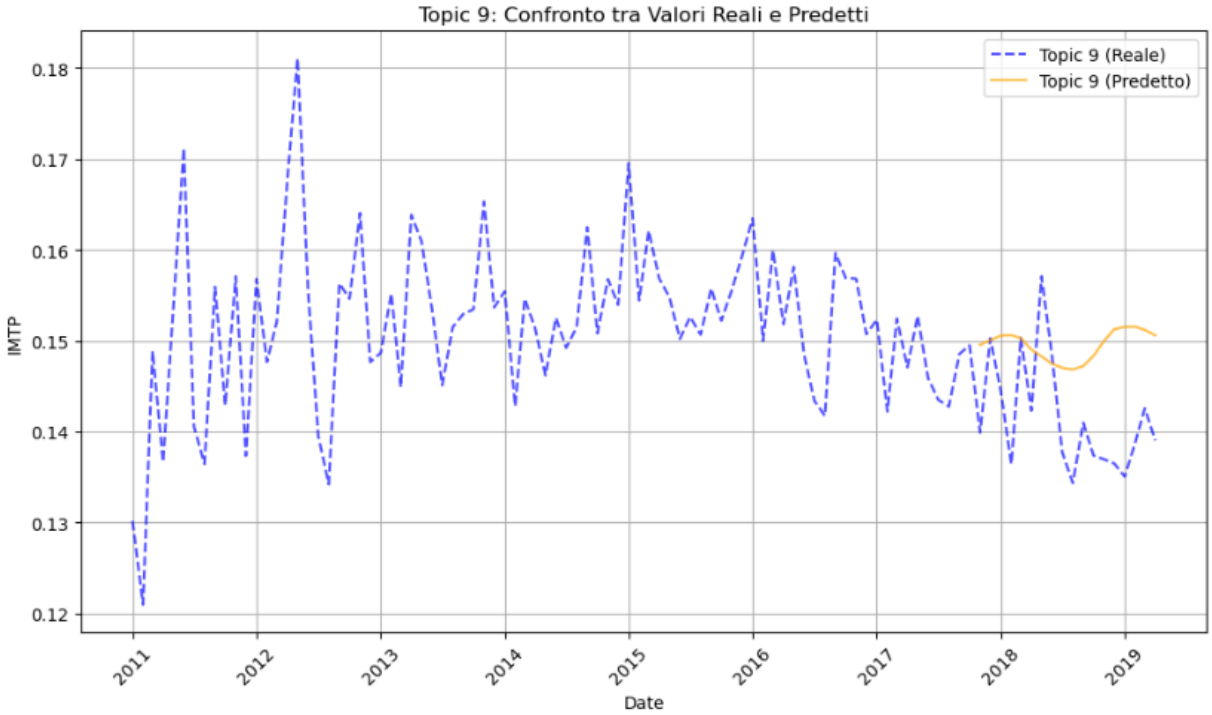

Figura 56 Previsione LSTM topic 9

La Figura 57 mostra la previsione per il topic 10.

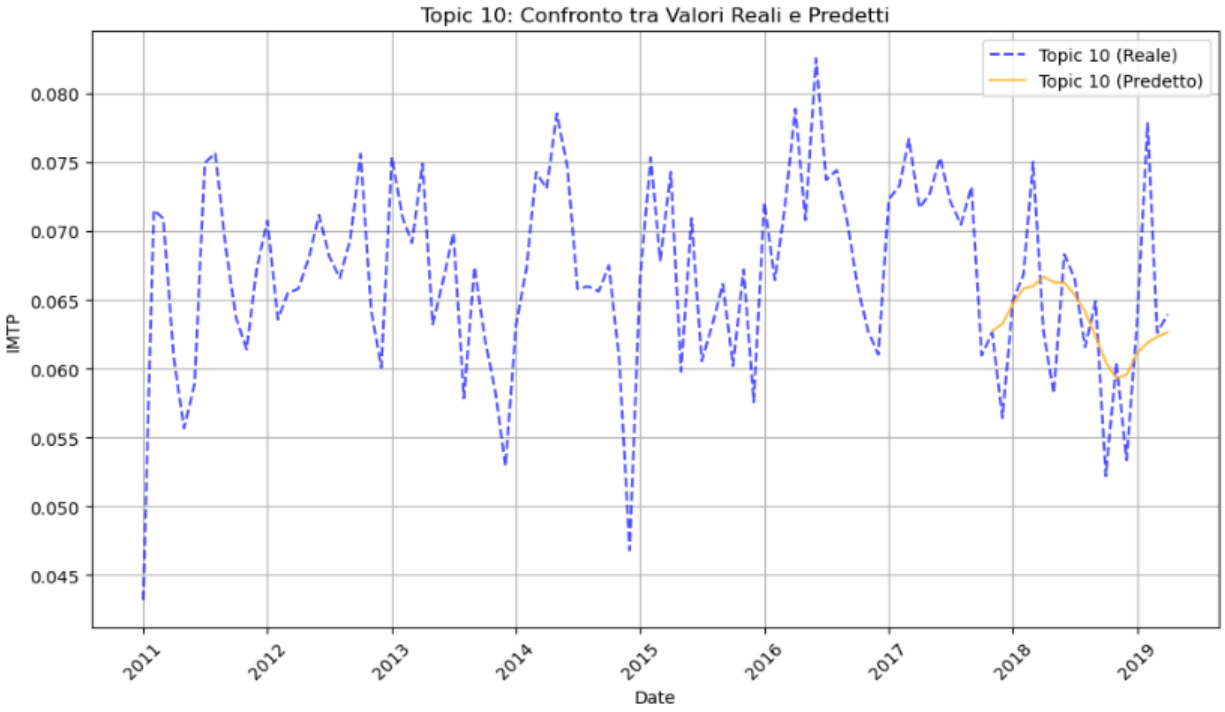

Figura 57 Previsione LSTM topic 10

La Figura 58 mostra la previsione per il topic 11.

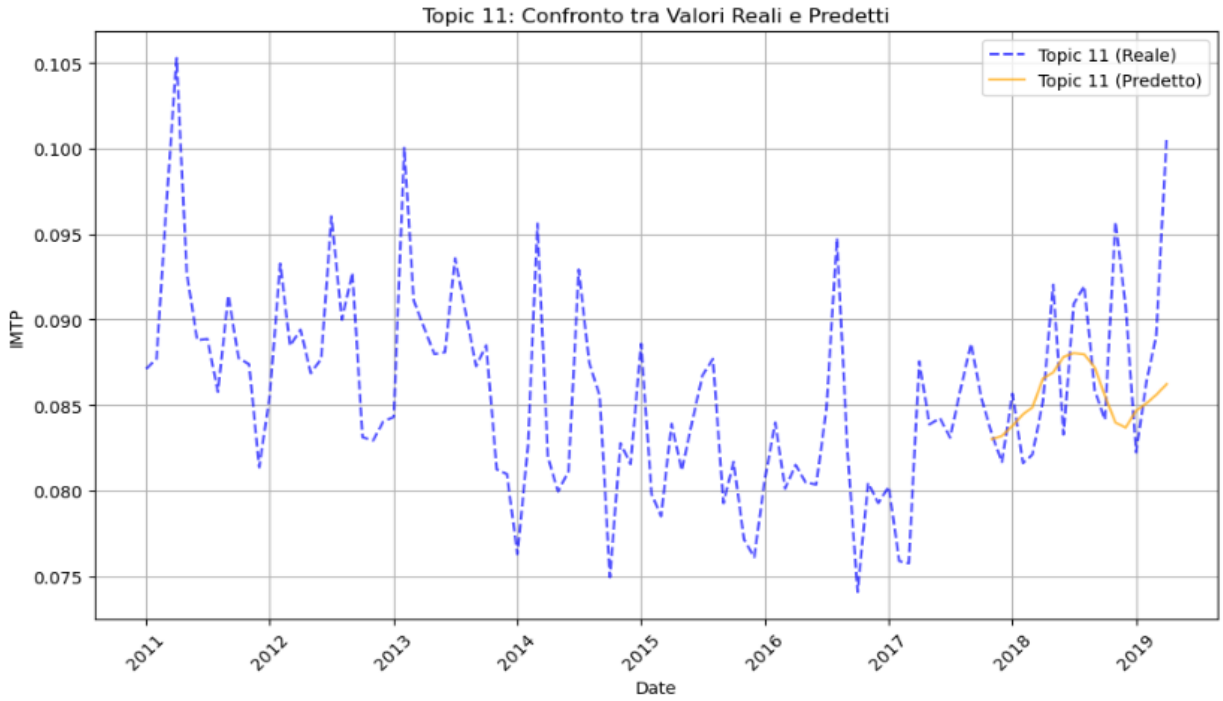

Figura 58 Previsione LSTM topic 11

La Figura 59 mostra la previsione per il topic 12.

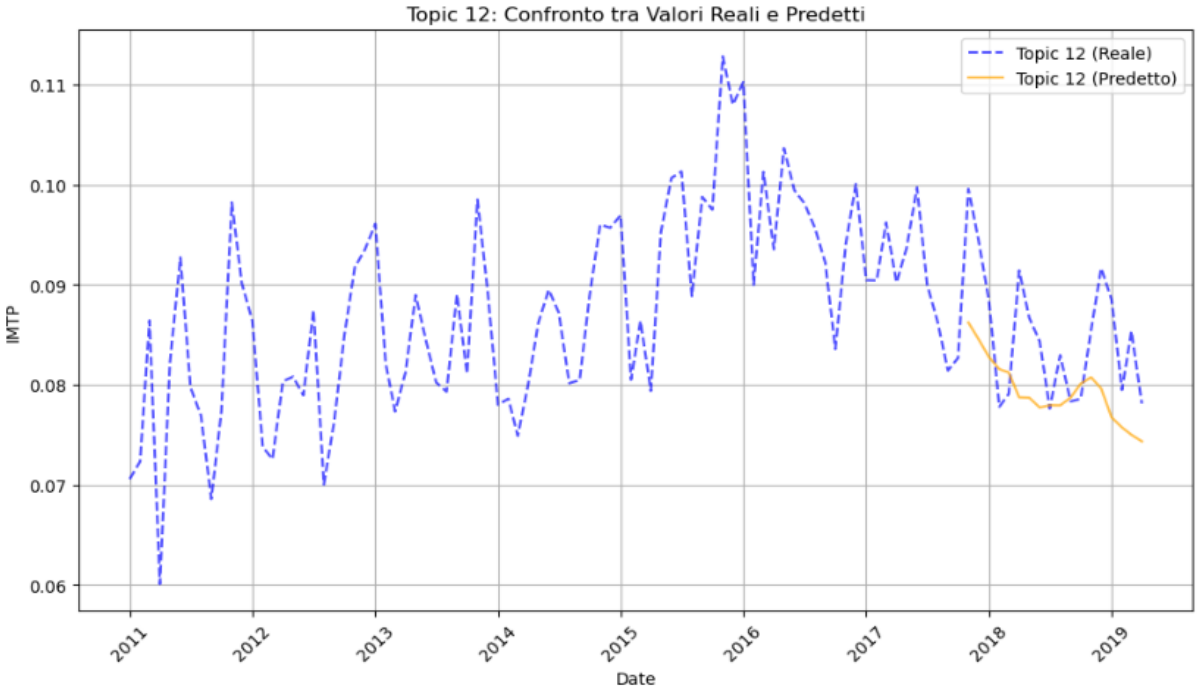

Figura 59 Previsione LSTM topic 12

La Figura 60 mostra la previsione per il topic 13.

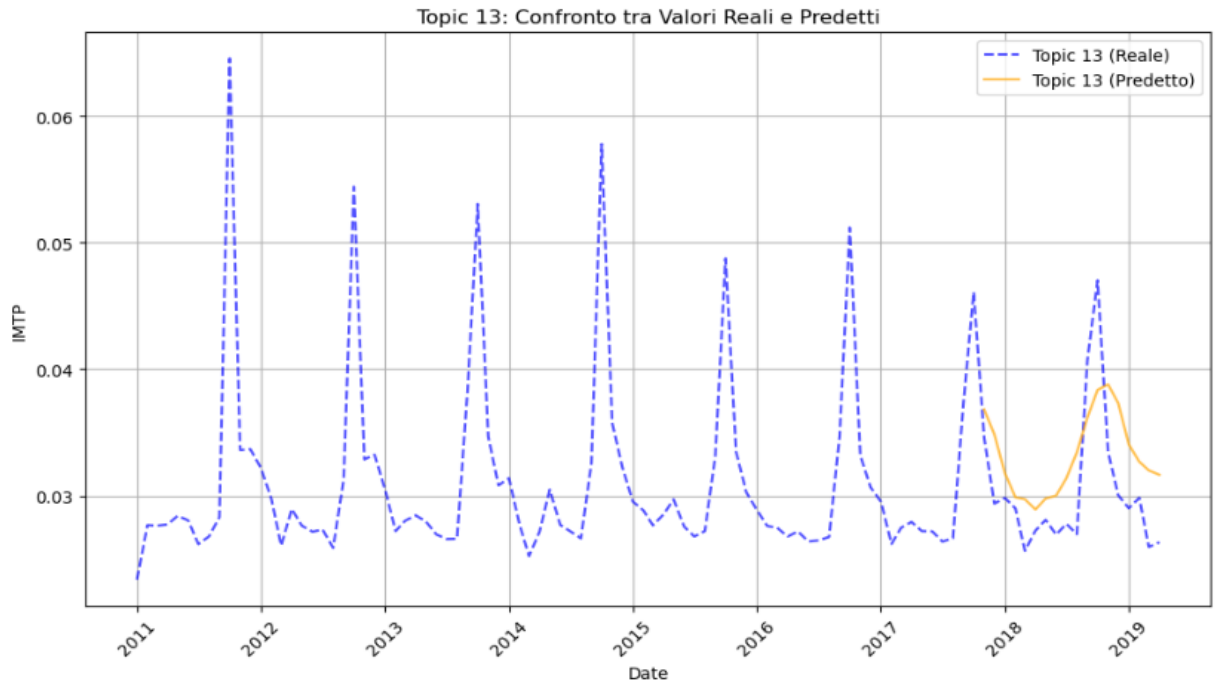

Figura 60 Previsione LSTM topic 13

5.2.3 Confronto delle due soluzioni

Le due soluzioni proposte possono essere adottate in base all'obiettivo dell'analisi e al livello di accuratezza desiderato:

- **Soluzione 1**: il modello LSTM si applica all'IMTP di ciascun topic estratto tramite l'algoritmo STM. Questo approccio permette di ottimizzare i parametri del modello per ogni topic specifico, garantendo previsioni accurate. La fase di addestramento e la verifica del rispetto del vincolo, per questo richiedono tempo in quanto vengono valutati i topic singolarmente.
- **Soluzione 2**: l'LSTM prende in input la Tabella 19, che contiene gli IMTP dei topic estratti, permettendo così un'analisi complessiva. Tuttavia, poiché i parametri dell'algoritmo vengono impostati un'unica volta per tutti i topic, diventa più complesso ottenere previsioni precise per ciascuno di essi.

Il management della qualità dovrà quindi scegliere la soluzione più adatta in base alle proprie esigenze:

I. se l'obiettivo è massimizzare la precisione delle previsioni, sarà preferibile la prima soluzione;

II. se invece si necessita un’analisi rapida e complessiva, si potrà optare per la seconda soluzione, accettando una minore accuratezza, ma garantendo internamente il rispetto del vincolo, secondo cui la somma degli IMTP su ciascun periodo deve essere pari a uno.

Conclusioni

Lo studio condotto ha evidenziato come l'analisi della *Digital Voice of Customer* (Digital VoC) rappresenti un valido strumento per il monitoraggio della qualità aziendale, superando i limiti delle metodologie tradizionali, quali i questionari e le interviste.

Attraverso l'applicazione dell'algoritmo *Structural Topic Modelling* (STM), è stato possibile identificare i temi chiave discussi dai clienti all'interno del corpus delle recensioni, che sono ricondotti alle determinanti di qualità del prodotto o servizio offerto dalle aziende.

Partendo dai risultati dell'STM è stato possibile valutare l'*Interval Mean Topical Prevalence* (IMTP), cioè quanto mediamente viene discusso ciascun topic in uno specifico instante temporale che, nel caso in esame, è stato considerato pari ad un mese.

L'obiettivo dell'elaborato era non solo estrapolare dal corpus delle recensioni i temi più rilevanti discussi, ma anche valutare la discussione dei topic nel tempo mediante l'applicazione di modelli capaci di condurre analisi predittive sui valori futuri. Lo studio si è focalizzato sulla risoluzione di un problema di regressione su serie storiche, considerando un intervallo temporale di circa 10 anni.

L'aspetto innovativo della ricerca risiede nell'applicazione di modelli predittivi e nello sviluppo di codici Python per il monitoraggio della qualità aziendale. Data l'importanza dell'analisi predittiva in diversi ambiti, come ad esempio, la previsione del prezzo delle azioni o l'analisi meteorologica, i codici sono stati implementati in modo da ottimizzare i parametri e garantire previsioni accurate.

L'obiettivo della ricerca è stato quello di sviluppare e testare modelli avanzati di *predictive analytics* per migliorare il monitoraggio della qualità, applicandoli a due casi studio: Disneyland e Ryanair.

A tal fine, sono stati implementati due modelli differenti:

i. *Seasonal Auto-Regressive Integrated Moving Average* (SARIMA) e
ii. la rete neurale ricorrente *Long Short-Term Memory* (LSTM).

Attraverso un adeguato processo di addestramento dei modelli, suddividendo il dataset in training e test, sono state condotte diverse iterazioni fino ad ottenere risultati soddisfacenti.

Tra i modelli analizzati, l'LSTM si è rivelato il più efficace grazie alla sua capacità di gestire sequenze di dati complesse e rumorose, fornendo previsioni più accurate rispetto SARIMA.

Quest'ultimo, infatti, ha mostrato limiti significativi, specialmente in presenza di fattori esogeni come la pandemia da COVID-19, approssimando i valori a una media senza cogliere adeguatamente le variazioni della serie.

L'integrazione di queste tecniche con gli attuali sistemi di gestione della qualità potrebbe apportare significativi vantaggi competitivi, consentendo alle aziende di adattarsi dinamicamente alle esigenze del mercato. Inoltre, lo studio apre la strada a nuove prospettive di ricerca, suggerendo la possibilità di estendere l'analisi predittiva alla valutazione del *Mean Rating Proportion* (MRP), ovvero dell'evoluzione della soddisfazione media del cliente associata a ciascun topic.

Infine, la ricerca ha rappresentato un'importante occasione di crescita professionale, permettendomi di consolidare le conoscenze acquisite durante il percorso accademico. Lo sviluppo dei codici e la loro ottimizzazione hanno richiesto un approfondimento delle tecniche di machine learning e del linguaggio Python, contribuendo così al mio arricchimento culturale.

Allegati

Allegato 1: Codice STM implementato per il database Disneyland

```
library(stm)

library(tm)

library(stringr)

library(igraph)

data <- read.csv2("Disneyland.csv", fileEncoding = "UTF-8")    # Lettura file csv

rimuovere <- read.csv2("Stopword.csv", fileEncoding = "UTF-8")

processed <- textProcessor(data$Review, metadata = data, customstopwords =
rimuovere$Stopwords, verbose=TRUE) # PRE-PROCESSING

out <- prepDocuments(processed$documents, processed$vocab, processed$meta,
lower.thresh=15, verbose=TRUE )

docs <- out$Review

vocab <- out$vocab

meta <-out$meta

# Creazione file con le reviews rimanenti.

write.csv (meta$Fonte,file='Review_ID_Rimanenti.csv')

write.csv (meta$Country,file='Rating_Rimanenti.csv')

write.csv (meta$Provider,file='Year_Month_Rimanenti.csv')

write.csv (meta$Type,file='Reviewer_Location_Rimanenti.csv')

write.csv (meta$Data,file='Review_Text_Rimanenti.csv')

write.csv (meta$Rating,file='Branch_Rimanenti.csv')

c=(5:100) # c vettore con numeri da 5 a 100

K<-c

storage <- searchK(out$documents, out$vocab, K, data = meta)

write.csv(unlist(storage$results ,file='optimizationresults.csv')

# Applicazione STM

k=13 # Impostare il numero di topic da estrarre

poliblogPrevFit <- stm(documents=out$documents, vocab=out$vocab,K=k ,max.em.its=75,
data=out$meta, init.type="Spectral")

write.csv (poliblogPrevFit$theta,file='matrice_review_topics.csv')

labelTopics(poliblogPrevFit, c(1:13), n=13) # Etichettatura dei topic

t=1 # Numero dei topic di cui si vogliono visualizzare i documenti più
significativi

d=10 # Numero  dei documenti da visualizzare

thoughts <-findThoughts(poliblogPrevFit, texts=out$meta$Review,n=d,topics=t)

a=topicCorr(poliblogPrevFit, method = c("simple", "huge"), cutoff = 0.10, verbose =
TRUE) # Correlazione tra i topic

plot.topicCorr(a)
```

Allegato 2: Labeling Disneyland

TOPIC	HIGHEST PROB	FREX	LIFT	SCORE
1	charact, photo, princess, pictur, mous, minni, greet, queu, children, walk, meet, girl, met	princess, minni, autograph, photo, meet, mous, pictur, charact, goofi, girl, elsa, photograph, greet	aurora, daisi, anna, rapunzel, eeyor, elsa, pavillion, dale, pluto, poppin, pavilion, mari, tigger	charact, aurora, princess, photo, minni, pictur, mous, autograph, meet, greet, goofi, photograph, donald
2	park, differ, compar, area, walk, smoke, origin, ride, fan, studio, disneyworld, size, usa	disneyworld, smoke, compar, origin, park, comparison, similar, size, uniqu, design, usa, version, differ	recal, los, angel, enforc, disneyworld, epcot, coast, smoker, butt, compact, smoke, properti, layout	park, recal, smoke, compar, disneyworld, differ, origin, area, version, similar, fan, usa, size
3	crowd, season, holiday, trip, summer, weekend, break, weather, way, stroller, school, hot, annual	crowd, summer, season, school, annual, peak, weekend, heat, stroller, januari, mid, crazi, holder	presid, labor, lineup, crowd, passhold, oct, crow, midweek, season, summer, renew, januari, congest	crowd, presid, season, summer, stroller, annual, weekend, school, holiday, break, hot, peak, weather
4	book, check, train, arriv, entranc, onlin, access, station, gate, disabl, bag, travel, secur	disabl, card, wheelchair, onlin, hall, airport, voucher, book, inform, guid, via, check, citi	asap, disabl, http, termin, wheelchair, proof, desk, card, deposit, letter, email, valle, driver	asap, book, disabl, onlin, check, voucher, wheelchair, card, airport, bag, station, citi, train
5	hotel, stay, excel, room, breakfast, clean, studio, walk, villag, night, half, servic, shuttle	excel, stay, hotel, facil, room, lodg, pool, breakfast, villag, santa, bed, york, stunt	architectur, newport, sequoia, buffalo, lodg, cheyenn, continent, york, pool, bed, davi, crockett, spur	hotel, architectur, stay, room, excel, breakfast, villag, facil, shuttl, buffet, lodg, bus, sequoia
6	close, disappoint, money, ride, price, staff, servic, rude, experi, work, poor, let, custom	poor, rude, custom, ridicul, worst, money, dirti, terribl, technic, complain, sad, rip, wors	econom, greed, profit, filthi, unaccept, earn, disgrac, corpor, farm, disgust, dirti, angri, appal	econom, money, price, disappoint, staff, rude, poor, custom, servic, close, ridicul, dirti, rip
7	magic, staff, friend, member, help, cast, christma, experi, wonder, clean, fantast, trip, birthday	magic, cast, birthday, member, atmospher, smile, detail, attent, christma, incred, friend, wow, wonder	truth, trueli, xmas, birthday, describ, dampen, unforgett, outstand, cast, boyfriend, blown, blew, magic	truth, magic, christma, member, cast, staff, friend, birthday, help, wonder, clean, fantast, decor
8	pass, fast, ride, plan, fastpass, system, popular, app, morn, hopper, get, open, singl	fast, pass, app, hopper, system, fastpass, popular, rider, singl, track, advantag, max, download	passrid, maxpass, fast, app, pass, max, rider, distribut, hopper, utilis, download, util, strategi	: pass, fast, passrid, fastpass, app, hopper, system, plan, ride, rider, popular, download, singl
9	kid, place, ride, famili, children, adult, experi, year, age, child, definit, trip, memori	kid, adult, age, childhood, place, famili, older, memori, worth, game, teenag, child, life	unnecessari, reliv, childhood, alik, atract, satisfi, grandpar, game, worth, lifetim, fond, grown, inner	kid, place, unnecessari, famili, adult, children, ride, age, year, childhood, child, memori, experi
10	mountain, space, ride, star, pirat, close, thunder, jone, tour, indiana, coaster, haunt, caribbean	mountain, pirat, jone, indiana, caribbean, thunder, war, space, roller, railroad, coaster, star, haunt	measur, submarin, sawyer, peril, railroad, caribbean, hyper, pirat, indiana, jone, war, mountain, nemo	mountain, measur, space, jone, indiana, pirat, thunder, haunt, star, mansion, caribbean, splash, coaster
11	food, expens, restaur, drink, eat, price, shop, meal, lunch, cost, snack, option, bottl	drink, bottl, burger, food, fri, ice, cream, sandwich, chicken, pizza, eat, snack, choic	immens, gluten, pasta, sauc, chicken, rib, veg, dish, veggi, rice, fri, muslim, beef	food, immens, drink, expens, eat, meal, restaur, price, euro, snack, bottl, burger, lunch
12	show, parad, firework, night, watch, light, ride, castl, even, street, view, rain, display	firework, show, king, lion, display, parad, watch, light, rain, view, night, spectacular, perform	grad, lion, wondrous, tung, king, chung, philharmag, firework, lantau, display, vantag, windi, golden	grad, firework, show, parad, night, lion, king, watch, light, mystic, castl, display, even
13	halloween, ride, theme, parti, event, treat, open, experi, special, set, various, toward, street	halloween, event, parti, theme, sight, various, sound, trick, nightmar, gear, treat, candi, toward	onward, pumpkin, deck, gear, sight, halloween, transform, trick, event, tradit, candi, parti, sound	onward, halloween, theme, event, parti, treat, trick, sight, decor, costum, various, gear, candi

Allegato 3: Labeling Ryanair

TOPIC	HIGHEST PROB	FREX	LIFT	SCORE
1	plane, flight, attend, wait, minut, board, water, staff, passeng, return	attend, water, toilet, bottl, rain, outsid, tarmac, sign, cold, whilst	economi, dusseldorf, soak, hungri, realiti, fuel, unload, rain, bathroom, outgo	economi, water, attend, plane, toilet, bottl, rain, wait, stand, tarmac
2	airlin, fli, book, year, travel, flight, use, time, return, holiday	year, holiday, budget, flown, old, midland, east, class, cant, regular	class, yrs, librari, luxuri, airway, add-on, east, midland, southend, oct	class, year, holiday, airlin, book, midland, east, flown, fli, old
3	flight, time, crew, friend, cabin, return, board, price, arriv, service	pleasant, effici, valu, excel, friend, legroom, clean, comfort, bit, earli	congest, unfound, gothenburg, genoa, wing, bumpi, newer, rygg, pitch, faultless	crew, unfound, pleasant, effici, excel, clean, cabin, legroom, comfort, valu
4	crew, passeng, cabin, staff, flight, help, member, assist, ask, aircraft	assist, member, english, wheelchair, safeti, special, languag, manner, steward, medic	verbal, conduct, drank, unwel, allergi, jacket, speaker, distress, stag, mood	verbal, crew, cabin, assist, wheelchair, member, passeng, jacket, languag, help
5	bag, luggag, pay, extra, hand, baggag, staff, paid, small, check	luggag, bag, suitcas, size, small, hand, fit, hold, weight, carri	precaut, overweight, heavier, conveyor, kilo, suitcas, bigger, fit, size, underneath	bag, luggag, precaut, size, suitcas, hand, carri, weight, pay, hold
6	flight, hour, delay, time, airport, arriv, stanst, late, fli, inform	delay, hotel, night, hour, connect, hrs, coach, leav, suppos	poznan, nant, sunni, lift, discount, midnight, dec, reschedul, coach, overnight	delay, poznan, hour, cancel, hotel, flight, compass, late, inform, night
7	board, gate, prioriti, queue, wait, minut, secur, time, flight, long	queue, gate, prioriti, secur, drop, plus, walk, board, close, open	reus, settl, boarder, unforgiv, croissant, squash, quicker, electron, lane, breez	reus, prioriti, gate, board, queue, secur, wait, walk, line, drop
8	check, pass, airport, onlin, pay, print, board, charg, check-	onlin, euro, counter, fee, pass, check-, check, app, printer	hide, arbitrarili, canadian, dollar, onlin, panel, counter, barcod, stamp, santorini	hide, print, check, onlin, fee, pass, euro, check-, counter, charg
9	flight, custom, servic, refund, book, chang, tri, email, cancel, company	refund, contact, chang, claim, email, repli, answer, chat, confirm, websit	com, fraud, que, para, por, refer, govern, covid, screw, refund	com, refund, cancel, email, contact, que, claim, voucher, chang, insur
10	seat, sit, flight, togeth, pay, plane, alloc, row, extra, next	togeth, seat, row, alloc, uncomfort, random, apart, sit, empti, constant	garish, vacant, fleet, tray, insuffici, harass, adjac, wizz, occupi, random	seat, garish, alloc, sit, togeth, row, leg, space, random, aisl

Allegato 4: codice SARIMA applicato al topic 13 “Halloween ed eventi speciali”

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from statsmodels.tsa.seasonal import seasonal_decompose
from statsmodels.tsa.stattools import adfuller
from pmdarima import auto_arima
import warnings
from statsmodels.tsa.arima.model import ARIMA
from sklearn.metrics import mean_squared_error
df = pd.read_csv('imtp13.csv', sep=';') # Caricare il file csv
df.set_index('date', inplace=True)
df = df.asfreq('MS') # Viene preso come riferimento il valore di inizio mese
df = df.sort_index() # I valori vengono ordinati in base alle date
# Rappresentazione grafica del dataset
plt.figure(figsize=(10,6))
plt.plot(df['imtp13'], label='imtp13')
plt.title('Serie Temporale IMTP T13')
plt.xlabel('Data')
plt.ylabel('IMTP')
plt.legend()
plt.show()
# Suddivisione della serie in stagionalità, trend e residui
result = seasonal_decompose(df['imtp13'], model='additive')
trend = result.trend
seasonal = result.seasonal
residual = result.resid
plt.figure(figsize=(12,8))
plt.subplot(411)
plt.plot(df['imtp13'], label='Originale') # Serie originale
plt.legend(loc='best')
plt.subplot(412)
plt.plot(trend, label='Trend') # Rappresentazione del trend
plt.legend(loc='best')
plt.subplot(413)
plt.plot(seasonal, label='Stagionalità') # Rappresentazione della componente
stagionale
plt.legend(loc='best')
```

```
plt.subplot(414)
plt.plot(residual, label='Residuo') # Rappresentazione del residuo
plt.legend(loc='best')
plt.tight_layout()
plt.show()
# Test di stazionarità: Test di Dickey-Fuller aumentato
def adf_test(series):
    result = adfuller(series.dropna())  # Dropna perché il test non accetta valori mancanti
    print('Test ADF:')
    labels = ['ADF Statistic', 'p-value', '#Lags Used', 'Number of Observations Used']
    for value, label in zip(result, labels):
        print(f'{label}: {value}')
    if result[1] <= 0.05:
        print("La serie è stazionaria (rifiutiamo l'ipotesi nulla)")
    else:
        print("La serie non è stazionaria (non rifiutiamo l'ipotesi nulla)")
adf_test(df['imtp13'])
# Differenziazione per ottenere la stazionarietà nella serie
df['imtp13_diff'] = df['imtp13'].diff().dropna()
adf_test(df['imtp13_diff'].dropna())
# Suddivisione del dataset in training e test
train_size = int(len(df) * 0.8)
train, test = df.iloc[:train_size], df.iloc[train_size:]
# Parametri del modello stabiliti in modo automatico dalla funzione auto_arima
model=auto_arima(df['imtp13'], seasonal=True, m=12, trace=True,
error_action='ignore', suppress_warnings=True)
print(model.summary())
best_order = model.order
best_seasonal_order = model.seasonal_order
# Addestramento del modello sui dati di training in base ai parametri scelti
sarima_model = ARIMA(train['imtp13'], order=best_order,
seasonal_order=best_seasonal_order)
sarima_model_fit = sarima_model.fit()
# Previsioni sul set di dati di test
forecast_steps = len(test)
forecast = sarima_model_fit.forecast(steps=forecast_steps)
forecast_index = test.index
```

```
# Si crea un unico dataset per il training e il test

combined_data = pd.concat([train, test])

# Plot del grafico

plt.figure(figsize=(12, 6))

# Si plottano i dati di training come parte della serie combinata

plt.plot(combined_data.index, combined_data['imtp13'], label='Dati di Training e
Test', color='blue')

# Plot delle previsioni come continuità

plt.plot(forecast_index, forecast, label='Previsione SARIMA', color='red')

# Si aggiungono etichette e titolo

plt.xlabel('Data')

plt.ylabel('IMTP')

plt.title('Previsione con SARIMA')

plt.legend()

plt.show()

# Calcolo dell'errore quadratico medio

mse = mean_squared_error(test['imtp13'], forecast)

rmse = np.sqrt(mse)

print('RMSE:', rmse)
```

Allegato 5: codice LSTM applicato al topic 13 “Halloween ed eventi speciali”

```
import numpy as np array
import pandas as pd dataframe
from sklearn.preprocessing import MinMaxScaler
from keras.models import Sequential
from keras.layers import LSTM, Dense
from sklearn.model_selection import train_test_split
import matplotlib.pyplot as plt
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
df = pd.read_csv('imtp13.csv', sep=';') # Caricare il file csv
data = df['imtp13'].values
# Funzione per creare sequenze di dati per la rete neurale
def create_sequences(data, window_size):
    X, y = [], []
    for i in range(len(data) - window_size):
        X.append(data[i:i + window_size])  # Finestra temporale di input
        y.append(data[i + window_size])    # Valore successivo da predire
    return np.array(X), np.array(y)
window_size = 10  # Dimensione della finestra temporale
# Creazione delle sequenze (X = input, y = target/output)
X, y = create_sequences(data, window_size)
# Suddivisione in training (80%) e test (20%)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42, shuffle=false)
# Rimodellamento dei dati per l'input della LSTM
X_train = np.reshape(X_train, (X_train.shape[0], X_train.shape[1], 1))
X_test = np.reshape(X_test, (X_test.shape[0], X_test.shape[1], 1))
# Creazione del modello LSTM
model = Sequential()
model.add(LSTM(150, return_sequences=False, input_shape=(X_train.shape[1], 1)))#150
unità LSTM
model.add(Dense(1))  # Strato denso di output con 1 unità (previsione)
# Compilazione del modello e calcolo della perdita quadratica media
model.compile(optimizer='adam', loss='mean_squared_error')
# Addestramento del modello
    train_history = model.fit(X_train, y_train, epochs=250, batch_size=16,
validation_data=(X_test, y_test), verbose=1, shuffle=False)
# Fare una previsione sui dati di test
predictions = model.predict(X_test)
```

```
# Calcolo delle metriche di errore

mae = mean_absolute_error(y_test, predictions)

mse = mean_squared_error(y_test, predictions)

rmse = np.sqrt(mse)

# Stampa delle metriche

print(f"Mean Absolute Error (MAE): {mae}")

print(f"Mean Squared Error (MSE): {mse}")

print(f"Root Mean Squared Error (RMSE): {rmse}")

# Visualizzazione dei risultati

date_labels = pd.date_range(start='2020-01-01', periods=len(data),
freq='M').strftime('%Y-%m').tolist() # Creazione delle etichette per l'asse x (mesi
e anni)

# Grafico dei valori effettivi e dei valori predetti

plt.figure(figsize=(14, 7))

plt.plot(date_labels, data, label='Valori Originali', color='blue', marker='o',
markersize=2, alpha=0.5)

plt.plot(date_labels[-len(y_test):], predictions.flatten(), label='Valori
Predetti', color='orange', marker='x')

plt.title('Confronto tra Valori Originali e Valori Predetti')

plt.xlabel('Data (Mesi e Anni)')

plt.ylabel('IMTP')

plt.xticks(rotation=45)  # Rotazione delle etichette dell'asse x per una migliore
leggibilità

plt.legend()

plt.grid()

plt.tight_layout()  # Evita che le etichette dell'asse x vengano tagliate

plt.show()
```

Allegato 6: codice LSTM seconda soluzione applicato al caso studio Disneyland

```
import pandas as pd

import numpy as np

from sklearn.model_selection import train_test_split

from sklearn.preprocessing import MinMaxScaler

from sklearn.metrics import mean_squared_error

import matplotlib.pyplot as plt

from tensorflow.keras.models import Sequential

from tensorflow.keras.layers import LSTM, Dense, Softmax

df = pd.read_csv("matrice.csv", sep=';')  # Caricare il file csv

dates = pd.to_datetime(df['date'], format='%Y-%m') # Converte le date in formato
datetime

data = df.drop(columns=['date']).to_numpy()

# Creare sequenze per LSTM

def create_sequences(data, sequence_length=12):

    X, y = [], []

    for i in range(len(data) - sequence_length):

        X.append(data[i:i + sequence_length])  # Prendi sequence_length righe

        y.append(data[i + sequence_length])   # La riga successiva come target

    return np.array(X), np.array(y)

sequence_length = 12 # Usa 12 mesi per prevedere il successivo

X, y = create_sequences(data, sequence_length)

# Suddivisione in training e test (80-20)

train_size = int(len(X) * 0.8)

X_train, X_test = X[:train_size], X[train_size:]

y_train, y_test = y[:train_size], y[train_size:]

test_dates = dates[sequence_length + train_size:]

# Costruire il modello LSTM

model = Sequential([

    LSTM(64, return_sequences=True, input_shape=(sequence_length, 13)),  # Primo
strato LSTM

    LSTM(32),  # Secondo strato LSTM

    Dense(13, activation='softmax')  # Uscita con 13 unità e softmax

])

model.compile(optimizer='adam', loss='mean_squared_error', metrics=['mse'])

model.fit(X_train, y_train, validation_data=(X_test, y_test), epochs=50,
batch_size=16, verbose=1)

# Fare previsioni

y_pred = model.predict(X_test)
```

```
# Calcolo dell'RMSE per ciascun topic

rmse_per_topic = np.sqrt(np.mean((y_test - y_pred) ** 2, axis=0))

print ("\nRMSE per ciascun topic:")

for i, rmse in enumerate (rmse_per_topic, start=1):

    print (f"Topic {i}: {rmse:.4f}")

# Creazione di un DataFrame per combinare i valori reali e predetti

combined_df = pd.DataFrame(data, columns=[f'Topic {i+1}' for i in
range(data.shape[1])])

combined_df['date'] = dates

for i in range(data.shape[1]):

    combined_df[f'Topic {i+1}_Pred'] = np.nan

combined_df.loc[sequence_length + train_size:, [f'Topic {i+1}_Pred' for i in
range(data.shape[1])]] = y_pred

# Grafici delle previsioni per ciascun topic

for i in range(data.shape[1]):

    plt.figure(figsize=(10, 6))

    plt.plot(dates, combined_df[f'Topic {i+1}'], label=f'Topic {i+1} (Reale)',
linestyle='dashed', color='blue', alpha=0.7)

    plt.plot(dates[sequence_length + train_size:], y_pred[:, i], label=f'Topic
{i+1} (Predetto)', color='orange', alpha=0.7)

    plt.title(f'Topic {i+1}: Confronto tra Valori Reali e Predetti')

    plt.xlabel('Date')

    plt.ylabel('IMTP')

    plt.legend()

    plt.grid()

    plt.xticks(rotation=45)

    plt.tight_layout()

    plt.show()
```

Bibliografia

[1] Barravecchia, F., Mastrogiacomo, L., & Franceschini, F. (2020). Categorizing quality determinants in mining user-generated contents. *Sustainability*, *12*(23), 9944.

[2] Barravecchia, F., Mastrogiacomo, L., & Franceschini, F. (2023). Product quality tracking based on digital Voice-of-Customers. *Total quality management & business excellence*, 34(11-12), pp.1386-1409.

[3] Barravecchia, F., Mastrogiacomo, L., & Franceschini, F. (2022). Digital voice-of-customer processing by topic modelling algorithms: insights to validate empirical results. *International Journal of Quality & Reliability Management*, *39*(6), pp.1453-1470.

[4]*"Metodi raccolta dati primari."* A Scuola di OpenCoesione, https://www.ascuoladiopencoesione.it/sites/default/files/asoc_files/1819/doc/3.5_Metodi_Raccolta_dati_primari.pdf. Accesso dicembre 2024.

[5] *"Text Mining."* IBM, https://www.ibm.com/it-it/topics/text-mining. Accesso dicembre 2024.

[6] Barravecchia, F., Mastrogiacomo, L., Tavani, L., & Franceschini, F. (2022). Statistical Process Control techniques to monitor quality determinants in digital Voice-of-Customer. *In Proceedings of the 5th ICQEM Conference*, pp.120-141.

[7] Jiang, D., Zhang, C., & Song, Y. (2023). *Probabilistic Topic Models: Foundation and Application*. Springer.

[8] Hofmann, T. (1999). PLSA Probabilistic Latent Semantic Analysis. *Uncertainity in Articial Intelligence* (UAI'99), pp. 289-296.

[9] Blei, D., Ng., A., & Jordan, M. (2003). Latent Dirichlet Allocation. *The Journal of Machine Learning Research.*

[10]*"Introduction to the Structural Topic Model (STM)."* Towards Data Science, https://towardsdatascience.com/introduction-to-the-structural-topic-model-stm-34ec4bd5383. Accesso dicembre 2024.

[11] Blei, D., & Lafferty, J. (2005). Correlated Topic Models. NIPS'05: *Proceedings of the 18th International Conference on Neural Information Processing Systems.*

[12] Roberts, M., Stewart, B., & Tingle D. (2019). stm: An R Package for Structural Topic Models. *Journal of Statistical Software*, 91(2), pp.1-40.

[13] Roberts M., Stewart, B. & Airoldi E., (2016) A Model of Text for Experimentation in the Social Sciences. *Journal of the American Statistical Association*, 111, pp.1-49.

[14] Barravecchia, F., Mastrogiacomo, L., Tavani, L., Franceschini, F. & Marimon, F. (2021). Mining quality determinants of product-service systems from user-generated contents. *Quality Engineering,* pp. 425-442.

[15] Barravecchia, F., Mastrogiacomo, L., & Franceschini, F. (2022). KA-VoC Map: Classifying product Key-Attributes from digital Voice-of-Customer. *Quality Engineering, 34* (3), pp.344–358.

[16]*"Cosa è l'apprendimento supervisionato."* IBM, https://www.ibm.com/it-it/topics/supervised-learning. Accesso dicembre 2024.

[17] *"Che cosa è un albero decisionale."* IBM, https://www.ibm.com/it-it/topics/decision-trees. Accesso dicembre 2024.

[18] Weidman, S. (2020). *Deep Learning*. O'REILLY (Milano).

[19]Saleem, S. *"Neural Networks in 10mins: Simply Explained."* Medium, https://medium.com/@sadafsaleem5815/neural-networks-in-10mins-simply-explained-9ec2ad9ea815. Accesso dicembre 2024. Accesso dicembre 2024.

[20] D'agostino, A. *"Valutazione delle performance di un modello di classificazione."* Diario di un Analista, https://www.diariodiunanalista.it/posts/valutazione-delle-performance-di-un-modello-di-classificazione/. Accesso dicembre 2024.

[21] D'agostino, A. *"Valutazione delle prestazioni di un modello di regressione."* Diario di un Analista, https://www.diariodiunanalista.it/posts/valutazione-delle-prestazioni-di-un-modello-di-regressione/. Accesso dicembre 2024.

[22] Bee Dagum, E. (2002). *Analisi delle serie storiche*. Springer (Bologna).

[23] J Hyndman, R. & Athanasopoulos., G. (2021). *Previsione: principi e pratica (III ediz.).* OTEXTS (Australia). https://otexts.com/fppit/seasonal-arima.html. Accesso dicembre 2024.

[24]*"Pandas Introduction."* W3Schools, https://www.w3schools.com/python/pandas/pandas_intro.asp. Accesso dicembre 2024.

[25] *"NumPy."* Python. (2023). https://wiki.python.org/moin/NumPy. Accesso dicembre 2024.

[26] Hunter, J., Dale, D., Firing, E., Droettboom, M., & il team di sviluppo di Matplotlib. (2024). *"Getting Started with Matplotlib."* https://matplotlib.org/stable/users/getting_started/index.html. Accesso dicembre 2024.

[27]*"statsmodels.tsa.seasonal.seasonal_decompose."* StatsModels, 2024. https://www.statsmodels.org/dev/generated/statsmodels.tsa.seasonal.seasonal_decompose.html. Accesso dicembre 2024.

[28]*"statsmodels.tsa.stattools.adfuller."* StatsModels, 2024. https://www.statsmodels.org/dev/generated/statsmodels.tsa.stattools.adfuller.html. Accesso dicembre 2024.

[29]“*pmdarima.arima.auto_arima."* Pmdarima, 2024.

https://alkaline-ml.com/pmdarima/modules/generated/pmdarima.arima.auto_arima.html. Accesso dicembre 2024.

[30]“*statsmodels.tsa.arima.model.ARIMA."* StatsModels, 2024. https://statsmodels.org/stable/generated/statsmodels.tsa.arima.model.ARIMA.html. Accesso dicembre 2024.

[31]*"mean_squared_error."* Scikit-learn, 2024.

https://scikit-learn.org/1.5/modules/generated/sklearn.metrics.mean_squared_error.html. Accesso dicembre 2024.

[32] *"Serie temporali di previsione: come utilizzare l’analisi e la previsione delle serie temporali per i tuoi dati finanziari."* FasterCapital, 2024. https://fastercapital.com/it/contenuto/Serie-temporali-di-previsione--come-utilizzare-l-analisi-e-la-previsione-delle-serie-temporali-per-i-tuoi-dati-finanziari.html. Accesso dicembre 2024.

[33] *"Econometria: Eleganza econometrica: il ruolo del test Dickey Fuller."* FasterCapital, 2024. https://fastercapital.com/it/contenuto/Econometria--Eleganza-econometrica--il-ruolo-del-test-Dickey-Fuller.html#Introduzione-all-econometria-e-al-test-di-Dickey-Fuller. Accesso dicembre 2024.

[34]*" Convalida della previsione come convalidare e testare modelli di previsione per la previsione degli investimenti."* FasterCapital, 2024.

https://fastercapital.com/it/contenuto/Convalida-della-previsione--come-convalidare-e-testare-modelli-di-previsione-per-la-previsione-degli-investimenti.html#suddivisione-dei-dati-e valutazione-delle-prestazioni-del-modello.html. Accesso dicembre 2024.

[35] Gahlawat, N. *"Recurrent Neural Networks for Dummies."* Plain English AI.

https://ai.plainenglish.io/recurrent-neural-networks-for-dummies-70991a87e5d7.Accesso dicembre 2024.

[36]*"sklearn.preprocessing.MinMaxScaler."* Scikit-learn,

https://scikit-learn.org/1.5/modules/generated/sklearn.preprocessing.MinMaxScaler.html. Accesso dicembre 2024.

[37] *"Models API."* Keras, https://keras.io/api/models/model/. Accesso dicembre 2024.

[38] *"Layers API."* Keras, https://keras.io/api/layers/. Accesso dicembre 2024.

[39]*"sklearn.model_selection.train_test_split."* Scikit-learn,

https://scikit-learn.org/1.6/modules/generated/sklearn.model_selection.train_test_split.html. Accesso dicembre 2024.

[40] Brownlee, J. "*How to reshape Input Data for Long Short-Term Memory Networks in Keras."* (2019). https://machinelearningmastery.com/reshape-input-data-long-short-term-memory-networks-keras/. Accesso dicembre 2024.

[41] De Bernardinis, R. "*Guida Creazione Modello LSTM Keras: Passo Passo."* (2024). https://www.riccardodebernardinis.com/blog/guida-creazione-modello-lstm-keras-passo-passo/. Accesso dicembre 2024.

[42] **"***Esempio pratico di Deep Learning con Python: Previsione del prezzo delle azioni."* Intelligenza artificiale Italia, https://www.intelligenzaartificialeitalia.net/post/esempio-pratico-di-deep-learning-con-python-previsione-del-prezzo-delle-azioni. Accesso dicembre 2024.

[43] Huang, F., Li, X., Yuan, C., Zhang, S., Zhang, J. & Qiao, S. (2022). Attention-Emotion-Enhanced Convolutional LSTM for Sentiment Analysis. *Transactions on Neural Networks and Learning Systems*, 33(9), pp. 4332-4345.

[44] Elsaraiti, M., & Merabet, A. (2021). A Comparative Analysis of the ARIMA and LSTM Predictive Models and Their Effectiveness for Predicting Wind Speed. *Energies*, *14* (20).

[45] Barsotti, G. "*Coronavirus e rimborso voli aerei. Ryanair cambia rotta: non effettua più rimborsi in denaro ed emette solo voucher*." ADUC. (2020). https://www.aduc.it/comunicato/coronavirus+rimborso+voli+aerei+ryanair+cambia_31065.php. Accesso dicembre 2024.

[46] "*Ryanair in crisi a causa del covid: riduzione dei voli e chiusura di basi*." Flycare (2020).

https://flycare.eu/ryanair-in-crisi-a-causa-del-covid-riduzione-dei-voli-e-chiusura-di-basi/. Accesso dicembre 2024.

[47] Pizzimenti, C. "*I voli cancellati? È perché i piloti scappano da Ryanair*". VANITYFAIR. (2017).https://www.vanityfair.it/viaggi-traveller/notizie-viaggio/2017/09/19/ryanair-voli-annullati-piloti. Accesso dicembre 2024.

[48]*"Preprocessing data."* Scikit-learn,

https://scikit-learn.org/stable/modules/preprocessing.html. Accesso dicembre 2024.

[49]*"Cross-validation and model selection."* Scikit-learn, https://scikitlearn.org/stable/modules/cross_validation.html#cross-validation-and-model-selection. Accesso dicembre 2024.

[50]*"tf.keras.Model."* TensorFlow, https://www.tensorflow.org/api_docs/python/tf/keras/Model. Accesso dicembre 2024.

[51] *"tf.keras.Layer."* TensorFlow, https://www.tensorflow.org/api_docs/python/tf/keras/Layer. Accesso dicembre 2024.

[52]"Softmax." Pytorch, https://pytorch.org/docs/stable/generated/torch.nn.Softmax.html. Accesso dicembre 2024.