POLITECNICO DI TORINO

Corso di Laurea Magistrale in Ingegneria Gestionale

Tesi di Laurea Magistrale

Estrazione delle determinanti di qualità dei servizi aerei tramite analisi di User Generated Contents.

Relatore: Prof. Luca Mastrogiacomo

Candidato: Mattia Branda

Anno accademico: 2019/2020

INDICE

INDICE DELLE FIGURE

INDICE DELLE TABELLE

PREMESSA

In passato l'economia era basata principalmente sulla concezione di prodotto e di vendita di un prodotto al cliente. A partire dagli ultimi decenni del secolo scorso in ogni settore dell'economia si è concepita una sempre più vasta concezione del servizio e della soddisfazione del cliente e ogni servizio viene concepito con l'intento di mettere il cliente al centro del processo di realizzazione e non che esso ne sia estraneo.
Con l'evoluzione degli strumenti di comunicazione e con l'avanzamento della tecnologia in ogni settore industriale è stato possibile creare una sempre più vasta gamma di servizi e le principali aziende manifatturiere hanno dovuto integrare tra loro il mondo dei prodotti e quello dei servizi e hanno dovuto sperimentare nuove strategie di marketing e di analisi del comportamento di un cliente nei confronti di un servizio.

Il seguente lavoro di tesi si pone l'obiettivo di individuare la qualità del servizio del trasporto aereo di passeggeri fornito da 10 compagnie aeree selezionate che costituiscono l'oggetto d'indagine, studiando il comportamento dei clienti tramite l'analisi degli User Generated Contents, ossia contenuti che vengono immessi in rete su forum, blog, Social Network, ecc.
La tecnica utilizzata per compiere quest'indagine qualitativa del servizio è la Text Mining Analysis, ovvero metodologie in grado di estrarre informazioni su quale sia il pensiero e i principali argomenti trattati (topics) negli UGC.
Nel dettaglio lo strumento che è stato utilizzato per poter fare questo tipo di analisi è stato l'algoritmo STM, implementato tramite il software statistico R, che ha permesso di estrarre le informazioni dagli UGC raccolti dalla rete.
Infine è stato presentato un focus sull'analisi temporale delle informazioni estratte tramite l'algoritmo STM, ovvero è stato studiato il comportamento in un determinato arco temporale di ogni topic discusso all'interno delle recensioni raccolte.

Il lavoro di tesi si sviluppa in 5 capitoli.
Il primo capitolo tratta la descrizione dei servizi, il concetto di qualità nei servizi e di qualità nel trasporto aereo di passeggeri e infine vi è un breve foucs sul mercato del trasporto aereo di passeggeri europeo e mondiale.

Nel secondo capitolo sono esposte le metodologie, i modelli e gli algoritmi utilizzati per compiere la Text Mining Analysis delle informazioni estratte dalla rete.

Nel terzo capitolo vengono esposti i risultati ottenuti facendo un confronto sulle 10 compagnie aeree suddividendole in base all'area geografica di appartenenza e al segmento di mercato in cui esse operano.
Inoltre è presente un focus sul settore low cost ed un confronto tra i due principali players Europei (Easyjet e Ryanair).
Al termine del terzo capitolo è presente un'analisi delle correlazioni che vi sono tra i vari argomenti (topic) discussi nelle recensioni raccolte e le famiglie di topic (clustering) che si possono osservare.

Il quarto capitolo si concentra sull'analisi dell'andamento temporale di ciascun topic sul totale delle recensioni del dataset preprocessato (escluse le recensioni raccolte che non hanno una data di pubblicazione e classificate come NULL).

Infine il quinto capitolo presenta le conclusioni che è stato possibile desumere dalle analisi effettuate nei capitoli precedenti.

CAPITOLO 1

1.1. I SERVIZI

Il settore dei servizi sta affrontando in questi ultimi anni un grande momento di crescita e sta modificando il modo di concepire l'economia, i rapporti tra gli individui e la società.
Le economie dei Paesi industrializzati sono sempre più orientate verso il settore dei servizi. L'indice di terziarizzazione dell'economia, misurato dal rapporto tra occupati nel settore terziario e occupazione totale, ha così superato il 70% nei paesi più avanzati, come gli Stati Uniti o la Gran Bretagna mentre si è assestato tra il 60% e il 70% in Italia e negli altri Paesi Europei. Nei servizi, i settori più tradizionali, come il commercio, gli alberghi e i pubblici esercizi, hanno mantenuto un peso rilevante: tra il 30 e il 35% nei diversi paesi industriali. Una percentuale poco inferiore al 30% è rappresentata dai servizi non destinabili alla vendita. La parte restante, tra il 30 e il 40%, è rappresentata dai servizi più avanzati e ad elevato valore aggiunto: credito e assicurazioni, comunicazioni, insegnamento e ricerca, servizi alle imprese (Enciclopedia Treccani - settore terziario).
Non esiste una definizione univoca e specifica di servizio, ma molteplici sono state le definizioni che sono state accostate nel tempo al concetto di servizio.
Il servizio è un bene intangibile, deteriorabile e non immagazzinabile che necessita di un sistema molto complesso di erogazione al quale partecipa il cliente (King 1992).
Il servizio è in ogni lavoro produttivo che non si concretizza in nessun genere di hardware (Ishikawa 1985).
Normann (1985) pone l'accento sull'importanza del cliente nella fase di erogazione del servizio, introducendo il concetto di "moment of truth". La maggior parte dei servizi sono il risultato di atti sociali, che si svolgono a diretto contatto tra il cliente e il fornitore del servizio. La qualità percepita dal cliente si realizza nel "moment of truth", dove il fornitore di servizi e il cliente si confrontano nell'arena. Al momento sono completamente soli. Ciò che accade in quel momento non può essere direttamente influenzato dall'azienda.
Rosander (1989) ribadisce il ruolo attivo del fattore umano, interessandosi maggiormente alla parte intangibile del servizio (la fiducia, la cortesia, la responsabilizzazione del front-line, ecc.) quale fattore determinante per il raggiungimento della soddisfazione del cliente.
La norma UNI EN ISO 8402 (1995) definisce il servizio come il " risultato di attività svolte all'interfaccia tra fornitore e cliente e di attività proprie del fornitore, per soddisfare le esigenze del cliente".
Il servizio è interpretato come un processo costituito da una sequenza logica di attività ben identificabili, osservabili, valutabili e misurabili (Barbarino, Leonardi, 1997).

1.2. CLASSIFICAZIONE DEI SERVIZI.

La prima classificazione che è possibile effettuare si basa sul grado di partecipazione del cliente al processo di fornitura (Gupta, Chen, 1995). I servizi possono essere classificati in:

- Servizi puri: un servizio si definisce puro quando il cliente deve essere presente durante l'erogazione del servizio (es. i ristoranti, il cinema, il sistema del trasporto aereo, ecc.).

- Servizi misti: essi prevedono sia il contatto diretto tra il cliente e il personale (front-office), sia il lavoro back office, invisibile al cliente.

- Servizi semi-manifatturieri: servizi in cui non esiste alcun tipo di contatto con la clientela. Tutto il lavoro viene svolto in maniera invisibile al cliente, e il modo con cui viene prodotto il servizio non influenza il modo in cui questo viene percepito dai clienti (es. bancomat, carte di credito e le compagnie telefoniche).

Una seconda classificazione può essere effettuata tra servizi pubblici e privati (Dai prodotti ai servizi-Fiorenzo Franceschini, 2000).

I servizi pubblici sono quelli elargiti dallo Stato o da altri Enti pubblici (statali, provinciali, regionali, ecc.) che non hanno scopo di lucro.
Vengono forniti gratuitamente, o a costi accessibili a tutti gli individui (ad esempio la scuola dell'obbligo, la giustizia, la sanità, ecc.).
In passato, la nozione di servizio pubblico è stata caratterizzata da una concezione soggettiva: era considerato servizio pubblico quello prestato da parte di un pubblico potere. Si è, in seguito, affermata una concezione oggettiva che, indipendentemente dalla natura del soggetto erogatore, riconosce il carattere di servizio pubblico in virtù del suo regime, dettato proprio per il soddisfacimento delle esigenze della collettività (Enciclopedia Treccani - Servizi pubblici).

I servizi privati possono essere suddivisi in due grandi categorie.
Alla prima categoria appartengono tutte le imprese che hanno come business principale la produzione di servizi rivolti ad altre imprese o privati. Alla seconda categoria appartengono le aziende produttrici di beni materiali che forniscono servizi a loro supporto (ad esempio servizi di assistenza post-vendita, le garanzie, i call center, ecc.).

Il servizio preso in analisi in questo lavoro di tesi è il servizio di trasporto aereo. Il trasposto aereo è un settore di importanza strategica per lo sviluppo economico e sociale perché contribuisce ad aumentare il traffico di passeggeri e merci e quindi la competitività di un Paese e garantisce ai cittadini il diritto alla mobilità in tutto il mondo (Ministero delle Infrastrutture e dei Trasporti).

Il trasporto aereo di merci e persone è regolato dall'Autorità di Regolazione dei Trasporti. La normativa europea disciplina i diritti dei passeggeri del trasporto aereo principalmente attraverso il regolamento (CE) n. 261/2004 del parlamento europeo e del consiglio dell'11 febbraio 2004, che istituisce regole comuni in materia di compensazione ed assistenza ai passeggeri in caso di negato imbarco, di cancellazione del volo o di ritardo prolungato.

1.3. CARATTERISTICHE DEI SERVIZI.

Tra prodotti e servizi si possono riscontrare innumerevoli analogie e differenze. Una sintesi delle principali analogie e differenze tra prodotti e servizi è stata effettuata da A.C. Rosander (1985) e H.C. Schwartz.
Un prodotto è tangibile, immagazzinabile e trasportabile, mentre un servizio è intangibile, non immagazzinabile e non trasportabile. Per un prodotto il cliente finale non partecipa o partecipa in piccola parte al processo produttivo che porta all'erogazione del prodotto e il processo di produzione ed erogazione sono separati, mentre nell'erogazione di un servizio l'acquisto di un servizio implica una prestazione immediata, i processi di produzione ed erogazione sono quasi sempre contemporanei e il cliente a differenza del mondo dei prodotti è parte integrante del processo produttivo.
Nell'ambito dei prodotti il progetto del prodotto è centrato sul cliente, mentre il progetto del processo è centrato sull'operatore, nei servizi sono entrambi focalizzati sul cliente.
Nei prodotti vi è poca variabilità a differenza dei servizi in cui l'erogazione del servizio presenta una fonte di elevata variabilità (ad esempio un volo aereo, il servizio è lo stesso mentre l'erogazione del servizio può essere percepita in modo diverso da cliente a cliente).
Nel mondo dei prodotti la produzione è indipendente dal consumo e vi è una grande facilità nell'applicazione di standard, misure, ispezioni e controlli a differenza dei servizi in cui l'erogazione e l'utilizzo del servizio sono simultanei vi è grande difficoltà nell'applicazione di standard, misure e controlli.
Nella produzione e vendita di un prodotto non è importante il rapporto tra operatore e cliente mentre per un servizio le relazioni tra operatore e cliente rappresentano una fonte di criticità e possono aumentare o diminuire la qualità e il valore del servizio fornito (ad esempio il rapporto tra hostess - clienti condiziona notevolmente il soddisfacimento o meno del cliente e quindi la qualità del servizio offerto).
Nei servizi le capacità interpersonali e relazionali sono più importanti di quelle tecniche e l'addestramento che viene fornito agli operatori che andranno ad erogare un servizio è di tipo psicoattitudinale piuttosto che tecnico questo perché la maggior parte degli operatori tratta direttamente con il cliente e quindi il rapporto operatore - cliente è alla base della qualità di un servizio.
Infine a differenza dei prodotti non è possibile applicare delle economie di scala significative ai servizi (Dai prodotti ai servizi - Fiorenzo Franceschini, 2000).

1.4. QUALITÀ NEI SERVIZI.

Nel 1985 Parasuraman, Zeithaml e Berry hanno ideato il primo modello concettuale per la descrizione della qualità nei servizi.
Per la formulazione empirica del modello, gli autori utilizzarono sia indagini su gruppi di clienti (focus groups), sia interviste con dirigenti di aziende di differenti comparti: servizi bancari per il pubblico, carte di credito, mediazioni di valori mobiliari, riparazioni e manutenzioni di apparecchi elettrodomestici.
I risultati dell'indagine empirica hanno permesso di predisporre un modello concettuale di validità generale, capace di evidenziare i legami e gli elementi condizionanti nel processo di erogazione della qualità nei servizi.
Per definire il modello sono state predisposte alcuni GAP che vengono trattati di seguito.
Il GAP 1 (scostamento tra qualità attesa e qualità ipotizzata): indica la differenza di percezione della qualità del servizio tra i manager di un'azienda che ipotizzano un livello di qualità del servizio (qualità ipotizzata) e la qualità che si aspetta di ricevere il cliente da un servizio (qualità attesa).
Il GAP 2 (scostamento tra qualità ipotizzata e qualità di progettazione): indica la differenza che si genera poiché i manager sanno quali sono i requisiti richiesti dai clienti ma essi trovano difficoltà nello stabilire specifiche operative adeguate a causa di vincoli progettuali e di mercato.
Il GAP 3 (scostamento tra qualità di progettazione e qualità realizzata): indica il divario che vi è tra la qualità che si presuppone di avere in fase di progettazione e la qualità che possiede il servizio e ciò è dovuto all'alta variabilità che caratterizza il mondo dei servizi.
Il GAP 4 (scostamento tra qualità realizzata e qualità di marketing): la pubblicità e i vari canali di comunicazione di un'azienda sono in grado di influenzare le aspettative del consumatore.
Lo scostamento tra la qualità realizzata e la qualità di marketing dipende principalmente da inadeguate comunicazioni tra le diverse funzioni aziendali e all'interno dei singoli reparti e comunicazioni esterne (pubblicità, promozioni, ecc.) tendenti a massimizzare la qualità offerta al cliente e/o focalizzate solo sulle componenti di reale eccellenza, oppure poco chiare, fuorvianti o inesatte.
Il Δ_{totale} (GAP 5 in figura…) = f (GAP 1, GAP 2, GAP 3, GAP 4): indica la qualità che l'utente finale percepisce del servizio ed è la differenza tra la qualità attesa e quella percepita ed è funzione dei GAP precedenti.

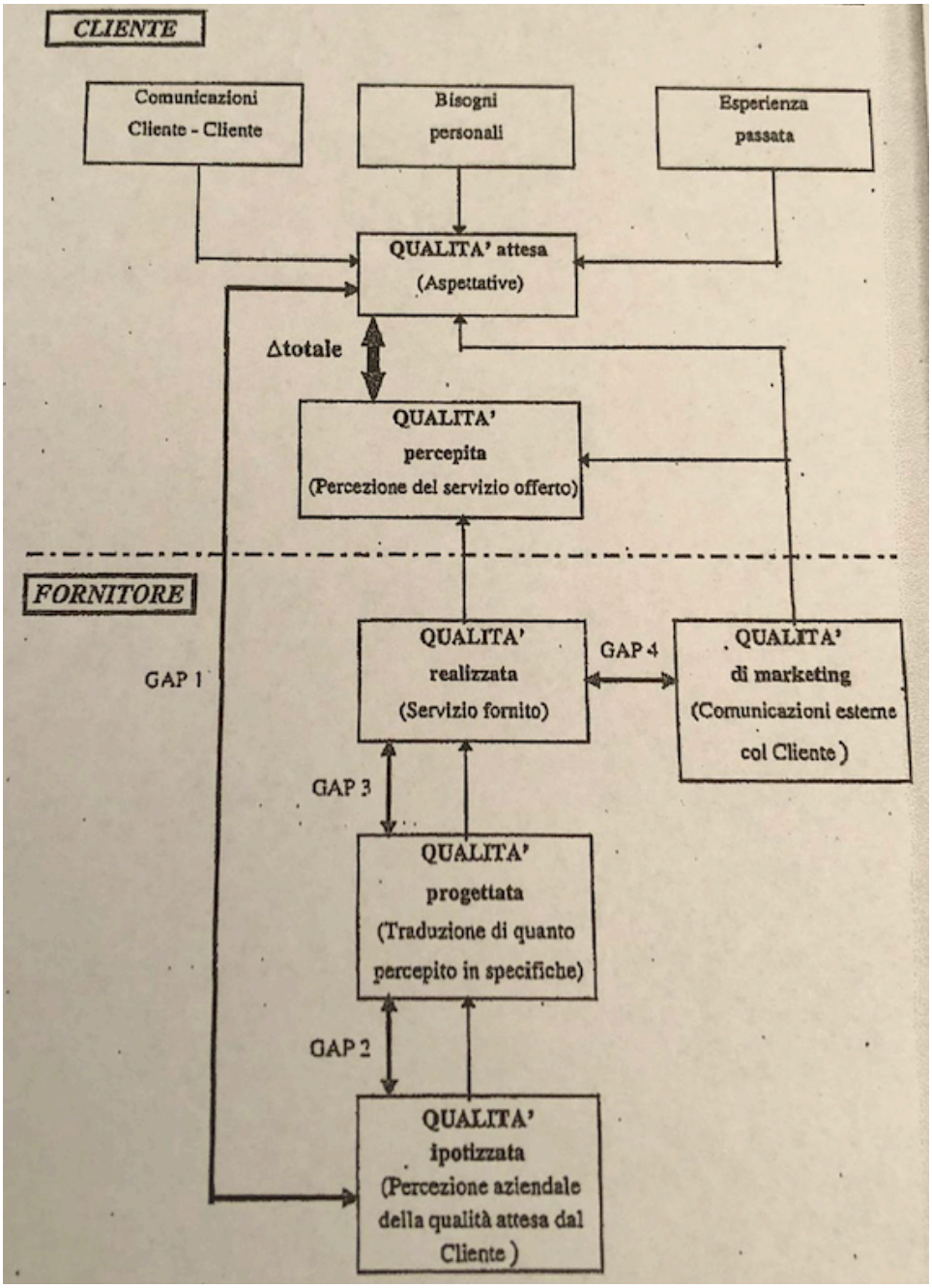


Figura 1. *Modello PZB.*

I focus groups hanno mostrato che, a prescindere dal servizio, gli utenti adottano criteri simili per valutare la qualità.
Questi criteri rientrano in 10 categorie denominate “fattori determinanti della qualità del servizio” e sono elencate e descritte in Appendice A.

Nel 1988, nasce il modello SERVQUAL, con l’intento di capitalizzare le conoscenze teoriche del modello PZB e fornire uno strumento tangibile per poter misurare la qualità di un servizio. SERVQUAL è un questionario a più voci (scala multi-item), pensato in modo da garantire la significatività delle informazioni raccolte, l’affidabilità dei dati e la capacità di indagare su tutte le componenti e le caratteristiche del servizio di riferimento.
Il primo passo della costruzione di SERVQUAL è la definizione della qualità del servizio come scostamento tra attese e percezioni.

In seguito si sono identificate le dieci determinanti che costituiscono il dominio del costrutto sulla qualità del servizio e la conseguente generazione di 97 items in rappresentanza delle dieci dimensioni.

Poi vi è la raccolta di dati su un campione di 200 utenti per 5 diversi tipi di servizi. Successivamente è stata effettuata la purificazione iterativa della scala con lo scopo di aumentare l'indicatore statistico α di Cronbach:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} \sigma_{Y_i}^2}{\sigma_X^2}\right)$$

dove k è il numero di items, σ^2_x è la varianza del punteggio totale e σ^2_{yi} è la varianza dell'item i-esimo per il campione di individui in esame.
L' α di Cronbach è un indicatore statistico utilizzato nei test psicometrici per misurarne l'attendibilità, ovvero per verificare la riproducibilità nel tempo, a parità di condizioni, dei risultati da essi forniti. In genere valori alti di attendibilità sono da considerarsi quelli che vanno da 0.70 in su.
Dopo la purificazione iterativa della scala, essa è costituita da 34 enunciati rappresentativi di 7 dimensioni. Per purificare ulteriormente la scala sono stati raccolti dati relativi alle attese e percezioni per un nuovo campione di 200 intervistati per ciascuna tipologia di servizio ed è stata valutata la scala con 34 enunciati.
Successivamente la scala è stata ulteriormente purificata identificando una scala a 22 enunciati per la misurazione della qualità del servizio e si è ottenuta la prima versione di SERVQUAL.
I 22 enunciati si riferiscono alle 5 nuove dimensioni della qualità da analizzare di un servizio: elementi tangibili, affidabilità, capacità di risposta, capacità di rassicurazione, empatia.
Infine gli ultimi due step sono stati la valutazione dell'affidabilità e della dimensionalità di SERVQUAL e il controllo della validità di SERVAQUAL.

Nel 1992 Cronin e Taylor riesaminarono alcune ipotesi introdotte da Parasuraman, Zeithaml e Berry nel modello SERVQUAL e svilupparono un nuovo strumento denominato SERVPERF (SERvice PERFormance).
L'idea principale alla base del nuovo modello è la netta distinzione tra la qualità del servizio e la soddisfazione del cliente.
L'analisi di Cronin e Taylor si incentra su tre punti:

- l'individuazione del modello più idoneo per effettuare una misura del costrutto della qualità di un servizio
- studio della sequenzialità dei concetti di soddisfazione del cliente e di qualità percepita
- effetto della soddisfazione del cliente e della qualità sulle intenzioni d'acquisto

1.5. QUALITÀ NEL TRASPORTO AEREO.

La qualità nel settore del trasporto aereo può essere valutata tramite 12 fattori di qualità del servizio presenti nella Carta della mobilità (ENAC-Ente Nazionale per l'Aviazione Civile):

- sicurezza del viaggio
- sicurezza personale e patrimoniale
- regolarità del servizio (e puntualità dei mezzi)
- pulizia e condizioni igieniche
- comfort del cliente
- servizi aggiuntivi
- servizi per viaggiatori con handicap
- informazione alla clientela
- aspetti relazionali e comportamentali
- servizi di sportello
- integrazione intermodale
- attenzione all'ambiente

Per ogni fattore sono individuati adatti indicatori di qualità. In corrispondenza ai singoli indicatori sono specificati i valori degli standard, da considerare come impegni benché non previsti da obblighi normativi. Il livello di percezione globale di ciascun fattore di qualità è misurato in termini di percentuale di persone soddisfatte dalla componente del servizio presa in esame.
Gli indicatori di qualità relativi ai fattori di qualità precedentemente elencati e relativi alle attività di gestione aeroportuale (settore passeggeri) sono riportati in Appendice B.
La Carta della mobilità prevede, inoltre, l'attivazione di sistemi di monitoraggio degli standard, dei fattori di percezione globale e dell'andamento della fenomenologia collegata al reclamo per tre principali tipologie: "lamentela per insoddisfazione, reclamo per inadempienza dell'azienda, richiesta di tutela per il riconoscimento dei diritti del cliente, oltre ai tempi medi di risposta".
La gestione dei reclami è un aspetto molto importante per gestire al meglio il rapporto con il cliente ed eventuali disguidi avuti da esso con la compagnia aerea.
La gestione dei reclami è parte integrante della Carta dei servizi. Coerentemente a quanto prescritto dalle norme UNI 10600, i soggetti erogatori predispongono un apposito modello per la redazione del reclamo (completo dei recapiti e delle informazioni utili a tal fine), specificando tempi e modalità di risposta (entro 30 gg. è prescritto un riscontro scritto) e precisando le casistiche che danno diritto a forme di risarcimento, con le informazioni correlate (coperture assicurative, entità dei risarcimenti, procedure, ecc.).
Il soggetto erogatore del servizio si occupa della raccolta dei dati relativi ai reclami ricevuti, classificati per tipologie (lamentela per insoddisfazione, reclamo per inadempienza dell'azienda, richiesta di tutela per il riconoscimento dei diritti del cliente) ed elaborati al fine di individuare gli indici più significativi (tasso di reclamo, tempo medio di risposta, tempo medio di soluzione delle controversie, confronto con i due periodi precedenti, ecc.), da mettere a disposizione dei passeggeri e dell'utenza in generale.
La raccolta e l'elaborazione dei reclami sono oggetto d'esame da parte del gestore aeroportuale e del Comitato per la regolarità e la qualità dei servizi, in quanto importante fonte di informazione per l'identificazione qualitativa e quantitativa delle cause dei disservizi e dei settori più critici tra quelli monitorati. I resoconti sui reclami ricevuti vengono resi disponibili, su richiesta, ai soggetti interessati.

1.6. TRASPORTO AEREO.

Il trasporto aereo consiste nel trasferire persone e merci da un punto ad un altro utilizzando un aereo ed è un fattore essenziale della globalizzazione economica e del progresso sociale.
In Europa in base ad un'analisi del Corriere del 2018 su dati aziendali (figura 2) il Gruppo Lufthansa detiene una quota di mercato del 13,4 % seguita da Ryanair con l'11,2 %, IAG con il 10,5% e Air France con l'8,1%.

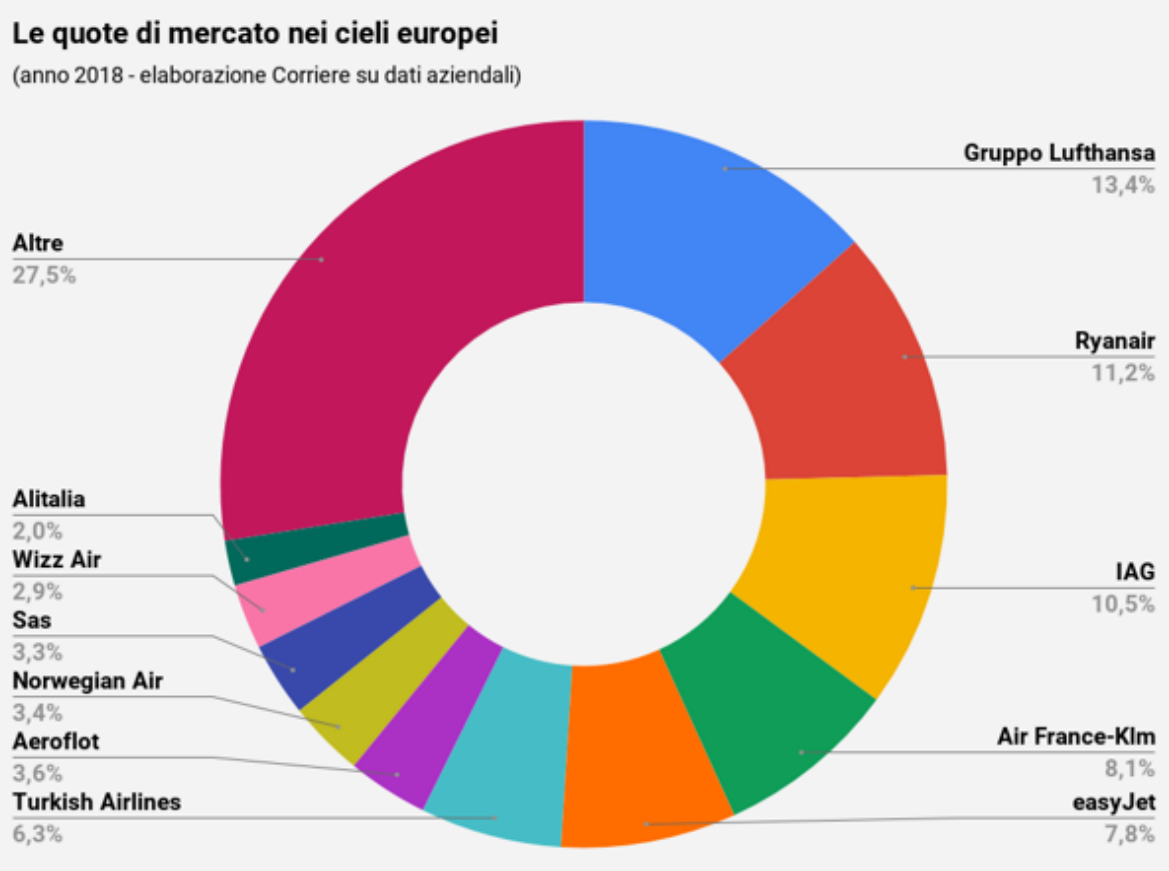


Figura 2. *Quote di mercato nel trasporto aereo delle compagnie aeree in Europa (2018)*

Nel settore low cost un'analisi di mercato dell'ottobre 2018 su dati Oag (figura 3) ha evidenziato che la compagnia che detiene una quota di mercato considerevole è Southwest Airlines con l'11,5% seguita da Ryanair con l'8,9%, Easyjet con il 6,6% e IndiGo con il 4,5%. Ryanair e Easyjet, tra le compagnie oggetto del campione della tesi, sono tra i principali players a livello mondiale nel settore low cost superate solo dal colosso americano Southwest Airlines e sono i due principali players a livello europeo.

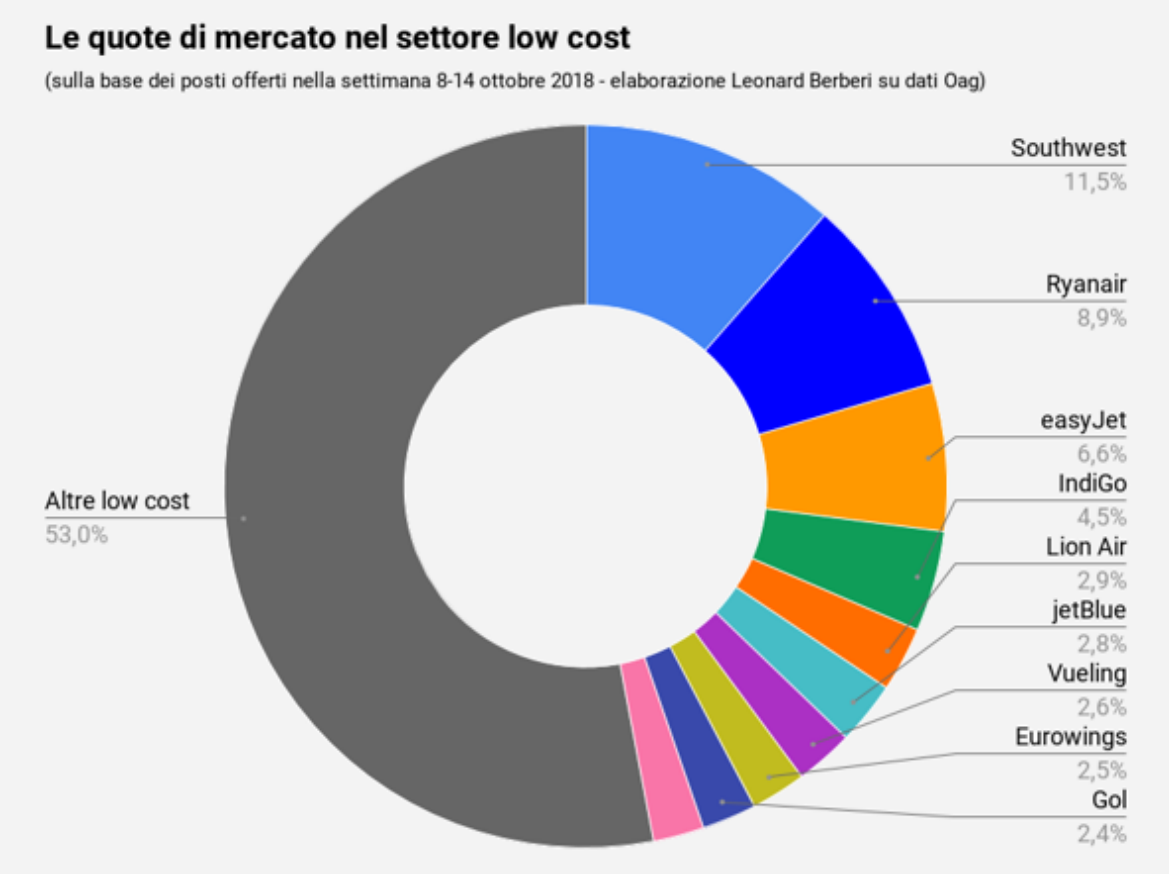


Figura 3. *Quote di mercato nel trasporto aereo nel settore low cost*

Nel presente lavoro di tesi sono state analizzate le principali compagnie europee, i due principali players europei del settore low cost e i “superconnectors” (Emirates, Qatar Airways e Etihad Airways) che sono tra i principali players nel trasporto aereo nel Medio Oriente.
Un’analisi di mercato sulle quote di mercato detenute dalle compagnie aeree (Statista, 2019) ha evidenziato che Emirates detiene 8,7% del mercato del Medio Oriente e si colloca tra i principali tre palyers del trasporto aereo di quest’area.
Etihad Airways possiede una quota di mercato pari al 3,5 % e Qatar Airways una quota pari al 3% e si collocano nei dieci principali players del trasporto aereo di passeggeri del dell’area medio orientale.

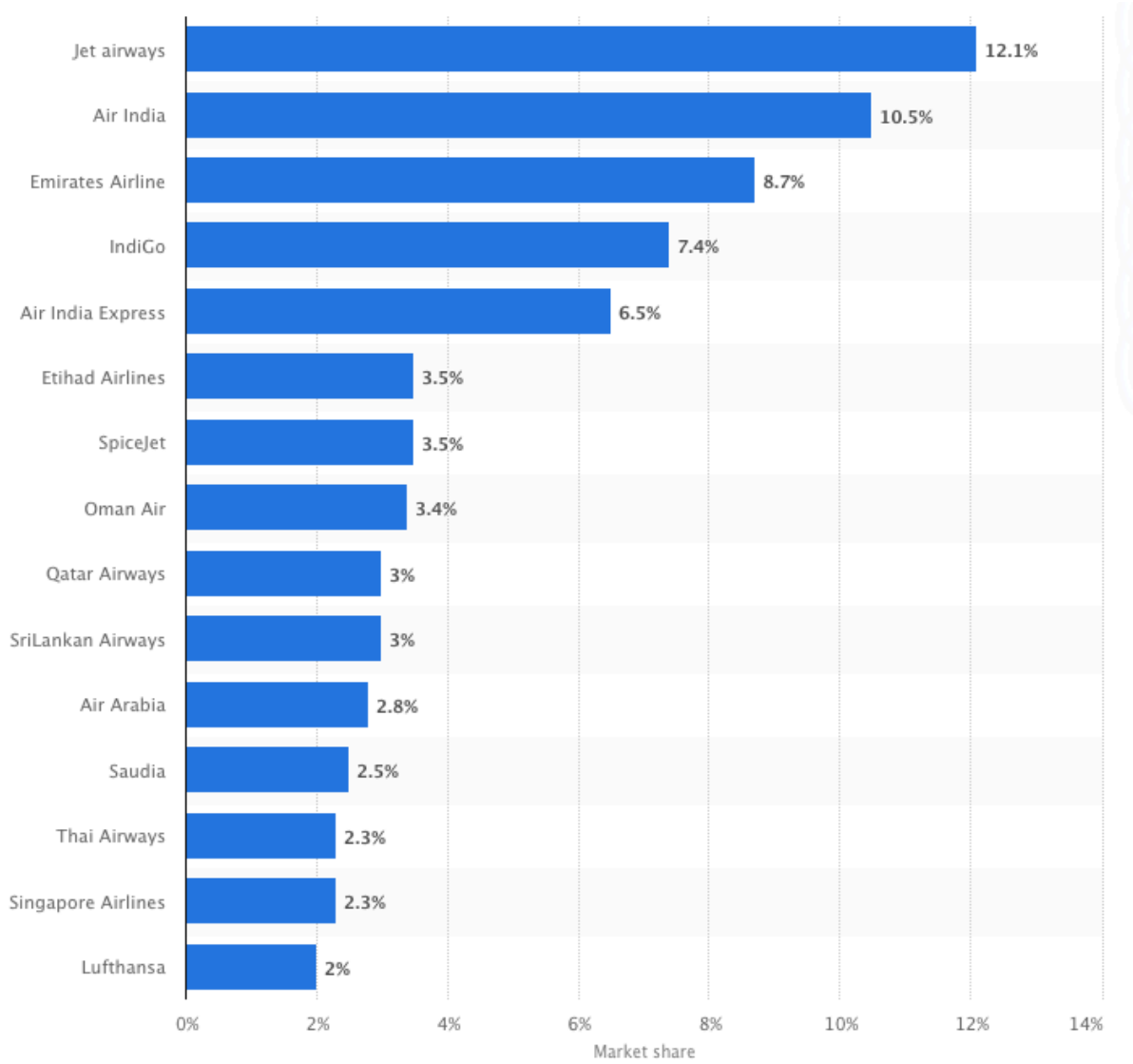


Figura 4. *Quote di mercato delle principali compagnie aeree dell'India e del Medio Oriente (2019, Statista).*

CAPITOLO 2

2.1. USER GENERATED CONTENT (UGC).

Gli user generated contents, o UGC, sono contenuti di varia natura (video, testi, immagini, audio, ecc.) che sono "postati" e liberamente condivisi dagli utenti su blog, forum, social media, ecc.
La principale caratteristica degli UGC è l'essere resi disponibili e liberamente accessibili ad altri utenti.
L'originalità e la creatività sono le altre due caratteristiche che contraddistinguono gli user generated contents poiché chi rilascia questi contenuti aggiunge sempre qualcosa di nuovo rilasciando un contenuto su un forum o un blog.

Gli UGC sono il "postulato della remix colture" (Lessing, 2009) e la "cultura convergente" (Jenkins, 2006).

Gli UGC sono "l'esempio più vivido dell'autocomunicazione di massa" (Castells, 2014) che ha smantellato il modello broadcast e unidirezionale dell'industria mediatica.

Gli UGC possono portare benefici e ritorni concreti. Gli UGC sono impiegati ormai all'interno di strategie aziendali di content marketing o di piani editoriali di testate ben rinomate e, in casi come questi, sono previste speciali policy per la retribuzione del creativo. In qualche altro caso chi crea video, testi, immagini e li condivide in rete è, in realtà, un professionista che intende sfruttare il Web alla ricerca di visibilità e di possibili guadagni.
Gli User Generated Contents hanno creato delle figure ibride, quelle dei pro-am, che sono professionisti che si dilettano nella produzione di contenuti gratuiti, anche qualitativamente inferiori a quelli per cui si farebbero pagare e che lasciano liberamente a disposizione degli altri utenti. Per la maggior parte degli utenti, creare un "meme" e renderlo virale in rete, scrivere una guida TripAdvisor o correggere le imperfezioni di una voce Wikipedia sono atti di «generosità digitale»: è per questo che milioni e milioni di utenti creano "surplus cognitivo" (Shirky, 2010) impiegando il loro tempo libero in attività in rete come queste che, sebbene non prevedano una retribuzione monetaria, hanno un ritorno in termini di appartenenza, di affiliazione, di comunità, di sentire di aver contribuito con poco e a costo (quasi) nullo ad una causa e ottenere un ritorno, in alcuni casi, in termini di popolarità.
Alcuni UGC presentano un contenuto di veridicità e affidabilità maggiore rispetto ad altri, che possono essere fuorvianti, discriminatori o non pertinenti.
Secondo uno studio di Olapic (piattaforma di Visual Marketing) realizzato su un campione di 4.500 internauti (utenti abituali di Internet) compresi tra 16 e 49 anni e residenti negli Stati Uniti e nei principali paesi europei, viene accordata maggior fiducia ai contenuti pubblicati dagli utenti (Figura 5):

- Immagini UGC (52%);
- Video UGC (27%);
- UGC testuali (12%);
- Classici messaggi pubblicitari (5%).

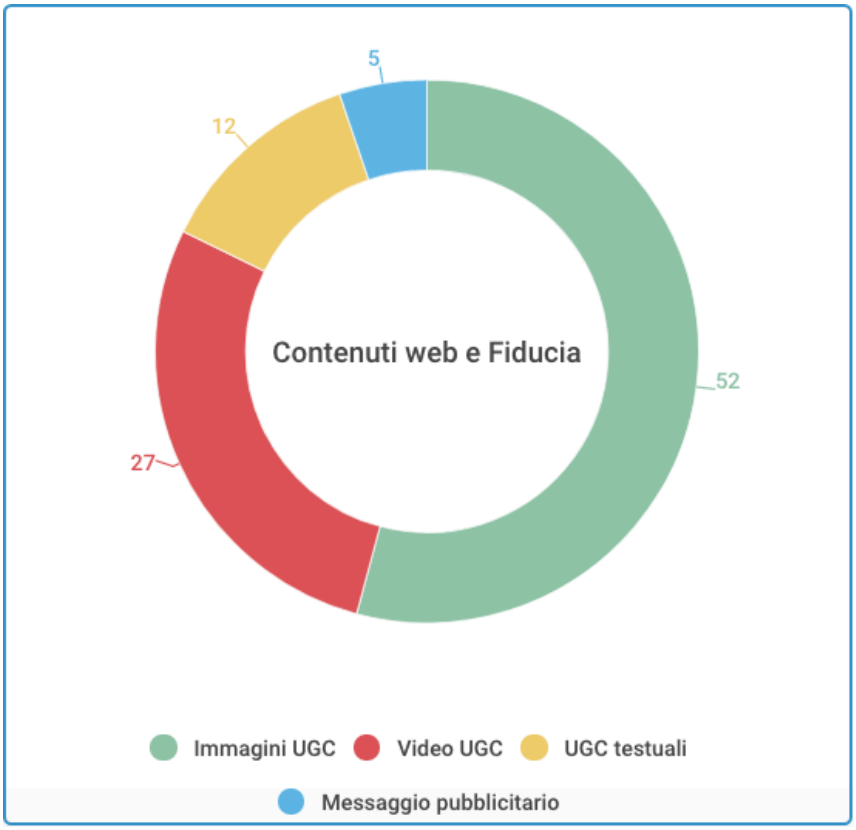


Figura 5. *Studio di Olapic sugli UGC.*

Il 40% degli intervistati ha affermato di aver prodotto contenuti "taggando" il proprio brand preferito, nel 34% dei casi lo ha fatto perché ha giudicato positiva l'esperienza di un brand. In tutto questo, solo il 14% degli intervistati ha detto di aver utilizzato "hashtag" proposti direttamente da un brand.
Il 76% ha affermato che gli User Generated Content sono "più onesti" della comunicazione tradizionale (contenuti prodotti dalle campagne di marketing delle aziende). Una percentuale variabile tra il 53 e il 70% (a seconda del paese di residenza) ritiene che gli UGC fatti bene possano aumentare la predisposizione all'acquisto di un prodotto o servizio descritto o raccontato.

Nel web è possibile trovare tre tipi di recensioni: le recensioni positive, le recensioni negative e le false recensioni.
Le recensioni positive infondono un senso di fiducia e tranquillità nell'utente che sta per acquistare un determinato prodotto o servizio.
Un'indagine sulle recensioni online di Review Trackers del 2018 ha rilevato che le recensioni negative hanno convinto il 94% dei consumatori intervistati a evitare un particolare servizio. Più sorprendentemente, l'85% dei consumatori nel sondaggio si è fidato delle recensioni online tanto quanto delle raccomandazioni di amici e parenti, il che suggerisce che le recensioni online esercitano un'influenza simile a quella dei 'referral' personali. Anche se il 68% dei consumatori si fida maggiormente delle recensioni se vede che sono sia positive che negative, sono le recensioni negative quelle maggiormente da temere in quanto sono quelle che più facilmente si diffondono tramite il passaparola.
Nonostante l'effetto passaparola, le recensioni negative ispirano maggiore credibilità nell'utente che le legge e aumentano la fiducia dei clienti, infatti gli utenti danno maggiore peso ad un servizio che ha sia recensioni positive che negative piuttosto che uno che possiede solo recensioni positive o solo recensioni negative.

Infine vi sono le false recensioni, ossia recensioni il cui contenuto è errato, fuorviante o discriminatorio.
Oltre a essere scorretto, sollecitare recensioni false da parte di utenti fittizi viola i termini di servizio dei siti di recensioni. Alcuni siti multano questo comportamento e i consumatori se ne accorgono ancor prima del sito stesso.

La sentiment analysis si basa sugli UGC per cercare di capire le percezioni di un utente su un servizio, che possono essere espresse tramite un UGC in maniera positiva o negativa.

Nel settore del trasporto aereo le recensioni degli utenti possono influenzare il comportamento di altri utenti nella scelta della compagnia aerea da prendere per effettuare un viaggio anche se i principali driver nella prenotazione di un volo continuano a rimanere il prezzo, la possibilità di portare più bagagli ed eventuali comfort a bordo.
Il processo di condivisione di UGC nel settore del trasporto aereo è descritto nello schema seguente.

Attraverso la condivisione della propria opinione tramite una recensione su un blog (Booking, TripAdvisor, Expedia, ecc.) su compagnie aeree e aeroporti si aiuteranno gli altri utenti a prendere una decisione migliore sulla scelta del volo e della compagnia aerea da scegliere per compiere un viaggio.
L'analisi delle recensioni rilasciate in rete ha un vantaggio sia per i clienti che per le compagnie aeree. I clienti possono trarre beneficio dalle recensioni di altri utenti per compiere la scelta migliore sul volo da prendere mentre le compagnie aeree tramite l'analisi degli UGC riferiti ad esse possono capire quali sono i punti di forza apprezzati dai clienti e quali le criticità che devono essere migliorate.

2.2. ALGORITMO STRUCTURAL TOPIC MODEL (STM).

Nell'ultimo decennio i modelli probabilistici sull'analisi dei topic, come l'algoritmo Latent Dirichlet Allocation (LDA), sono diventati uno strumento comune per la comprensione dei testi.
Sebbene originariamente sviluppati per scopi descrittivi ed esplorativi, i ricercatori vedono sempre più il valore dei modelli di analisi dei topic come uno strumento per la misurazione delle variabili linguistiche, politiche e psicologiche latenti presenti all'interno di testi (recensioni, commenti, articoli, ecc.).
L'elemento caratterizzante di questo lavoro è la presenza di informazioni aggiuntive a livello documentale (es. autore, rating, data, ecc.) sulle quali la variazione della prevalenza o del contenuto attuale è di interesse reticolare. Questo comporta generalmente l'esecuzione di un'implementazione standard dell'LDA e poi l'esecuzione di una valutazione post-hoc della variazione con una covariata di interesse.
Sono stati sviluppati numerosi casi speciali di questo quadro di riferimento per particolari tipi di struttura del corpus che riguardano sia la prevalenza dell'argomento (ad esempio, il tempo, l'autore, ecc.) che i contenuti attuali (ad esempio, l'ideologia, la geografia, ecc.).
Gli utenti sono stati lenti ad adottare questi modelli perché spesso è difficile trovare un modello che si adatti esattamente al corpus del testo di riferimento.
Si è sviluppato il modello STM che accoglie la struttura del corpus del testo attraverso covariate a livello documentale che influenzano la prevalenza e/o il contenuto dell'argomento.
Il modello generalizza diversi approcci esistenti in letteratura e permette agli utenti di incorporare la struttura specifica del corpus da analizzare senza sviluppare nuovi modelli da zero.
Il modello STM (figura 6) combina ed estende tre modelli esistenti: il correlated topic model (CTM), il Dirichlet - Multinomial Regression topic model (DMR) e lo Sparse Additive Generative topic model (SAGE).
Il precedente modello logistico sulla prevalenza topica è sostituito da un modello lineare logistico – normale. La distribuzione delle parole è ricollocata con una multinomiale tale che la distribuzione di un token è la combinazione di tre effetti (topic, covariate, interazione con la covariata dell'argomento).

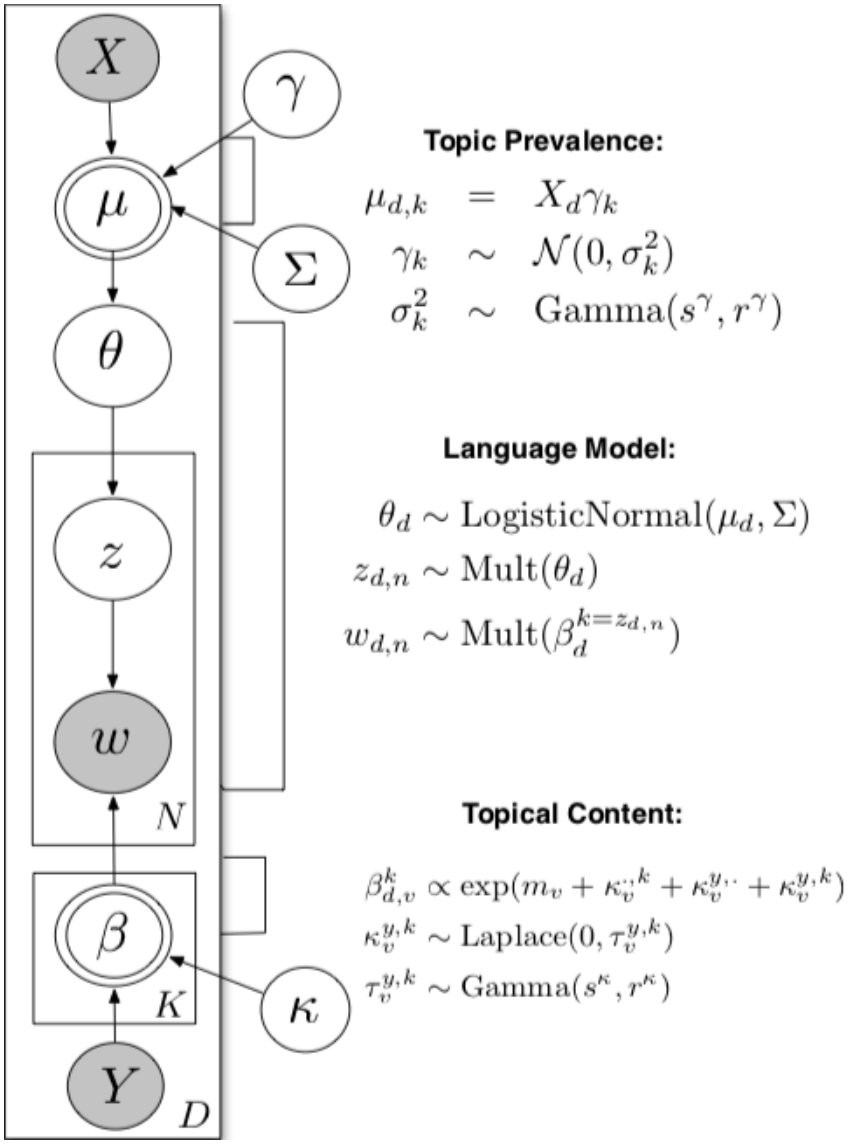


Figura 6. *Schema del modello STM.*

In STM si tratta una frase come un'unità di struttura di base, e si assume che tutte le parole di una frase condividano lo stesso aspetto attuale. Inoltre, si assume che due segmenti adiacenti siano altamente correlati; nello specifico, STM pone una forte dipendenza transazionale tra gli argomenti: la scelta dell'argomento per ogni frase si basa direttamente sull'assegnazione dell'argomento alla frase precedente, cioè sulla proprietà di Markov del primo ordine. Prendendo le intuizioni di HMM-LDA (Hidden Markov Model su LDA) che non tutte le parole veicolano contenuti (alcune di esse possono essere solo il risultato di un'esigenza sintattica), si introduce un argomento funzionale fittizio z_B per ogni frase del documento. Si utilizza questo argomento funzionale per catturare la distribuzione di parole indipendente dal documento, cioè lo sfondo del corpus del testo. Di conseguenza, in STM, ogni frase viene trattata come un mix di contenuto e argomenti funzionali.
Formalmente, si assume che un corpus sia costituito da D documenti con un vocabolario di dimensione V, e che ci siano k argomenti di contenuto incorporati nel corpus. In un dato documento d, ci sono delle frasi m e ogni frase i ha delle parole N_i. Si suppone che la probabilità di transizione dell'argomento p(z|z′) sia ricavata da una distribuzione multinomiale Mul(z′), e che la probabilità di emissione della parola sotto ogni argomento p(w|z) sia ricavata da una distribuzione multinomiale Mul($a_{z'}$).

Per ottenere una descrizione unificata del processo di generazione, viene aggiunto un altro argomento fittizio T-START in STM, che è l'argomento iniziale con la posizione "-1" per ogni documento ma non emette alcuna parola. Inoltre, poiché si presume che l'argomento funzionale si presenti in tutte le frasi, non si ha bisogno di modellarne la transizione con altri argomenti di contenuto. Si utilizza una variabile binomiale per controllare la proporzione tra contenuto e argomenti funzionali in ogni frase. Pertanto, ci sono transizioni di k+1 topic, una per ogni T-START e altre per k topic di contenuto; e k probabilità di emissione per gli argomenti di contenuto, con una ulteriore per l'argomento funzionale z_B (in totale k+1 distribuzioni di probabilità di emissione).
Condizionato sui parametri del modello θ = (α, β , π), il processo generativo di un documento in STM può essere descritto come segue:

- Per ogni frase s_i nel documento d:

(a) Creare il topic z_i dalla distribuzione multinomiale, condizionata alla frase precedente s_{i-1}:

$$z_i \sim Mul(\alpha_{z_{i-1}})$$

(b) Studiare ogni parola w_{ij} nella frase s_i dall'insieme del contenuto del topic z_i e del topic funzionale z_B:

$$w_{ij} \sim \pi p(w_{ij}|\beta, z_i)+(1-\pi)p(w_{ij}|\beta, z_B)$$

La probabilità congiunta di frasi e argomenti in un documento definito da STM è quindi data da:

$$p(S_0, S_1, \ldots, S_m, \mathbf{z}|\alpha, \beta, \pi) = \prod_{i=1}^{m} p(z_i|\alpha, z_{i-1})p(S_i|z_i)$$

dove, la probabilità di emissione dell'argomento z_i nella frase S_i è definito come:

$$p(S_i|z_i) = \prod_{j=0}^{N_i} \left[\pi p(w_{ij}|\beta, z_i) + (1-\pi)p(w_{ij}|\beta, z_B)\right]$$

Il processo è graficamente illustrato nella figura 7.

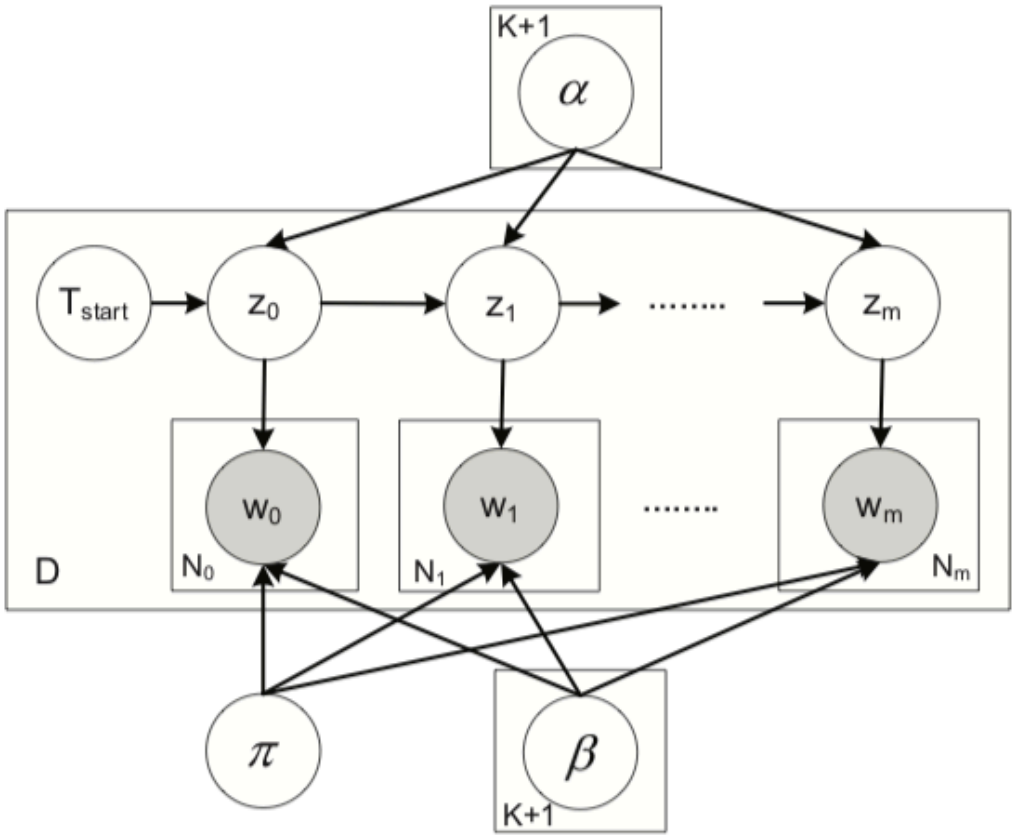


Figura 7. *Rappresentazione grafica di come opera il modello STM.*

Dalla definizione di STM, si può notare che la struttura del documento è caratterizzata da una catena di argomenti specifici del documento, e racchiudere le parole in un'unica frase a condividere lo stesso argomento di contenuto garantisce la coesione semantica degli argomenti. Anche se non si modella direttamente il mix di argomenti per ogni documento come fanno i topic models tradizionali, i modelli di raccolta delle parole all'interno dello stesso documento sono segnati dalla propagazione degli argomenti attraverso delle transizioni. Questo può essere facilmente compreso quando si scrive la probabilità posteriore dell'assegnazione dell'argomento per una particolare frase:

$$
\begin{aligned}
& p(z_i|S_0, S_1, \ldots, S_m, \Theta) \\
=& \frac{p(S_0, S_1, \ldots, S_m|z_i, \Theta)p(z_i)}{p(S_0, S_1, \ldots, S_m)} \\
\propto\ & p(S_0, S_1, \ldots, S_i, z_i) \times p(S_{i+1}, S_{i+2}, \ldots, S_m|z_i) \\
=& \sum_{z_{i-1}} p(S_0, \ldots, S_{i-1}, z_{i-1})p(z_i|z_{i-1})p(S_i|z_i) \\
& \times \sum_{z_{i+1}} p(S_{i+1}, \ldots, S_m|z_{i+1})p(z_{i+1}|z_i)
\end{aligned}
$$

La prima parte dell'equazione precedente descrive l'influsso ricorsivo nella scelta dell'argomento per l'i-esima frase delle frasi precedenti, mentre la seconda parte descrive come le frasi successive influenzano l'attuale assegnazione dell'argomento. Intuitivamente, quando si deve decidere l'argomento di una frase, si osserva "indietro" e "avanti" su tutte le frasi del documento per determinarne una "adatta". Inoltre, a causa della proprietà di Markov del primo ordine, la dipendenza topica locale diventa più accentuata, cioè interagiscono direttamente attraverso la prova di transizione p (zi | z_{i-1}) e p (z_{i+1} | z_i). E tale interazione su frasi più lontane verrebbe smorzata dalla moltiplicazione di tali probabilità. Questo risultato è ragionevole, soprattutto in un documento lungo, dato che le frasi vicine hanno più probabilità di coprire argomenti simili rispetto a due frasi distanti tra loro.

2.3. PROCESSO DI ESTRAZIONE DATI.

Il primo passo della metodologia proposta consiste nella raccolta di recensioni e UGC (User Generated Content) dai social media (Facebook, Trip Advisor, Twitter, ecc.) e altri aggregatori di recensioni con l'obiettivo di creare un dataset che verrà analizzato tramite la Text Mining Analysis.
La tecnica utilizzata per raccogliere le recensioni che andranno a costituire il dataset viene definita Web Scraping.
Il web scraping (dall'inglese "to scrape" che significa "grattare", "raschiare", "racimolare") è una tecnica informatica di estrazione di dati da un sito web per mezzo di programmi software.
Di solito, tali programmi simulano la navigazione nel World Wide Web utilizzando l'Hypertext Transfer Protocol (HTTP) o attraverso browser, come Internet Explorer o Mozilla Firefox. I programmi che permettono di fare web scraping effettuano lo scraping tramite l'utilizzo di un browser programmabile, detto in gergo "headless browser" (letteralmente browser senza testa). Headless significa che è in grado di navigare come un comune browser, ma, non avendo un'interfaccia grafica, non mostrerà nulla all'utente dei dati elaborati.
Il vantaggio di questa tecnica è che per il sito web non c'è differenza tra un utente umano e un bot. Il principale scopo dello scraping è l'estrapolazione delle informazioni dal corpus di un testo, disponibile sulla rete internet. I dati sono estratti, elaborati e archiviati in un database (Figura 8).

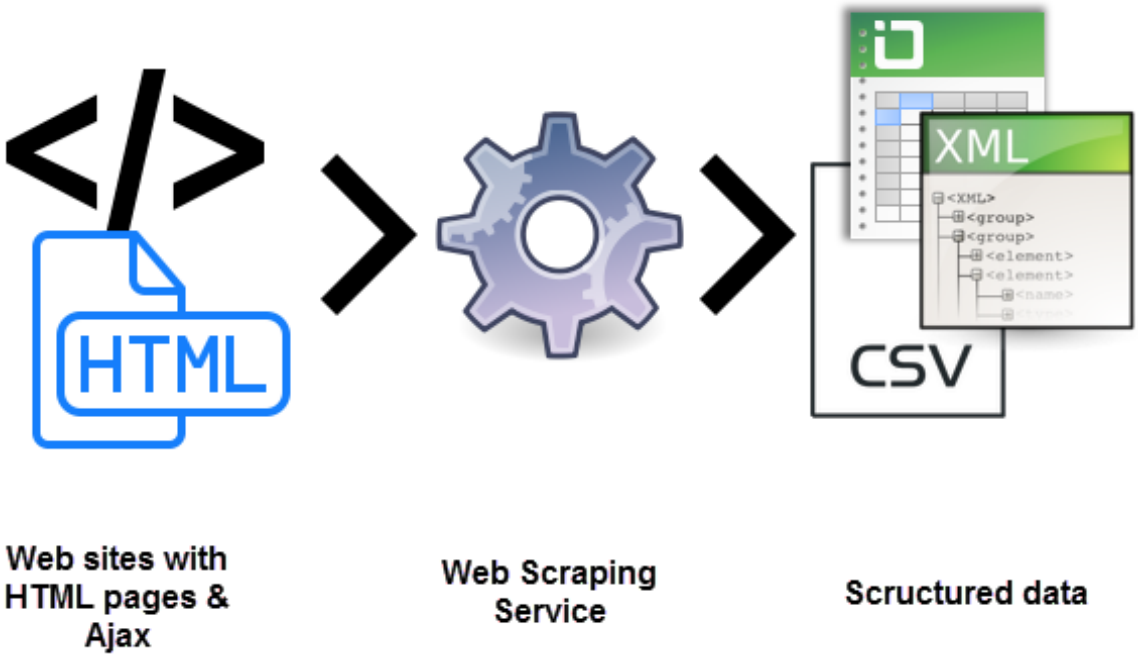


Figura 8. *Tecnica del Web Scraping.*

E' possibile fare scraping sui vari social e siti web tramite l'utilizzo di sofware gratuiti o a pagamento oppure attraverso la scrittura di algoritmi di programmazione.
Esistono molti software che permettono di fare scraping tra cui: Data Toolbar, Octoparse, Parsehub, Web Scraper.io, Google Spreadsheets, ecc.
Tutti questi software forniscono lo stesso servizio, ossia permettono di fare scraping ma hanno funzionalità e caratteristiche differenti.
Data ToolBar è uno strumento intuitivo di web scraping che automatizza il processo di estrazione dei dati web. Occorre puntare ai campi di dati che si desidera raccogliere e lo strumento compie l'operazione. Lo strumento dati è progettato per gli utenti aziendali di tutti i giorni e non richiede competenze tecniche.
Octoparse è uno strumento di scraping potente ed efficace che permette di estrarre diverse tipologie di dati da sorgenti online. Grazie ad un'interfaccia semplice e visuale è possibile configurare il tool in pochi passi ed impostare l'architettura di estrazione senza dover scrivere una singola riga di codice.
Parsehub è un software desktop disponibile per Windows, Mac e Linux dotato di caratteristiche molto avanzate tra cui la possibilità di sfruttare diversi IP (per evitare blocchi da parte del server), l'integrazione con sistemi di archiviazione (come Dropbox) e la scansione di siti realizzati con tecnologie come Javascript e Ajax (difficili da scansionare da altri strumenti).
Nel seguente lavoro si è utilizzato per la raccolta dati lo strumento Data Toolbar e un algoritmo di raccolta dati scritto in Python.

2.3.1 DATA TOOLBAR

Il 20% del totale del dataset (circa 4000 recensioni) è stato costituito tramite l'utilizzo del software Data Toolbar.

Data Toolbar è un software che “raschia”, raccoglie e converte i dati strutturati dalle pagine Web in un formato tabulare che può essere caricato in un foglio di calcolo o in un programma di gestione del database.
Dopo aver scaricato il software bisogna installarlo e avviarlo. Dopo aver avviato il programma, vi sarà una finestra del browser con il pulsante Data Tool (Figura 9) nell'angolo in alto a sinistra del browser. Per impostazione predefinita, il programma apre il browser Google Chrome ma è possibile passare a Firefox utilizzando il menu a discesa.

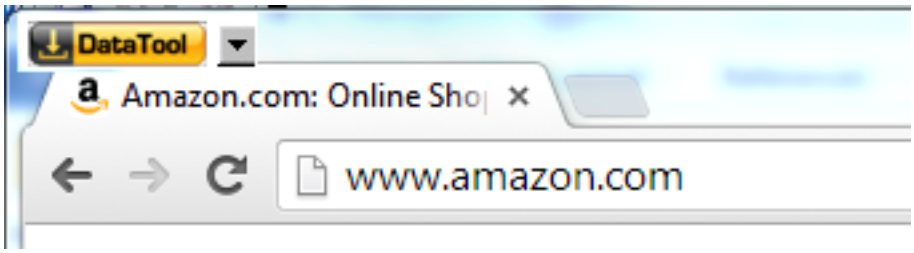


Figura 9. *Pulsante Data Tool.*

Cliccando sul pulsante Data Tool il software creerà un progetto predefinito associato al dominio di destinazione.
Ogni progetto richiede un URL di avvio. L'URL di inizio punta alla pagina web dove il progetto navigherà all'inizio. A volte l'URL di partenza è l'URL principale del sito web, ma spesso i dati richiesti si trovano in una sotto-pagina. Alcuni siti web consentono la navigazione o il reindirizzamento senza modificare l'URL visibile. In questi casi, non si ha un URL di partenza che punta direttamente alla pagina web di partenza preferita, ma si dovranno aggiungere delle funzioni al progetto di estrazione dati per navigare verso quella pagina web. La posizione predefinita del file di progetto dipende dal dominio URL iniziale. Per impostazione predefinita, il progettista del progetto salva il progetto utilizzando il nome del dominio di destinazione. Se necessario, il progetto del dominio predefinito può essere salvato con un nome personalizzato utilizzando la schermata “Proprietà del progetto”. I progetti personalizzati possono essere aperti utilizzando il menu a discesa. Oltre al progetto, il programma salva l'immagine dell'inizio e i cookie dell'ultima sessione.
Quando il progetto è aperto, tutta la navigazione e l'interazione con il browser deve avvenire solo attraverso il designer. Il designer emulerà tutti gli eventi del mouse e della tastiera in base agli input dell’utente. In modalità di progettazione, spostando il puntatore del mouse sulla pagina web, il designer evidenzia automaticamente gli elementi della pagina che possono essere contrassegnati come campi dati (Field). Cliccando su un campo di testo, un link o un'immagine si crea automaticamente una nuova riga nella griglia dati. Il clic su un link non provoca la navigazione. Selezionando nuovi campi, vengono create automaticamente altre colonne. È possibile rimuovere i contenuti aggiunti o modificarne le proprietà. Se per qualche motivo si ha bisogno di interagire direttamente con il browser, bisogna impostare la modalità di selezione su “Off”.
Quando il primo elemento di immissione viene aggiunto al modello, il programma assegna automaticamente un'azione all'elemento. È sempre possibile rimuovere l'azione o aggiungerla ad un altro elemento. Le informazioni provenienti da elementi con azione o elementi di input allegati non vengono raccolte.
Nella schermata principale di Data Toolbar ci sono cinque colonne (Figura 10). L'icona più a sinistra mostra il tipo di elemento. Cliccando sull'icona si apre la finestra delle proprietà degli elementi. La colonna "Field" è un nome modificabile. La colonna "Data" mostra le informazioni raccolte o inserite.

La colonna "Action" mostra se c'è un'azione allegata. Nella maggior parte dei casi, un'azione è un equivalente del clic del mouse sull'elemento web. Cliccando sul pulsante dell'azione si apre un nuovo modello. Le azioni vengono automaticamente allegate all'elemento di invio quando l'elemento viene selezionato. È possibile attaccare o staccare manualmente un'azione a qualsiasi elemento, ma il più delle volte si tratta di un elemento di collegamento. L'ultima colonna della griglia è il pulsante di cancellazione della riga.

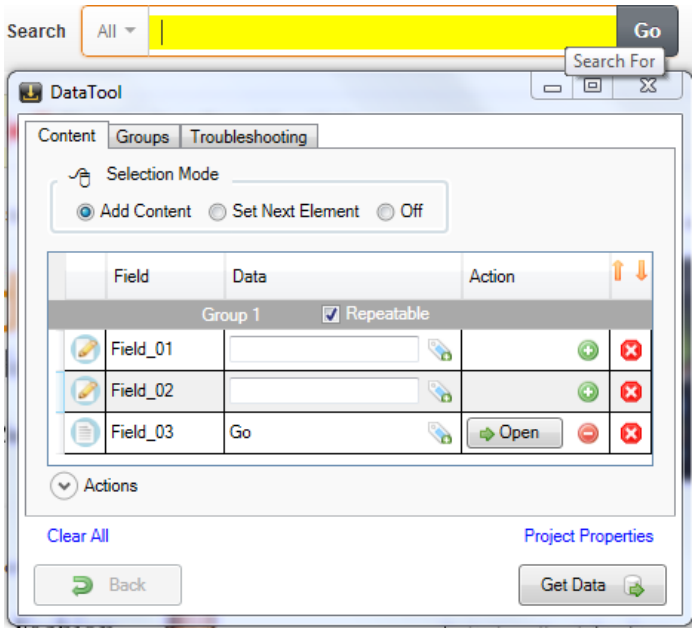


Figura 10. *Schermata principale di Data Toolbar*

Dopo aver selezionato dalla pagina web i vari elementi (testo, immagini, ecc.) che si vogliono acquisire bisogna cliccare sul pulsante “Get Data” (Figura 10) che permette di ottenere dalla pagina web i dati richiesti.
Data Toolbar permette di fare il “Paging” ossia di raccogliere i dati scorrendo le pagine successive di un sito web o di un social network.
I risultati della ricerca hanno di solito un link alla pagina successiva che apre la pagina successiva nel set di navigazione. In questo caso, è necessario impostare la modalità di selezione su “Set Next Element” (Figura 11) e selezionare il link “Next Page”(Figura 12) nella pagina web.

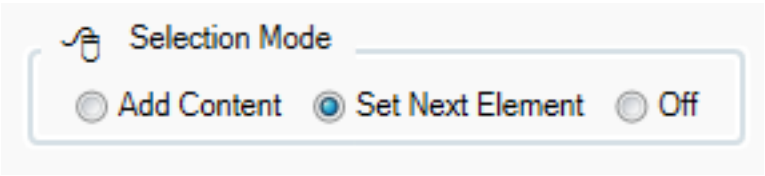


Figura 11. *Modalità Set Next Element in Data Toolbar*

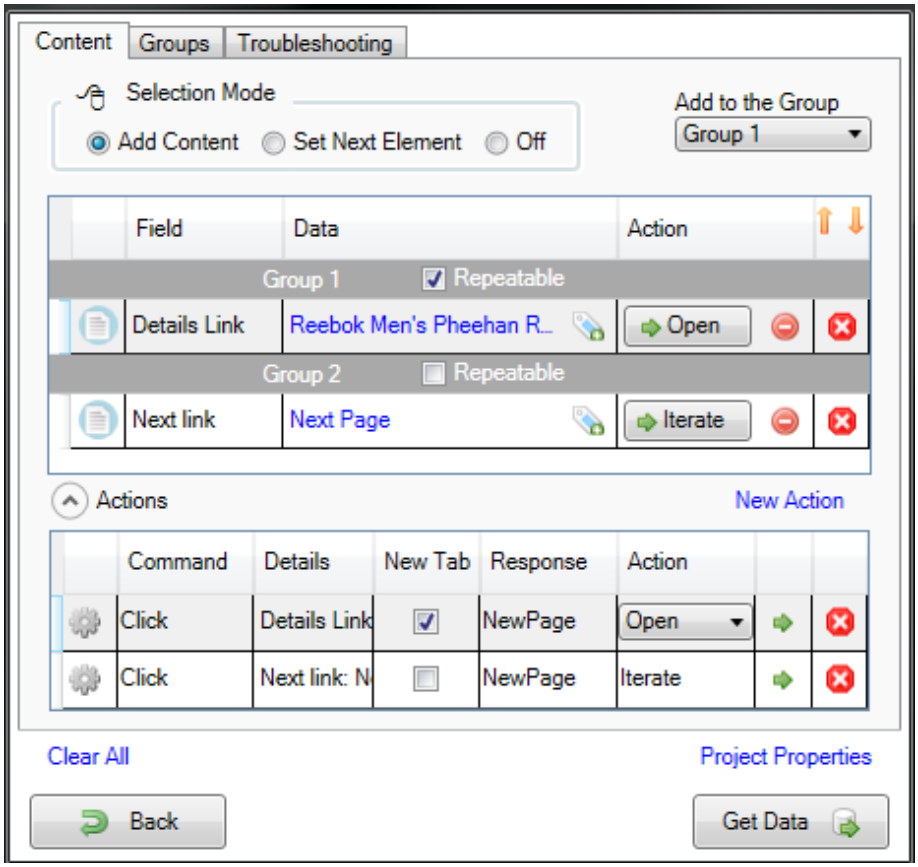


Figura 12. *Link "Next Page" selezionato dalla pagina web con associata nella colonna "Action" il pulsante "Iterate".*

Una volta acquisiti i dati se si è soddisfatti del risultato ottenuto si può procedere a salvare i dati premendo il pulsante "Salva". Il formato di output dipende dall'estensione del file di output. Il programma può salvare i dati come file CSV, XML, HTML, XLSX o script SQL. Questi formati possono essere facilmente importati in un file Excel, in un database o in un foglio di calcolo di Google.

2.3.2 ALGORITMO DI PROGRAMMAZIONE IN PYTHON

L'80 % (circa 18000 recensioni) del dataset di recensioni sulle 10 compagnie aeree considerate è stato ottenuto attraverso un algoritmo di programmazione in Python che scorre le pagine del social Trip Advisor e raccoglie le recensioni disponibili da ciascuna pagina e ha permesso di costituire la restante parte del Dataset.
Python è un linguaggio di programmazione ad alto livello, orientato agli oggetti adatto per sviluppare applicazioni distribuite, scripting, computazione numerica e system testing.
L'algoritmo utilizzato per la costituzione del dataset è stato reperito da una fonte online (GitHub) ed è riportato in Appendice D.
L'algoritmo va lanciato dal terminale (linea di comando) indicando il link della pagina di TripAsdvisor, il nome della compagnia aerea, il numero di recensioni da estrarre e la creazione di un file .CSV.
Successivamente l'algoritmo apre il browser e la pagina di TripAdvisor richiesta e inizia a scorrere le pagine raccogliendo da ciascuna pagina le 5 recensioni presenti.

Il processo di raccolta delle recensioni è stato abbastanza lungo poiché bisogna impostare un tempo iniziale time.sleep = 15 s in modo da poter caricare la prima pagina e poi un time.sleep = 7 s tra una pagina e l'altra per poter caricare la pagina successiva.
Il processo in alcuni casi presentava delle criticità poiché più è alto il numero di immagini e contenuti multimediali in una pagina e maggiore sono i secondi necessari per poter caricare la pagina. Se il numero di secondi necessari per caricare la pagina successiva di TripAdvisor, a causa della presenza di un numero elevato di contenuti multimediali (immagini, video, inserzioni pubblicitarie), è superiore a 7 secondi l'algoritmo si arresta ed è necessario aumentare il numero di secondi di time.sleep per poter riuscire a caricare quella pagina e poter andare avanti nella raccolta delle recensioni e degli altri UGC (nome compagnia, rating, data del volo, luogo di partenza e destinazione, autore della recensione, ecc.) necessari per l'analisi.
Dopo aver raccolto i dati essi vengono salvati in un file CSV che verrà poi esportato in R per poterlo utilizzare per la Text Mining Analysis.

2.4. DATASET

Dopo aver terminato la fase di web scraping e aver raccolto le recensioni delle 10 principali compagnie aeree del mondo, i dati raccolti sono stati salvati in un file Excel.
Le 10 compagnie prese in considerazione e per le quali sono state raccolte le recensioni sono: Alitalitalia, Lufthansa, Air France, British Airways, Qatar Airways, Etihad Airways, Easyjet, Ryanair, Emirates e Iberia.
Il dataset è costituito da 20451 recensioni e i dati sono suddivisi nel file Excel attraverso 10 variabili.
Le variabili sono : la compagnia aerea di riferimento ("Company"), l'autore della recensione ("author"), se presente la data di quando è stato effettuato il volo ("flightDate") altrimenti viene indicato il valore "NULL", la città di partenza del volo ("flightFromCity"), la città di destinazione del volo ("flightToCity"), il tipo di volo ("flightType"), la classe ("flightClass"), il rating dato dall'autore della recensione al volo su una scala da 1 a 5 ("mark") , il titolo ("title) assegnato dall'autore alla recensione e il corpo della recensione("text") (Figura 2.3.1).

	A	B	C	D	E	F	G	H	I	
1	Company	author	flightDate	flightFromCity	flightToCity	flightType	flightClass	mark	title	text
2	Lufthansa	Maurizio	mar-20	Roma	New York City	Internazionali	Economy	5.0	Idraulico	Divertimento con amico Che o conocido km talian chero visitar/© ny per la noche no capisci Che devo scrivere dime tu ora no so C
3	Lufthansa	Sandro-Cristina	ott-19	San Francisco	Monaco di Baviera	Internazionali	Economy	4.0	Viaggio rilassante	Siamo partiti da San Francisco alle 20.55, di conseguenza anche un pochino stanchi dalla giornata. Dopo il decollo, verso le 22.00, h
4	Lufthansa	maxniuz	dic-19	Venezia	Budapest	Europa	Economy	5.0	Nel cielo e fra le stelle	Per lo scorso Capodanno, ho volato fino a Budapest e al ritorno fino a Venezia con questa meravigliosa compagnia che ha dimostra
5	Lufthansa	patrizialazzaro	feb-20	Roma	Monaco di Baviera	Europa	Economy	5.0	puntuali per principio	Compagnia di volo caratterizzata da estrema puntualtv†, non c'v¤ che dire. Cortesia ed efficienza del personale di cabina, sempre
6	Lufthansa	gironimarika	set-19	Bologna	Francoforte	Europa	Economy	4.0	Merita	Merita, compagnia valida e soprattutto puntuale.Aerei nuovi spaziosi.Consiglio questa compagnia aerea
7	Lufthansa	Quest606509	feb-20	Milano	Miami	Internazionali	Economy	4.0	buono	Buono:puntuali,puliti,chiari,anche il costo buono,il personale gentile attenti. Una cosa se gli annunci in aereo fossero detti anche in
8	Lufthansa	Carmelo D	mar-20	Amsterdam	Roma	Europa	Economy	4.0	Sempre una buona compagnia	Sempre una buona compagnia aerea ad un prezzo accettabile.Personale gentile e professionale.Partito e arrivato in orario.
9	Lufthansa	Isabella	feb-20	Venezia	Francoforte	Europa	Economy	4.0	Weekend di San Valentino	Volo da Venezia a Francoforte diretto per weekend di San Valentino. Partenza puntuale. Avevo fatto check in on line, comodo, ma
10	Lufthansa	gianfri23	ago-19	Napoli	Monaco di Baviera	Europa	Premium Economy	5.0	Volo Napoli /Monaco/Napoli Agosto 2019 con mia moglie inferma per rottura piede	Dopo la rottura del piede di mia moglie abbiamo deciso di soggiornare in Austria per 10 giorni in un paesino del Tirolo , nel periodo a
11	Lufthansa	Marco G	feb-20	Ancona	Amsterdam	Europa	Economy	5.0	Ottima esperienza	Volo Ancona-Amsterdam via Monaco: ottima esperienza, con partenze ed atterraggi puntuali. Voli confortevoli, personale di bord
12	Lufthansa	Marina C	mag-19	Milano	Amburgo	Europa	Economy	5.0	Un ottimo volo!	Gentilissima accoglienza a bordo, descrizione del da parte del comandante e deliziosa cortesia da parte delle hostess. Nonostante l
13	Lufthansa	Fabrizio C	dic-19	Firenze	Bangkok	Internazionali	Economy	1.0	Modifica Vettore in peggio dopo l'acquisto del biglietto	Qualche mese dopo l'acquisto del biglietto Lufthansa ci comunica che nella tratta intercontinentale v¤ stato modificato il vettore da
14	Lufthansa	aspide78	feb-20	Monaco di Baviera	Chicago	Internazionali	Economy	1.0	Poca considerazione dei Passeggeri	Ho utilizzato questo volo per lavoro e mi v¤ toccata la classe Economy, i posti sono molto stretti inoltre 24 ore prima della partenza
15	Lufthansa	Caterina	ago-19	Milano	Nairobi	Internazionali	Economy	1.0	LUFTHANSA:THERE IS NO WORSE WAY TO FLY	Non volerv'si mai piv'n con Lufthansa. Nella mia vita ho preso tanti voli ma non mi v© mai capitato niente del genere. Durante il mi
16	Lufthansa	Riccardo S	ago-19	Los Angeles	Francoforte	Internazionali	Economy	4.0	viaggio piacevole	volo molto piacevole. personale gentile e scherzoso, le uniche cosa negative sono i pasti che non erano molto buoni e i programmi
17	Lufthansa	Marta M	mar-20	Francoforte	Firenze	Europa	Economy	1.0	Sprecisi	Viaggio da incubo,, dalla Danzica per Firenze con scalo a Francoforte,, aero come solito in ritardo solo pochi minuti per prendere la
18	Lufthansa	FuMattiaPascal	mar-20	Miami	Monaco di Baviera	Internazionali	Economy	3.0	Buon servizio ma pessimo mangiare	Ottimi servizi, velivolo perfetto , pulito, stabile e personale sempre cortese e sorridente. Perv's mangiare orribile, cibo scotto e incoll
19	Lufthansa	quadrifoglio2009	feb-20	Francoforte	Porto	Europa	Economy	4.0	esperienza positiva	Prima esperienza sulla Lufthansa e sinceramente v¤ stata molto positiva. Il personale v¤ molto gentile e l'aereo in ottime condizior
20	Lufthansa	Luigi P	feb-20	Orlando	Milano	Internazionali	Economy	4.0	Lufthansa	IL volo come al solito era praticame.nte completo, solo pochissimi posti liberi. Partito ed arrivato in perfetto orario. La qualtv† del c
21	Lufthansa	ardesign70	NULL	Monaco di Baviera	Milano	Europa	Economy	2.0	Prenoti Lufthansa e ti trovi su dolomite airlines	Dalla quinta fila spostati alla nona....per problemi di bilanciamento ???? Volo Lufthansa ma in realtv† Dolomite airlines in ritardo. S
22	Lufthansa	ardesign70	NULL	Cracovia	Monaco di Baviera	Europa	Economy	3.0	Senza parole	Ho viaggiato con un collega per lavoro ed avevamo fatto il check in online per scegliere i posti....dovevamo essere in 2 posti vicini
23	Lufthansa	Matteo	feb-20	Roma	Tokyo	Internazionali	Economy	1.0	Meglio viaggiare dentro una gabbia.	La prima tratta del viaggio quella breve v¤ andata bene, ma il vero livello di una compagnia aerea si vede sulle lunghe tratte. E nor
24	Lufthansa	Carlo C	dic-19	Francoforte	Los Angeles	Internazionali	Economy	4.0	Natale a Los Angeles	Viaggio con amici lungo ma tutto sommato confortevole Spazio decisamente superiore alle economy in cui ho viaggiato Vasta scel
25	Lufthansa	ardesign70	NULL	Milano	Francoforte	Europa	Economy	4.0	Ok	Volo ok, personale di terra e di aria cortese. Aereo comodo. Per spuntino una brioche con uvetta e da bere ampia scelta. Interessa
26	Lufthansa	Pamela T	NULL	Bologna	Buenos Aires	Internazionali	Business	5.0	Ottima	Viaggiare con la business v¤ una vera pacchia. Ho avuto la fortuna di beccare un prezzo molto basso e sono riuscita ad aggiudicam
27	Lufthansa	camifarina	feb-20	Milano	Miami	Internazionali	Economy	4.0	Volo tranquillo	Ho volato per la prima volta con questa compagnia ! Aereo 380 silenziosissimo !! L intrattenimento di bordo lascia molto a desidera
28	Lufthansa	Luca P	gen-20	Milano	Fort-de-France	Internazionali	Economy	4.0	Viaggio a Francoforte	Puntuali sia per l'andata che per il ritorno, checkin organizzato a zone che penalizza gli ultimi a salire, i quali non trovano quasi mai
29	Lufthansa	tommy g	NULL	Napoli	Paderborn	Europa	Business	5.0	La migliore	Ci volo da sempre, secondo me v¤ la migliore compagnia presente in europa al momento. Fantastiche le lounges, piloti meravigliosi
30	Lufthansa	dmourglia56	feb-20	Torino	Tromsv∏	Europa	Economy	5.0	Ottima!	Partiti in orario con la consociata Air Dolomiti, scalo a Frankfurt da dove il volo v¤ partito in orario. Cosv¨ al ritorno, sin in anticipo a F

Figura 13. *Rappresentazione dei dati nel file Excel.*

Il numero di recensioni raccolte per ciascuna compagnia aerea non è lo stesso ma è differente in base a quante recensioni è stato possibile reperire dal web su ciascuna compagnia aerea.
Sono state raccolte: 2501 recensioni per Lufthansa, 3002 recensioni per Alitalia, 1550 recensioni per Qatar Airways, 2998 recensioni per Ryanair, 1800 recensioni per Emirates, 4000 recensioni per Easyjet,1400 recensioni per Air France, 1600 recensioni per British Airways, 600 recensioni per Etihad Airways e 1000 recensioni per Iberia (Figura 14).
Come è possibile desumere dal grafico (Figura14), nel campione di compagnie aeree considerate, la compagnia aerea di cui sono state reperite più UGC (User Generated Content) è stata Easyjet mentre la compagnia aerea con il numero più basso di recensioni trovate è stata Etihad Airways.

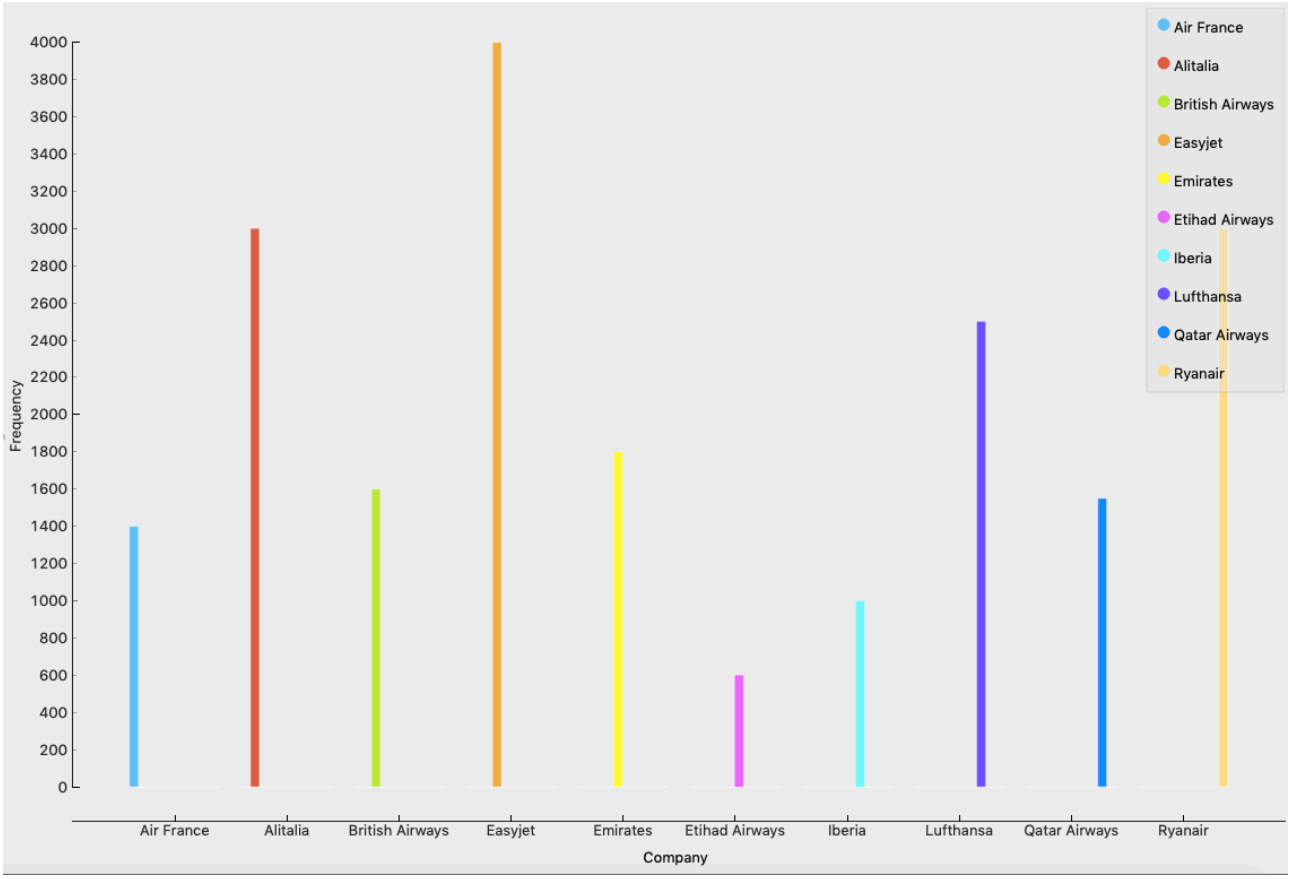


Figura 14. *Grafico rappresentante il numero di recensioni per ciascuna compagnia aerea.*

2.5. ANALISI PRELIMINARE SUL DATASET.

Tramite i dati presenti nel dataset è stato possibile fare un'analisi preliminare del campione di compagnie aeree prese in considerazione.
Considerando la variabile "flightType" è stato possibile tracciare un istogramma (Figura 15) che ha sull'asse delle ascisse la variabile "flightType" e sull'asse delle ordinate la frequenza ("Frequency").
Dal grafico è possibile notare che i due tipi di volo maggiormente compiuti dai clienti sono i voli in Europa e Internazionali e questo è dovuto al fatto che nel campione considerato la maggior parte delle compagnie aeree che ne fanno parte sono europee e quindi coprono tratte all'interno dell'Europa e viaggi internazionali.
Nel tipo di volo "Europa" la compagnia Easyjet rappresenta il picco ma ciò è anche dovuto al fatto che la percentuale di recensioni all'interno del dataset per la compagnia Easyjet è la più alta (19,55%). Anche Ryanair ha una frequenza di voli in Europa elevata. Nel tipo volo "Internazionali" si può notare che Emirates risulta avere la frequenza più elevata e nonostante la percentuale di recensioni raccolte su Emirates sia dell'8,8 % è la compagnia che nel campione considerato compie più viaggi internazionali, seguita da Qatar Airways con una frequenza di 1500 viaggi internazionali sui 1550 dati raccolti. Ciò permette di affermare che quasi la totalità del fatturato delle due compagnie proviene da viaggi internazionali.
Nella categoria "Nazionali" la frequenza più elevata la possiede Alitalia seguita da Easyjet e Ryanair. Le altre categorie non hanno valori o valori molto bassi o nulli (es. Medio Oriente e Russia) e ciò è dovuto al fatto che nel campione non vi sono compagnie asiatiche, americane o africane e quelle presenti non coprono molte tratte in tali paesi.

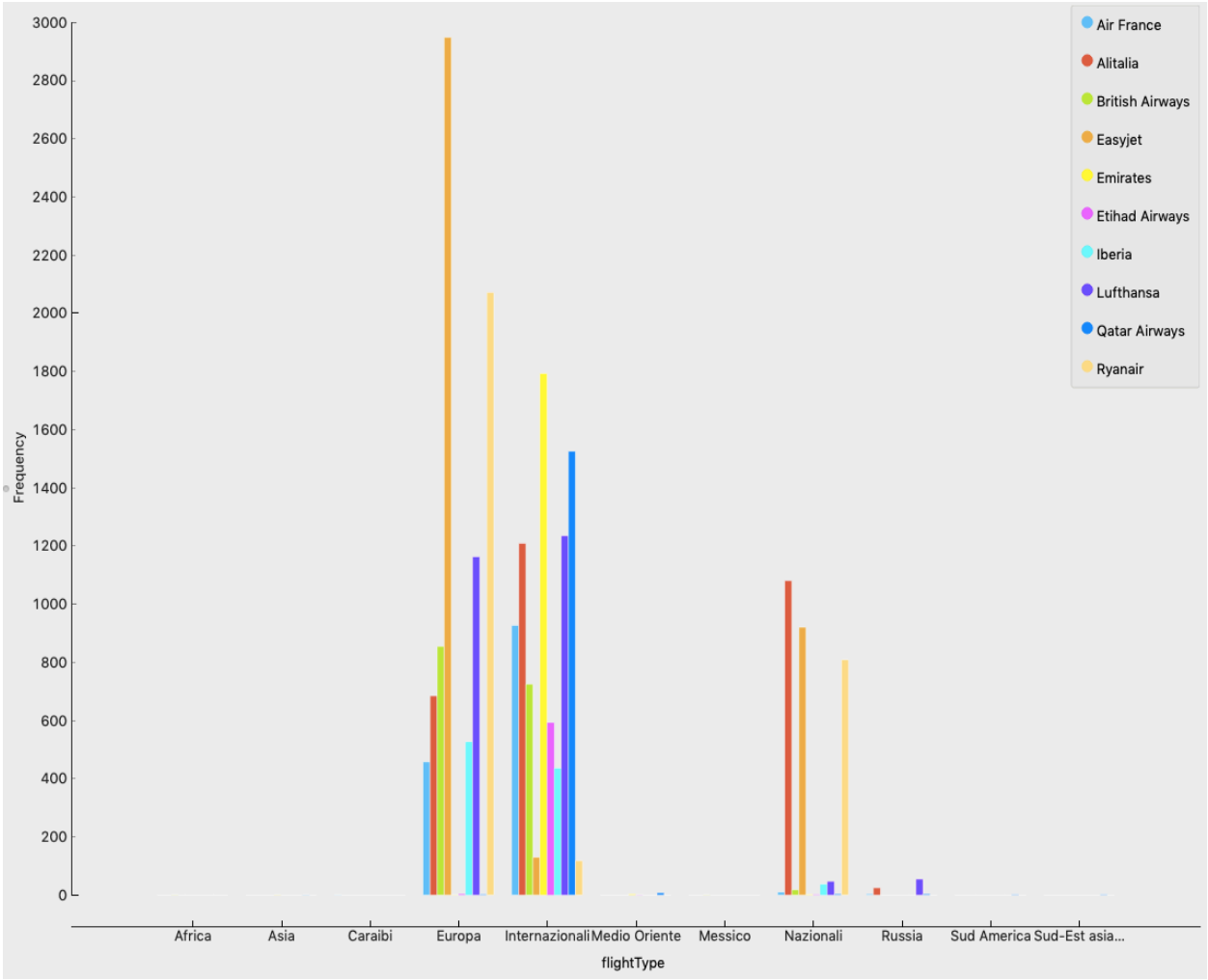


Figura 15. *Istogramma flightType – Frequency.*

La seconda variabile del campione che è possibile analizzare è “fligh Class”, ossia la classe in cui viaggiano le persone durante un volo.
Il grafico in figura 16 ha sull’asse delle ascisse la variabile “flightClass” e sull’asse delle ordinate la frequenza (“Frequency”).
Come è possibile notare dal grafico la maggior parte dei clienti di queste compagnie aeree viaggiano in “Economy” (89,65 % del campione considerato) che corrisponde al tipo di classe con il prezzo del biglietto aereo più basso. Nella classe economy spiccano i valori di Easyjet e Ryanair che sono due compagnie che sono due compagnie aeree che forniscono voli low cost. Nelle altre categorie della variabile flightClass i valori sono molto bassi poiché i clienti prediligono la classe Economy e solo pochi di loro volano in Business (6,96) e Premium Economy (2,98%) mentre la percentuale di persone che vola in Prima Classe si attesta intorno allo 0 %(0,40%).
Questo istogramma mostra che il fattore principale sul quale viene effettuata la scelta dal cliente sul tipo di biglietto da acquistare e fondata quasi totalmente sul prezzo.

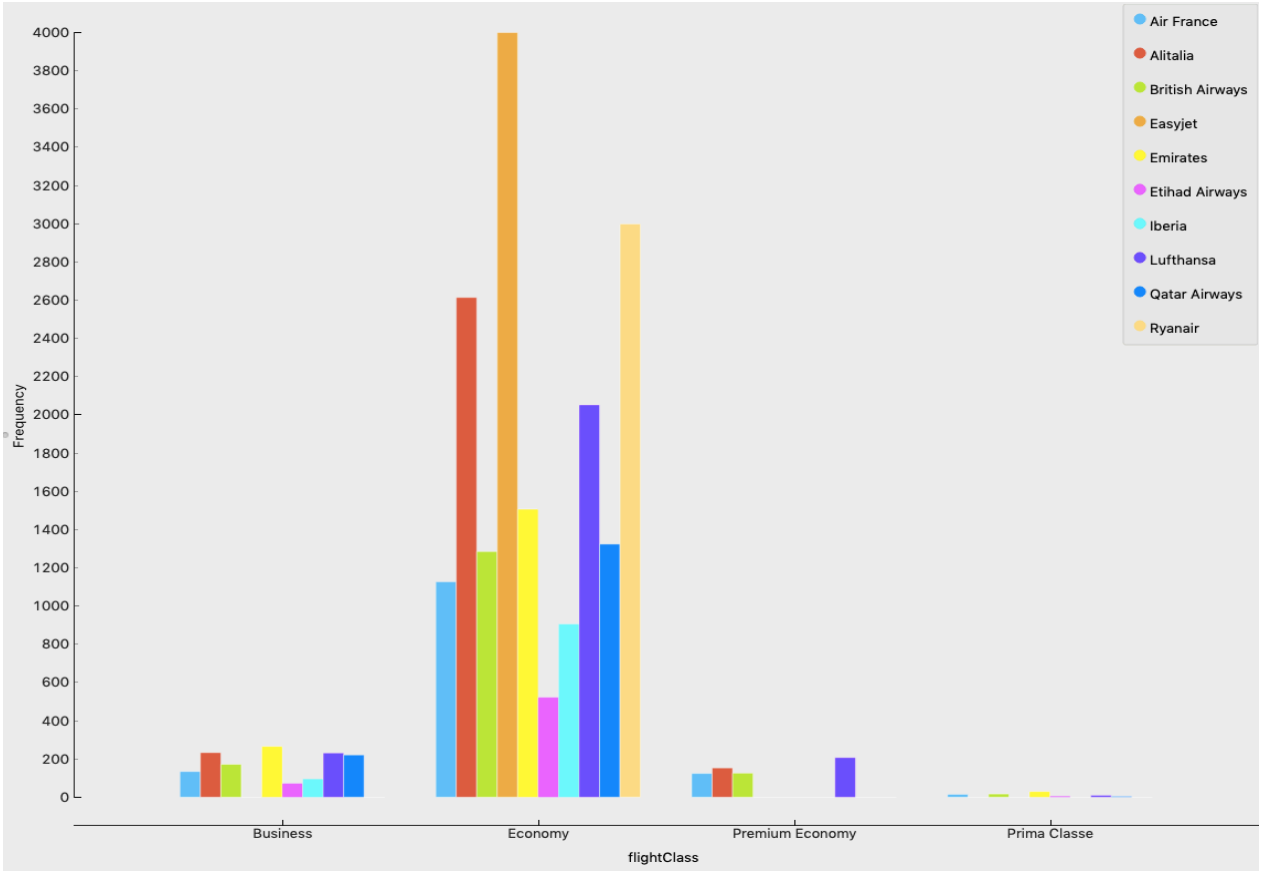


Figura 16. *Istogramma flightClass-Frequency*

L'ultima variabile che è possibile analizzare e che fornisce un focus oltre al testo della recensione è il rating che l'autore della recensione può rilasciare, ossia un voto da 1 a 5 (Figura 17).
Il 13,83% delle recensioni ha preso voto 1, in particolare le compagnie low cost Easyjet e Ryanair hanno preso le più alte percentuali di 1 di tutto il dataset di riferimento, rispettivamente il 3,61% e il 3,70% e ciò sta ad indicare che nonostante il prezzo delle compagnie low cost sia il più conveniente i servizi a bordo, il personale, la comodità e altri fattori potrebbero essere inferiori alle altre compagnie aeree.
Il 7,80% del totale delle recensioni ha ottenuto voto 2 e si possono notare di nuovo i valori di Easyjet (1,35%), Ryanair (1,65%) e Alitalia (1,30%) mentre le altre percentuali sono inferiori all'1%. Il 14,96% delle recensioni ha ottenuto il voto 3 e anche in questo caso i valori più elevati sono quelli di Easyjet, Ryanair e Alitalia ma ciò è condizionato anche dal fatto che i campioni di recensioni di queste tre compagnie sono i più numerosi.
Il 31,72% degli autori delle recensioni ha assegnato la valutazione 4 al volo che ha effettuato. In particolare si possono notare i valori di Easyjet (7,01%), Alitalia (4,69%), Lufthansa (4,41%) e Ryanair (4,15%) mentre le altre compagnie aeree hanno valori inferiori al 3%.
Il valore 5 è stato ottenuto dal 31,69% del totale del dataset.
La compagnia aerea ad avere il maggior numero di votazioni 5 è Emirates (6,25%) e nonostante il campione di recensioni di tale compagnia sia inferiore ad altre compagnie, Emirates risulta essere la migliore compagnia di volo per coloro che hanno rilasciato una recensione. Tra le compagnie che hanno ottenuto un rating di 5 vi è Alitalia (4,55%), Lufthansa (4,50%) e Qatar Airways (4,45%).

In conclusione i due elementi più significativi sono il valore 1 e 5. Le compagnie con rating 1 del dataset sono le compagnie low cost Easyjet e Ryanair e rispecchiano la realtà poiché esse offrono un servizio base, ossia il volo ad un prezzo competitivo ma ottengono un rating basso perché i servizi a bordo (cibo, bevande, giornale, ecc.) costano molto e anche gli altri servizi come la comodità del posto e altri comfort sono inferiori alle altre compagnie aeree. Le compagnie con rating 5 del dataset di riferimento sono Emirates, Lufthansa, Qatar Airways che sono tra le compagnie più importanti e prestigiose al mondo e ciò rispecchia la realtà. Alitalia ha una percentuale alta di rating 5 ma ciò è dovuto in parte al fatto che il campione di recensioni di Alitalia è tra i maggiori del dataset (3002 recensioni circa il 14,7%). I rating 4 e 5 mettono in evidenza che nonostante Etihad Airways sia composto da un campione modesto di recensioni (600 circa il 3%) ha delle percentuali più elevate nel rating 4 e 5 e ciò rispecchia la realtà perché tale compagnia risulta essere un competitor nello stesso segmento di mercato di Emirates e Qatar Airways. Iberia (1000 circa 4,8%) ha il più basso valore di rating 5 e anche un basso valore di rating 4 mentre British Airways (1600 circa l' 8%) e Air France (1400 circa il 7%) hanno dei valori simili ma più alti per i valori di rating 3,4 e 5.

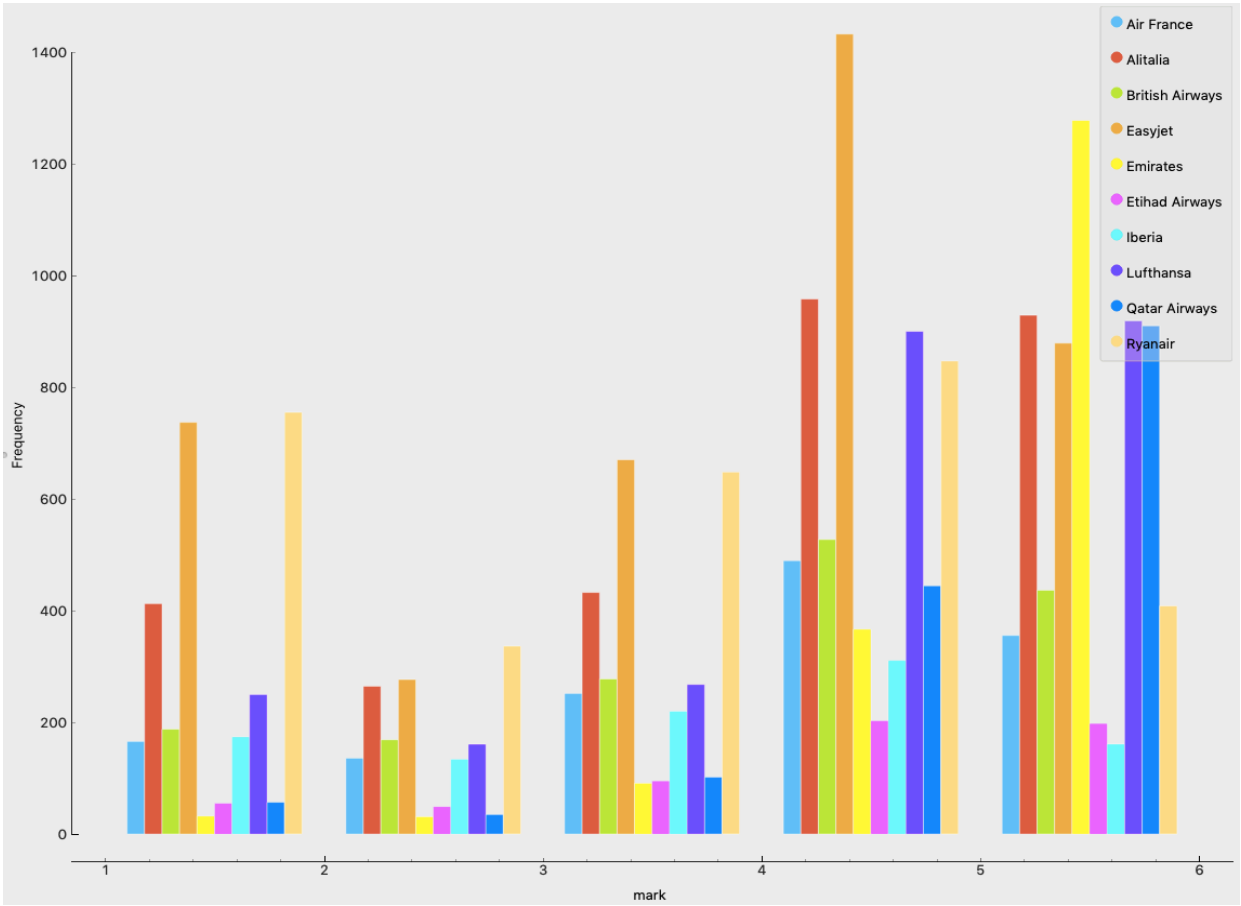


Figura 17. *Istogramma mark-Frequency*

2.6. ELABORAZIONE DATI.

Per analizzare il dataset degli User Generated Contents ottenuti ci si è avvalsi dell'utilizzo di un software statistico R.
R è un linguaggio di programmazione e un ambiente di sviluppo specifico per l'analisi statistica dei dati ed è disponibile in diverse versioni ed è scaricabile dai più diffusi sistemi operativi (Windows, Mac, Linux).
Una volta acquisiti gli UGC e costituito il dataset in Excel per poter analizzare il testo delle recensioni è stato necessario compiere sul testo di tutte le recensioni una fase di pre processamento tramite una funzione di R.

La fase di preprocessamento e preparazione del testo delle recensioni del dataset è stata particolarmente complessa poiché la funzione utilizzata era preimpostata sulla lingua inglese ("en") e l'insieme di stopwords eliminate dal testo erano in lingua inglese.
E' stato necessario modificare la funzione cambiando la lingua in italiano poiché il dataset di recensioni è in italiano ("italian") e la funzione utilizzata per ogni lingua possiede un dizionario di stopwords che elimina dal testo.
Inoltre è stato necessario eliminare un ulteriore gruppo di parole che erano presenti nel testo ma non apportavano alcun valore all'analisi utilizzando la sezione "customstopwords" della funzione utilizzata in R, creando un vettore contenente le ulteriori parole da eliminare dall'analisi.
Preprocessare un testo per essere poi analizzato significa "pulirlo" dai termini molto utilizzati come ad esempio la parola "volo" presente nel totale delle recensioni 21143 volte o la parola "compagnia" presente 10118 volte, dalle congiunzioni, avverbi, articoli, pronomi e in generale dalle parole "vuote" definite stopwords, ma anche la rimozione degli spazi aggiuntivi e della punteggiatura.
La funzione utilizzata in R si chiama "textProcessor" (Figura 18) ed appartiene al pacchetto "stm" che deve essere richiamato come una libreria in R ("library ("stm")") per poter essere utilizzata.

```
textProcessor(documents, metadata = NULL, lowercase = TRUE,
  removestopwords = TRUE, removenumbers = TRUE,
  removepunctuation = TRUE, ucp = FALSE, stem = TRUE,
  wordLengths = c(3, Inf), sparselevel = 1, language = "en",
  verbose = TRUE, onlycharacter = FALSE, striphtml = FALSE,
  customstopwords = NULL, custompunctuation = NULL, v1 = FALSE)
```

Figura 18. *Funzione textProcessor.*

Innanzitutto, il dataset contenente le recensioni è stato "tokenizzato", ossia il testo è stato eleaborato in modo da ottenere una suddivisione dello stesso in una collezione di parole definite "tokens".
Successivamente è avvenuta la rimozione della punteggiatura, dei numeri, delle stopwords predefinite dalla funzione e delle customstopwords (Appendice E) e delle parole con meno di tre caratteri (wordLenghts = c(3,inf)) con l'obiettivo di ridurre il più possibile il vocabolario e rimuovere tutti gli elementi non significativi ai fini dell'analisi.

Infine la funzione ha operato lo “stemming” e il “rooting” per ridurre le parole alle loro radici e infine la normalizzazione, che include la rimozione di errori di battitura e la conversione delle parole in minuscolo.

In conclusione, il dataset di recensioni disponibili per l’analisi sono N = 20446 mentre il dataset iniziale era composto da 20451 recensioni.

2.7. SCELTA DEL NUMERO OTTIMO DI TOPIC.

Terminata la fase di pre-processing e prima di determinare i topic tramite l’applicazione dell’algoritmo STM è stato necessario stabilire il numero ottimo di topic da prendere in considerazione per l’analisi.
Per determinare il numero ottimo di topic da analizzare è stata utilizzata in R la funzione “searchK” che ha permesso di ottenere come output quattro indicatori: Held-Out Likelihood, Residuals, Semantic Coherence e Lower Bound (Figura 19).
L’indicatore held-out likelihood misura la probabilità di held-out del modello e mostra che il modello migliora man mano che il numero di topic aumenta per poi assestarsi intorno ad un range di valori e continuare con andamento costante all’interno di questo range con l’aumentare del numero di topic.
Il grafico Number of Topics (K) vs Residuals mostra i residui del modello per ogni numero di topic compreso tra 5 e 50. L’indicatore Semantic Coherence (coerenza semantica) è massimizzato quando le parole più probabili in un dato argomento coesistono insieme ed è una metrica che è simile al giudizio umano per stabilire la qualità di un topic.
Come è possibile osservare dal grafico avere un’alta coerenza semantica è relativamente facile, però, se si hanno solo pochi argomenti (es. da 5 a 10 topic) dominati da parole molto comuni, quindi si vuole guardare sia alla coerenza semantica che all’esclusività delle parole rispetto agli argomenti.
Il Lower Bound indica la misura interna della misura di idoneità del modello in base al numero di topic.

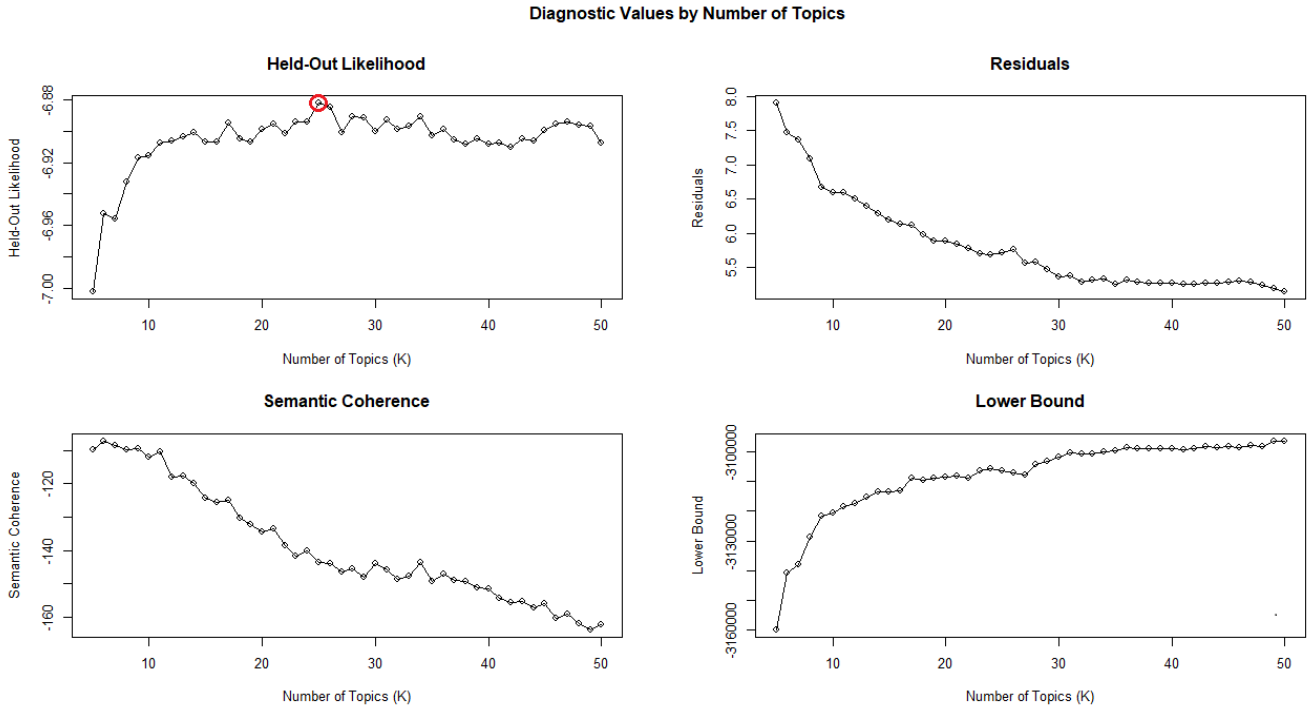


Figura 19. *Grafici Held-Out Likelihood, Residuals, Semantic Coherence, Lower Bound.*

Il metodo che è stato utilizzato per scegliere il numero di topic ottimo da considerare per l'analisi si basa sul grafico Held-Out Likelihood vs Number of Topics. Per avere il numero ottimo di topic si è preso il numero di topic che ha il valore massimo nel grafico, ossia K=25. Sono stati considerati per l'analisi delle recensioni K=25 topic.

2.8. ETICHETTATURA.

Questa è la fase antecedente all'analisi dei risultati, in cui ci si occupa di dare un titolo a ciascun topic fornito dall'algoritmo STM applicato in R.
Questa fase dipende principalmente dall'analista che assegnerà a ciascun topic un titolo.
Non vi è un metodo unico per l'assegnazione dei vari titoli ai topic. Per compiere tale procedura si è osservato le parole (Highest Prob., FREX, Lift, Score) che componevano ciascun topic e si è cercato di assegnare l'etichetta che meglio le raggruppava.
Non essendo un metodo unico e universale, soggetti diversi possono assegnare etichette diverse a ciascun topic.
Il concetto principale da tenere presente in questa fase è quello di assegnare a ciascun topic l'etichetta che meglio raggruppa e rappresenta l'insieme di parole che costituisce ciascun topic.

In Appendice F è rappresentata la lista contenente i 25 topics con le 7 parole Highest Prob, FREX, Lift e Score per ciascun topic fornite dall'output di R e le etichette assegnate ai vari topic.

La fase di etichettatura non ha mostrato particolari difficoltà tranne che nel caso del topic 24. L'identificazione di un'etichetta che raggruppasse tutte le parole del topic è risultata più difficile poiché all'interno del topic 24 vi erano parole riguardanti il ritardo aereo (es. ritardo, aereo, ecc), parole che si riferivano all'utilizzo a bordo dell'aria condizionata (es. aria, condizionata, ecc.) e aggettivi dispregiativi riferiti all'inefficienza e alla scortesia del personale (es. pessimo, personale, scortese, vergogna, peggior, ecc) di bordo.
Per poter assegnare un'etichetta nel modo corretto al topic 24 si è osservato le recensioni in cui il topic 24 era il più discusso e si è notato che nella maggior parte delle recensioni il cliente che ha prodotto quella recensione lamentava un problema durante il volo, un disguido con la compagnia, l'inefficienza e/o la scortesia (e in alcuni casi l'arroganza) del personale e altre lamentele e così a tale topic è stata assegnata l'etichetta "Lamentele dei clienti".
Inizialmente il topic 24 era stato contrassegnato con l'etichetta "Aria condizionata" poiché si pensava ad un problema legato a questo servizio a bordo poi si è notato che in alcune recensioni si lamentavano problemi con l'aria condizionata, ma non nella maggior parte delle recensioni in cui il topic 24 era il più discusso si parlava di tale problema e quindi si è optato per l'etichetta "Lamentele dei clienti".
E' possibile stabilire in maniera più precisa quale etichetta assegnare a ciascun topic utilizzando la funzione findThoughts appartenente al pacchetto STM in R, che permette di visualizzare per ciascun topic le recensioni in cui il topic i-esimo è il più discusso e vedendo le recensioni che si riferiscono maggiormente al topic è possibile assegnare un'etichetta al topic in maniera più accurata.
La Figura 20 mostra due recensioni in cui il topic 1 è prevalente e come si può notare esse si riferiscono alla prenotazione del volo ed è per questo che al topic 1 è stata assegnata l'etichetta "Prenotazione volo (1)".

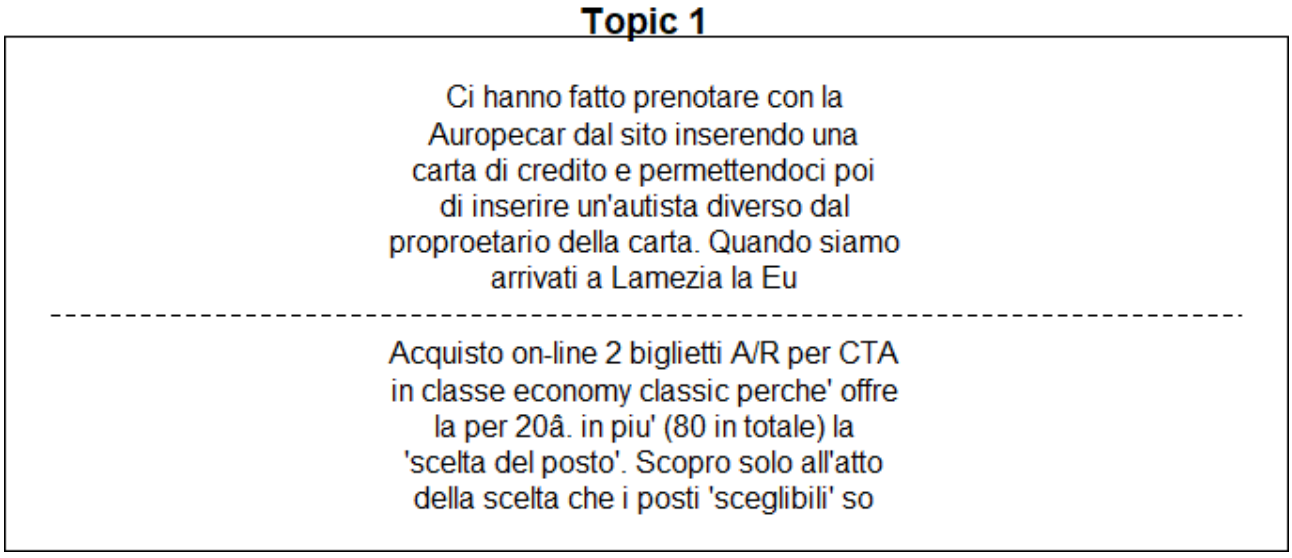


Figura 20. *Esempio di due recensioni in cui il topic 1 è prevalente*

La Figura 21 mostra le due recensioni in cui il topic 23 è prevalente e come si può notare esse si riferiscono prevalentemente all'intrattenimento e agli extra che si possono consumare a bordo durante il volo ed è per questo che al topic 23 è stata assegnata l'etichetta "Intrattenimento".

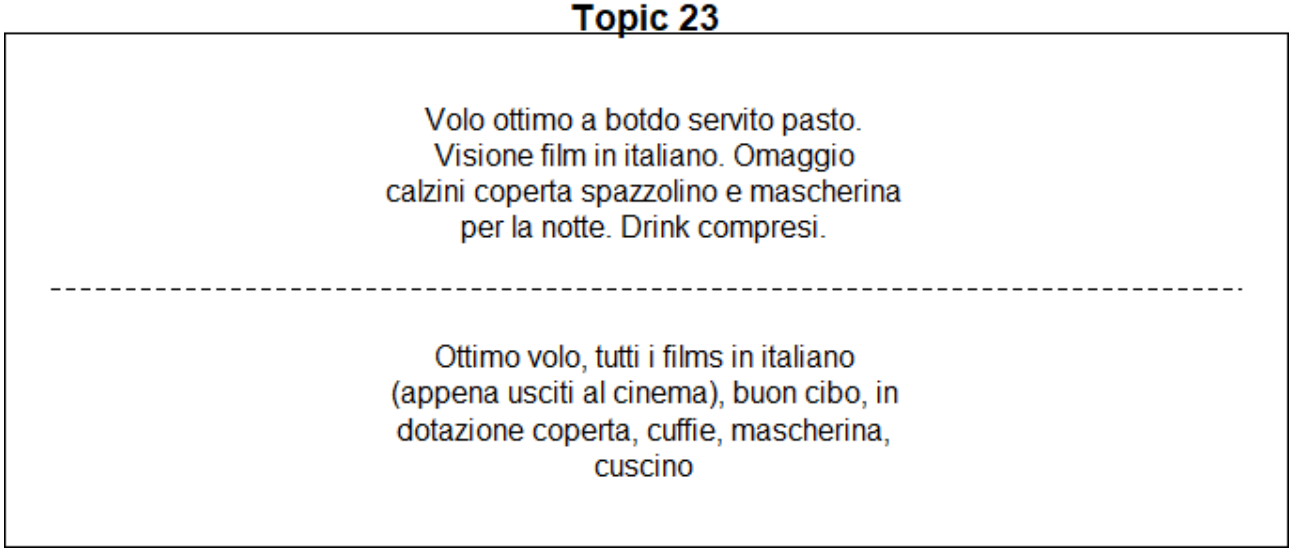


Figura 21. *Esempio di due recensioni in cui il topic 23 è prevalente*

Sono stati riportati gli esempi per il topic 1 e il topic 23, ma è possibile ottenere due o più recensioni in cui il topic i-esimo è prevalente per poter assegnare a ciascun topic l'etichetta più rappresentativa delle parole dello stesso.
Per l'assegnazione dell'etichetta ad un topic in primo luogo si devono osservare le parole che compongono il topic i-esimo e assegnare un'etichetta che le raggruppi, in secondo luogo in caso di difficoltà, si può utilizzare la funzione findThoughts per poter trovare l'etichetta adeguata.

2.9. VALIDAZIONE DELL'ALGORITMO

Dopo aver preprocessato le recensioni del dataset, aver scelto il numero ottimo di topic e aver generato e assegnato i topic alle recensioni si è proceduto con la fase di validazione dell'algoritmo STM.
La fase di validazione dell'algoritmo consiste in un procedimento applicato dall'analista per confrontare le assegnazioni dei topic alle recensioni compiute dall'algoritmo con la realtà.
In questa fase è stato considerato il concetto di validità di convergenza, ossia il grado di correlazione che vi è tra le misure di uno stesso costrutto ottenute con metodi differenti e indipendenti tra loro (F. Franceschini, 2001).

Per poter effettuare la fase di validazione dell'algoritmo sono stati estratti 4 campioni di 50 recensioni ciascuno e i 4 campioni sono anche stati analizzati insieme come un unico campione di 200 recensioni (circa l'1% del dataset) .
La procedura consiste nel leggere le recensioni di ogni campione estratto ed assegnare manualmente uno o più topics dei 25 topics utilizzati che l'analista ritiene discussi all'interno della recensione di riferimento.
Sono state utilizzate alcune definizioni per valutare se i risultati ottenuti dall'analista sono uguali oppure no a quelli ottenuti dall'algoritmo STM in R:

- TRUE POSITIVE (TP): se il topic assegnato dall'analista alla recensione corrisponde con il topic con la percentuale di prevalenza θ maggiore assegnato dall'algoritmo STM.
- TRUE NEGATIVE (TN): se l'analista non riesce ad assegnare con chiarezza un topic alla recensione e l'algoritmo non identifica una percentuale di prevalenza media θ nettamente superiore alle altre per cui un topic si può definire prevalente all'interno della recensione.
- FALSE POSITIVE (FP): l'algoritmo assegna come prevalente un topic mentre l'analista leggendo la recensione ne assegna uno differente come prevalente nella recensione.
- FALSE NEGATIVE (FN): l'algoritmo non identifica una percentuale di prevalenza nettamente superiore per un topic mentre l'analista sì.

Di seguito viene riportata una recensione del campione di 200 recensioni del documento Excel "Validazione dell'algoritmo":

" Ci volo da sempre, secondo me e' la migliore compagnia presente in europa al momento. Fantastiche le lounges, piloti meravigliosi e competenti. Customer service unico e disponibile. Aerei comodi con molto spazio per le gambe, soprattutto nella business class"

L'algoritmo STM indica che per questa recensione il topic prevalente è il topic 20 (Prenotazione volo (2)) come mostra la riga seguente che mostra le percentuali di prevalenza dei vari topic nella recensione e in grassetto la percentuale di prevalenza più elevata corrispondente al topic 20:

[0,0100 ; 0,1103 ; 0,0098 ; 0,0191 ; 0,0204 ; 0,03996 ; 0,0158 ; 0,0245 ; 0,0100 ; 0,0503 ; 0,0612 ; 0,0989 ; 0,0130 ; 0,0313 ; 0,0115 ; 0,0147 ; 0,1204 ; 0,0092 ; 0,0092 ; **0,1260** ; 0,0903 ; 0,0253 ; 0,0479 ; 0,0101 ; 0,0198]

L'analista leggendo la recensione nota che in essa non si parla della prenotazione del volo ma della cortesia e delle competenze del personale, della comodità del posto durante il

viaggio e del tipo di classe e quindi l'esperto avrebbe assegnato altri topic a questa recensione, come il topic 2 (Tipo di classe con θ = 11,03%), il topic 17 (Posto a bordo con θ = 12,04 %), il topic 14 (Competenze personale (1) con θ = 3,13 %) o il topic 21 (Competenze personale (2) con θ = 9,03 %).
Nel topic 20 assegnato come prevalente nella recensione riportata come esempio non vi è alcuna delle parole presenti nella recensione, invece nei topic elencati in precedenza compaiono le parole "competenti", "disponibile", "unico", "comodi", "spazio", "gambe", "class" ed è per tale motivi che essi sarebbero stati più appropriati da assegnare come prevalenti alla recensione precedente.
Quello precedente è un esempio di FALSE POSITIVE (FP) il cui l'algoritmo STM associa alla recensione un topic come prevalente mentre l'analista ne avrebbe associati altri ma non quello assegnato dall'algoritmo STM.
Questo procedimento è stato applicato alle recensioni di ciascun campione estratto dal documento Excel (Validazione dell'algoritmo) ed è stato possibile stabilire il numero di TP, TN, FP e FN presenti in ciascun campione e riassunti nella tabella seguente.

Tabella 1. *Numero di recensioni TP, TN, FP, FN, per il campione 1.*

NUMERO CAMPIONE	TIPO	NUMERO DI RECENSIONI
1	TRUE POSITIVE (TP)	30
	TRUE NEGATIVE (TN)	7
	FALSE POSITIVE (FP)	9
	FALSE NEGATIVE (FN)	4
	TOTALE	50

Tabella 2. *Numero di recensioni TP, TN, FP, FN, per il campione 2.*

NUMERO CAMPIONE	TIPO	NUMERO DI RECENSIONI
2	TRUE POSITIVE (TP)	28
	TRUE NEGATIVE (TN)	7
	FALSE POSITIVE (FP)	11
	FALSE NEGATIVE (FN)	4
	TOTALE	50

Tabella 3. *Numero di recensioni TP, TN, FP, FN, per il campione 3.*

NUMERO CAMPIONE	TIPO	NUMERO DI RECENSIONI
3	TRUE POSITIVE (TP)	16
	TRUE NEGATIVE (TN)	9
	FALSE POSITIVE (FP)	13
	FALSE NEGATIVE (FN)	12
	TOTALE	50

Tabella 4. *Numero di recensioni TP, TN, FP, FN, per il campione 4.*

NUMERO CAMPIONE	TIPO	NUMERO DI RECENSIONI
4	TRUE POSITIVE (TP)	17
	TRUE NEGATIVE (TN)	8
	FALSE POSITIVE (FP)	17
	FALSE NEGATIVE (FN)	8
	TOTALE	50

Tabella 5. *Numero di recensioni TP, TN, FP, FN, per il campione totale.*

NUMERO CAMPIONE	TIPO	NUMERO DI RECENSIONI
TOTALE (formato dalle recensioni dei 4 campioni precedenti)	TRUE POSITIVE (TP)	87
	TRUE NEGATIVE (TN)	33
	FALSE POSITIVE (FP)	50
	FALSE NEGATIVE (FN)	30
	TOTALE	200

In seguito per valutare la bontà dell'algoritmo STM sono stati utilizzati 4 indicatori (Costa, 2007) per ciascuno dei campioni e per ogni campione è stato prodotto un diagramma radar per i 4 indicatori (Appendice G).

Il primo è l'indicatore RECALL, esso rappresenta il rapporto tra il numero di veri positivi (TP) e la somma di veri positivi (TP) e falsi negativi (FN).

$$\text{RECALL (R)} = \frac{TP}{TP + FN}$$

Il secondo indicatore è PRECISION (P), esso indica il rapporto tra i veri positivi (TP) e la somma di veri positivi (TP) e falsi positivi (FP).

$$\text{PRECISION (P)} = \frac{TP}{TP + FP}$$

Il terzo indicatore utilizzato è F-MEASURE (F) ed è definito come la media ponderata dei due precedenti indicatori (RECALL e PRECISION).

$$\text{F – MEASURE (F)} = 2 \times \frac{P \times R}{P + R}$$

Il quarto ed ultimo indicatore si chiama ACCURACY e rappresenta il rapporto tra la somma dei veri positivi (TP) e i veri negativi (TN) e il totale degli elementi (TP, TN, FP, FN) in cui può essere classificato un elemento.

$$\text{ACCURACY (A)} = \frac{TP + TN}{TP + TN + FP + FN}$$

Nelle tabelle seguenti sono riportati i valori dei 4 indicatori precedenti per i 4 campioni e per il campione totale.

Tabella 6. *Risultati degli indicatori RECALL, PRECISION, F – MESAURE, ACCURACY per il campione 1.*

NUMERO CAMPIONE	INDICATORE	RISULTATO
1	RECALL	88,24 %
	PRECISION	76,92 %
	F – MEASURE	82,19 %
	ACCURACY	74,00 %

Tabella 7. *Risultati degli indicatori RECALL, PRECISION, F – MESAURE, ACCURACY per il campione 2.*

NUMERO CAMPIONE	INDICATORE	RISULTATO
2	RECALL	87,50 %
	PRECISION	71,79 %
	F – MEASURE	78,87 %
	ACCURACY	70,00 %

Tabella 8. *Risultati degli indicatori RECALL, PRECISION, F – MESAURE, ACCURACY per il campione 3.*

NUMERO CAMPIONE	INDICATORE	RISULTATO
3	RECALL	57,14 %
	PRECISION	55,17 %
	F – MEASURE	56,14 %
	ACCURACY	50,00 %

Tabella 9. *Risultati degli indicatori RECALL, PRECISION, F – MESAURE, ACCURACY per il campione 4.*

NUMERO CAMPIONE	INDICATORE	RISULTATO
4	RECALL	68,00 %
	PRECISION	50,00 %
	F – MEASURE	57,63 %
	ACCURACY	50,00 %

Tabella 10. *Risultati degli indicatori RECALL, PRECISION, F – MESAURE, ACCURACY per il campione totale.*

NUMERO CAMPIONE	INDICATORE	RISULTATO
TOTALE	RECALL	74, 36 %
	PRECISION	63,50 %
	F – MEASURE	68,50 %
	ACCURACY	60,00 %

In letteratura i ricercatori hanno definito alcune soglie entro le quali verificare se i valori ottenuti per i 4 indicatori rispettano le indicazioni fornite dalla letteratura.

L'indicatore RECALL secondo uno studio effettuato da Powers e D. Martin ha valori compresi tra il 51 % e l'87 %, mentre i ricercatori B. Pang, L. Lee e S. Vaithyanathan in un articolo ("*Thumbs up? Sentiment Classification using Machine Learning Techniques"*) sull'impiego di tecniche di machine learning nella sentiment analysis nel loro esperimento hanno ottenuto dei valori di RECALL compresi tra il 50 % e il 69 % e dei valori di ACCURACY compresi tra il 58 % e il 64 %.

In un articolo del 2004 ("*Protein NMR Recall, Precision, and F-measure Scores (RPF Scores): Structure Quality Assessment Measures Based on Information Retrieval Statistics*") sugli indicatori RECALL, PRECISION e F – MEASURE i ricercatori Yuanpeng J. Huang, Robert Powers, and Gaetano T. Montelione mostrano come l'indicatore RECALL possa assumere valori compresi tra il 72 % e l'83 %, come l'indicatore PRECISION possa avere valori compresi tra l'81 % e il 97 % e l'indicatore F – MESAURE possa avere valori compresi tra 0,81 e 0,89.

I ricercatori S. Velupillai, H. Dalianis, M. Hassel nell'articolo "*Developing a standard for de-identifying electronic patient records written in Swedish: precision, recall and F-measure in a manual and computerized annotation trial"* hanno ottenuto nel loro esperimento un valore di F – MESAURE di 0,65 e 0,80.

Gli studiosi W. Kasper e M. Vela nell'articolo "Sentiment Analysis for Hotel Reviews" mostrano come i valori di ACCURACY siano compresi tra il 54% e il 67 % e i valori di F – MESAURE siano compresi tra 0,66 e 0,81.

Infine se il valore di ACCURACY è superiore al 55 % significa che si è ottenuto un buon risultato di validazione dell'algoritmo STM (K. Nassirtoussi, 2014).

I valori ottenuti nel lavoro di tesi per i 4 indicatori rientrano nelle soglie precedentemente citate negli articoli presi in analisi.
L'indicatore RECALL è leggermente superiore alle soglie nel primo e nel secondo campione rispetto alla letteratura considerata. L'indicatore PRECISION dei campioni considerati è leggermente inferiore alla letteratura considerata.
I valori di F – MESAURE rientrano in almeno uno dei range della letteratura presa in esame.
Infine l'indicatore ACCURACY è superiore alla soglia minima per poter ottenere un buon risultato di validazione dell'algoritmo (55%) eccetto nei campioni 3 e 4 in cui è di poco inferiore (50%).

CAPITOLO 3

3.1. PRESENTAZIONE GENERALE DEI RISULTATI SUL TOTALE DELLE RECENSIONI.

Dopo aver concluso la fase di etichettatura dei 25 topic è stato applicato l'algoritmo STM al dataset composto da 20446 recensioni e da 4076 vocaboli rimanenti dopo la fase di preprocessamento riferite ad un campione di 10 compagnie aeree.
Dopo aver applicato la funzione stm in R è stato possibile ottenere per ciascun topic la percentuale di prevalenza (θ) di ogni topic in ciascuna recensione e questo permetteva di stabilire quale fosse il topic più discusso per ciascuna recensione.
Avendo le percentuali di prevalenza di ciascun topic per ognuna delle 20446 recensioni è stato possibile calcolare la percentuale di prevalenza media sul totale delle recensioni, la varianza e la deviazione standard.
In base alla percentuale di prevalenza ogni topic rispetto a ciascuna recensione viene indicato nel modo seguente (rappresentazione delle percentuali di prevalenza dei topic per la prima recensione):

(0,02008372 ; 0,02899948 ; 0,01430772 ; 0,01903741 ; 0,11238538 ; 0,04593689 ; 0,07966489 ; 0,02158616 ; 0,01403555 ; 0,01944438 ; 0,01404838. ; 0,029294717 ; 0,01479479 ; 0,02967322 ; 0,04009451 ; 0,02380338 ; 0,01822202 ; 0,01607259 ; 0,00932929 ; ***0,26761979*** *; 0,04684495 ; 0,04555842 ; 0,02817411 ; 0,022440725 ; 0,01854753)*

Osservando le percentuali di prevalenza dei vari topic per ciascuna recensione è possibile notare quale sia il topic più discusso, ossia quello con la percentuale di prevalenza maggiore e si può notare che per l'esempio riportato il topic più discusso con la percentuale di prevalenza maggiore è il topic 20 (Prenotazione volo on-line) con un θ=26,76 %
Dopo aver effettuato questa operazione è stato possibile stabilire il numero di recensioni sul totale in cui il topic i-esimo è prevalente e la corrispondente percentuale sul totale delle recensioni del dataset (Tabella 11).

Tabella 11. *Risultati di prevalenza media, dev. standard, varianza, numero di recensioni sul totale in cui il topic i-esimo è prevalente, percentuale sul totale per le recensioni del dataset completo pre-processato.*

	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
Prevalenza media del topic i-esimo sul totale delle recensioni	2,85%	3,78%	1,99%	3,10%	2,83%
Deviazione standard	5,36%	5,51%	3,73%	5,59%	4,69%
Varianza	0,28%	0,30%	0,13%	0,31%	0,22%
Numero di recensioni	450	676	245	634	428

sul totale in cui il topic i-esimo è prevalente					
Percentuale sul totale delle recensioni	2,20%	3,30%	1,19%	3,10%	2,09%
	Topic 6	**Topic 7**	**Topic 8**	**Topic 9**	**Topic 10**
Prevalenza media del topic i-esimo sul totale delle recensioni	5,81%	4,47%	3,56%	1,97%	3,44%
Deviazione standard	5,22%	9,17%	4,39%	3,39%	6,33%
Varianza	0,27%	0,84%	0,19%	0,11%	0,40%
Numero di recensioni sul totale in cui il topic i-esimo è prevalente	683	1153	413	184	614
Percentuale sul totale delle recensioni	3,34%	5,63%	2,01%	0,89%	3%
	Topic 11	**Topic 12**	**Topic 13**	**Topic 14**	**Topic 15**
Prevalenza media del topic i-esimo sul totale delle recensioni	2,68%	5,26%	3,56%	3,79%	3,78%
Deviazione standard	4,70%	7,42%	5,45%	5,76%	7,40%
Varianza	0,22%	0,55%	0,29%	0,33%	0,54%
Numero di recensioni sul totale in cui il topic i-esimo è prevalente	427	1089	593	670	796
Percentuale sul totale delle recensioni	2,08%	5,32%	2,90%	3,27%	3,89%
	Topic 16	**Topic 17**	**Topic 18**	**Topic 19**	**Topic 20**
Prevalenza media del	2,30%	4,02%	5,17%	4,37%	3,29%

topic i-esimo sul totale delle recensioni					
Deviazione standard	4,33%	5,46%	9,96%	7,69%	6,48%
Varianza	0,18%	0,29%	0,99%	0,59%	0,42%
Numero di recensioni sul totale in cui il topic i-esimo è prevalente	341	631	1503	1196	624
Percentuale sul totale delle recensioni	1,66%	3,08%	7,35%	5,84%	3,05%
	Topic 21	**Topic 22**	**Topic 23**	**Topic 24**	**Topic 25**
Prevalenza media del topic i-esimo sul totale delle recensioni	10,59%	5,98%	6,71%	3,06%	1,50%
Deviazione standard	10,53%	7,64%	10,30%	5,66%	0,76%
Varianza	1,11%	0,58%	1,06%	0,32%	0,00580%
Numero di recensioni sul totale in cui il topic i-esimo è prevalente	3206	1225	2133	526	6
Percentuale sul totale delle recensioni	15,68%	5,99%	10,43%	2,57%	0,02%

Dopo aver osservato i dati è stato possibile notare che il topic con la prevalenza media più elevata è il topic 21(Competenze personale (2)) con =10,59% seguito dai topic 23 (Intrattenimento) con θ=6,71% e il topic 22 (Anticipo/ritardo volo) con θ=5,98%.
I topic con la percentuale di prevalenza media più bassa e i meno discussi sul totale delle recensioni sono il topic 3 (Famiglia) con θ= 1,99%, il topic 9 (Snack e bevande) con θ= 1,97% e il topic con la prevalenza media più bassa e prevalente in sole 6 recensioni è il topic 25 (Problemi riscontrati con la compagnia aerea) con θ=1,50%.
Il topic 21 è il più discusso ed è il più dibattuto in 3206 recensioni e rappresenta l'argomento su cui maggiormente pone l'attenzione la clientela mentre il topic meno dibattutto nel campione considerato è il topic 25, prevalente in appena 6 recensioni (Tabella 11).

La Figura 22 (Top Topics) mostra la classifica dal topic più discusso a quello meno discusso all'interno del dataset considerato ordinando i vari topic secondo la prevalenza media.

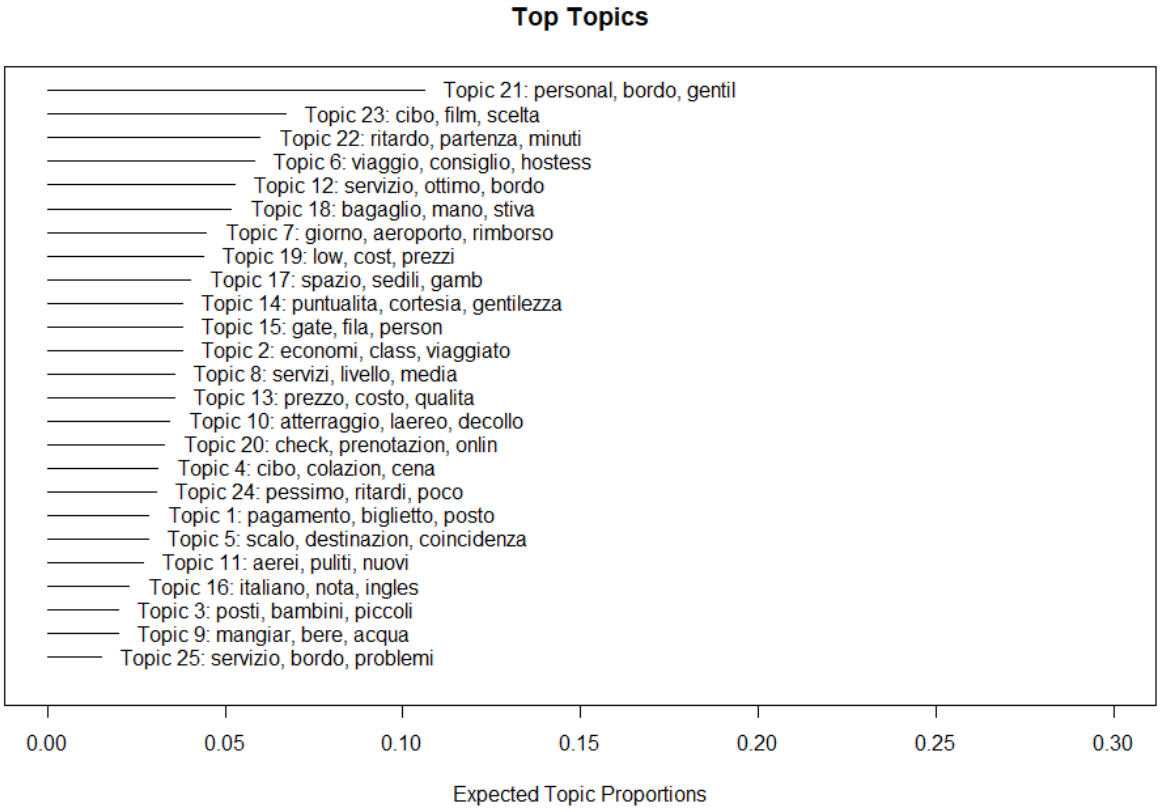


Figura 22. *Lista dei topic ordinati.*

E' stato poi possibile fare un istogramma (figura 23) rappresentante la media percentuale di prevalenza per ogni topic. L'istogramma sottostante per ogni topic mostra la media considerando solo le recensioni per cui il topic i-esimo è prevalente. Tutti i valori dell'istogramma sono compresi nell'intervallo [23,97% , 35,88%].
Il procedimento che è stato eseguito è stato quello di sommare le percentuali di prevalenza delle recensioni in cui il topic i-esimo è prevalente e il risultato ottenuto è stato diviso per il numero di recensioni in cui il topic i-esimo è prevalente.
Il topic 7 (Cancellazione volo) ha la percentuale di prevalenza media maggiore (circa 35,88%) ed è prevalente in 1153 recensioni (5,63% del dataset) ed ha una percentuale θ_{medio}=4,47% mentre il topic 6 (Comodità viaggio) ha la percentuale di prevalenza media inferiore (23,97%) ed è discusso in modo prevalente in 683 recensioni (circa 3,34% del totale) e ha un θ_{medio}=5,81%. Il topic 21 (Competenze personale (2)) nonostante sia il topic prevalentemente discusso (3206 recensioni) e abbia θ_{medio}=10,59 % ha una percentuale di prevalenza media in figura 23 pari a 29,76 % che risulta essere nella media perché il numero di recensioni per cui viene divisa la somma intermedia ottenuta è superiore rispetto alla somma intermedia del topic 6.
Il topic 6 è prevalente e quindi ha un massimo in 683 recensioni e la somma di questi massimi viene divisa per 683 mentre la somma dei massimi del topic 21 (954,0339) viene divisa per 3206 e questo spiega il perché nonostante il topic 21 sia il più discusso e quello con il θ_{medio} maggiore risulta avere un valore di media percentuale di prevalenza nella media.

Il topic 25 (Problemi riscontrati con la compagnia) è il meno discusso e ha il θ_{medio} inferiore e ha una percentuale di prevalenza media nel campione di recensioni in cui tale topic ha il θ massimo tra le più basse (25,35%).

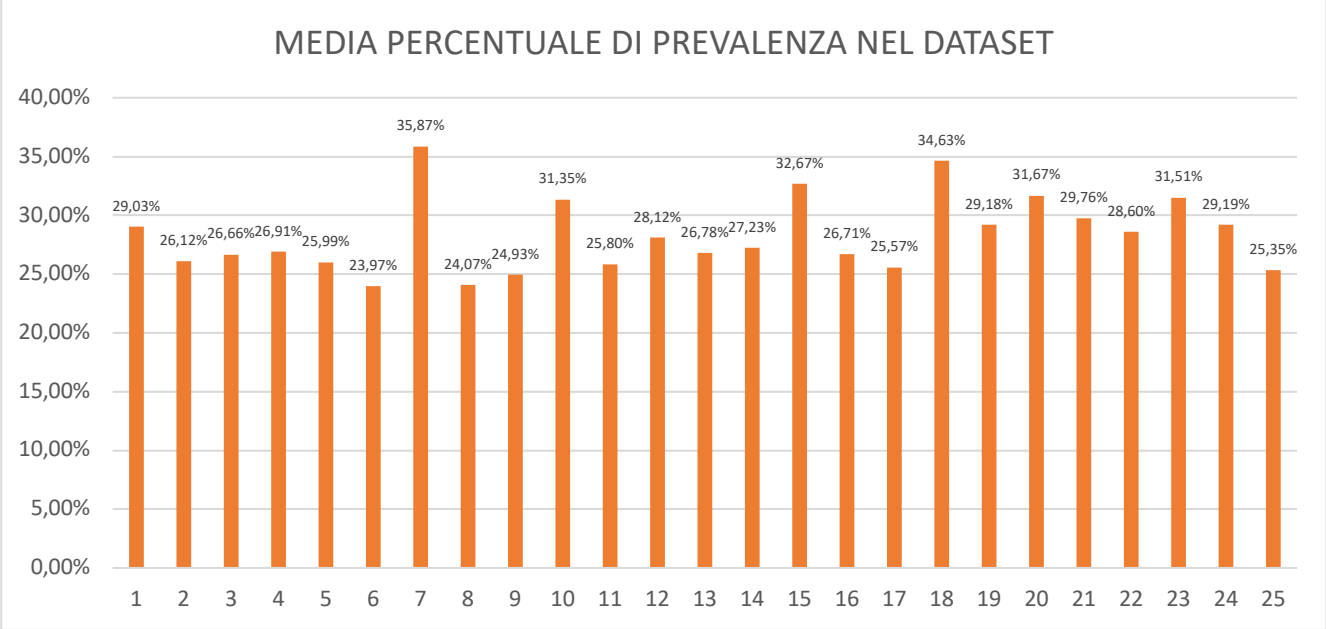


Figura 23. *Grafico media percentuale di prevalenza di ciascun topic nel dataset.*

3.2. ANALISI SETTORIALE.

Le recensioni raccolte nel dataset sulle 10 compagnie aeree considerate sono state suddivise in base alla compagnia a cui sono riferite.
Le compagnie aeree che costituiscono il campione di riferimento appartengono a segmenti di mercato differenti e ad aree geografiche differenti ed è opportuno analizzare i dati e trarre le conclusioni in base al segmento di mercato ed alla posizione geografica a cui ciascuna compagnia appartiene.
Sono stati considerati tre segmenti di mercato del trasporto aereo in cui collocare le 10 compagnie aeree:

- Medio Oriente: Etihad Airways, Emirates, Qatar Airways sono le principali compagnie aeree del Medio Oriente e sono definite le tre Big del Golfo.
  Queste compagnie aeree sono famose in tutto il mondo per gli extra e i vantaggi che offrono al cliente durante il volo (es. fiori freschi, la possibilità di fare una doccia a 30.000 piedi di altezza, lenzuola di seta Frette o i kit di Giorgio Armani dati ai clienti, ecc.).
  Queste tre compagnie vengono definite “super connettori” tra Europa e Asia.
  Mentre le compagnie aeree nazionali più tradizionali puntano soprattutto sui clienti che partono e tornano nel paese dove hanno la loro sede, i super connettori fanno arrivare i passeggeri nei loro hub e poi li smistano sui voli a lunga percorrenza verso le destinazioni finali.
  Queste tre compagnie offrono al cliente oltre al servizio base, ossia il volo, un’ulteriore esperienza a bordo.

- Europa: le compagnie che fanno parte di questa area geografica e coprono la maggior parte delle tratte europee oltre ad alcune tratte intercontinentali sono: Lufthansa, Alitalia, British Airways, Air France e Iberia.

- Europa low-cost: le compagnie che fanno parte di questo segmento di mercato sono Easyjet e Ryanair. Queste compagnie basano la propria strategia di mercato sul fornire al cliente voli low-cost, mentre mettere un bagaglio nella stiva e altri extra sono a pagamento.
  Easyjet e Ryanair sono due dei principali players in questo segmento di mercato e competono sul prezzo, sulle tratte internazionali, sulle tratte a corto e medio raggio, sugli aeroporti in cui sono presenti, sul numero di aerei, sul prezzo dei servizi aggiuntivi, ecc.

Oltre ad effettuare un’analisi su queste tre aree geografiche si è analizzata singolarmente Easyjet e Ryanair per poter capire, oltre all’analisi del segmento low cost in cui queste compagnie aeree operano, le differenze tra i due modelli di business planning attraverso la Text Mining Analysis.

3.2.1. RISULTATI MEDIO ORIENTE.

Le tre compagnie aeree appartenenti a questo segmento di mercato sono: Emirates, Qatar Airways e Etihad Airways.
Le recensioni totali riferite a Emirates dopo la fase di preprocessing sono 1800 (circa 8,8% del recensioni rimaste), quelle appartenenti a Qatar Airways sono 1550 (circa 7,6% delle recensioni rimaste) e quelle che si riferiscono a Etihad Airways sono 600 (circa 3% delle recensioni rimaste). In totale queste tre compagnie aeree sono associate a 3950 recensioni (circa il 19,4%) delle recensioni totali dopo la fase di preprocessing.
La tabella 12 mostra che il topic con la percentuale di prevalenza media maggiore è il topic 23 (Intrattenimento) ed è prevalente in 1338 recensioni (circa il 33,87% delle recensioni del campione Medio Oriente). Il topic 23 ha una deviazione standard dei valori di θ pari a 13,84% e una varianza pari a 1,91%.
Seguono il topic 23, il topic 12 (Qualità servizi (2)) e il topic 21 (Competenze personale (2)) rispettivamente con θ_{medio} = 12,33% e θ_{medio} = 9,77%. Il topic 12 è prevalente in 682 recensioni (circa il 17,26%) e il topic 21 è prevalente in 398 recensioni.
La tabella 12 mostra che il topic con il θ_{medio} più basso è il topic 18 (Bagaglio) con θ_{medio}=0,79%, dev. standard pari a 2,21%, varianza pari a 0,04% e prevalente in appena 20 recensioni (circa 0,50%).
Il topic meno prevalente della categoria Medio Oriente è il topic 25 (Problemi riscontrati con la compagnia aerea), prevalente in una sola recensione.
I dati della tabella sottostante confermano che in questo segmento di mercato per le tre compagnie considerate i principali argomenti trattati sono i servizi di intrattenimento durante il volo (es. musica, film, aperitivi, riviste, ecc.) , la qualità dei servizi forniti e le competenze e la cordialità del personale.
Gli argomenti che non vengono trattati o discussi in minima parte dai clienti quando rilasciano una recensione su una di queste tre compagnie sono: problemi riscontrati con la compagnia, il rapporto qualità/prezzo, anticipo / ritardo del volo, cancellazione volo, snack e bevande, ecc.

Tabella 12. *Risultati di prevalenza media, dev. standard, varianza, numero di recensioni sul totale in cui il topic i-esimo è prevalente, percentuale sul totale per le recensioni dell'area Medio Oriente.*

	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
Prevalenza media del topic i-esimo sul totale delle recensioni (3950)	1,15%	5,01%	2,10%	2,96%	3,13%
Deviazione standard	2,46%	6,31%	4,16%	4,93%	4,37%
Varianza	0,06%	0,39%	0,17%	0,24%	0,19%
Numero di recensioni sul totale in cui il topic i-esimo è prevalente	16	161	51	80	70

Percentuale sul totale delle recensioni (3950)	0,40%	4,07%	1,29%	2,02%	1,77%
	Topic 6	**Topic 7**	**Topic 8**	**Topic 9**	**Topic 10**
Prevalenza media del topic i-esimo sul totale delle recensioni(3950)	7,27%	2%	4,57%	1,60%	1,43%
Deviazione standard	5,67%	5,41%	4,76%	2,72%	2,80%
Varianza	0,32%	0,29%	0,22%	0,07%	0,07%
Numero di recensioni sul totale in cui il topic i-esimo è prevalente	159	86	89	27	24
Percentuale sul totale delle recensioni(3950)	4,02%	2,17%	2,25%	0,68%	0,60%
	Topic 11	**Topic 12**	**Topic 13**	**Topic 14**	**Topic 15**
Prevalenza media del topic i-esimo sul totale delle recensioni(3950)	5,80%	12,33%	1,81%	4,61%	1,58%
Deviazione standard	7,18%	10,68%	3,01%	5,98%	4,15%
Varianza	0,51%	1,14%	0,09%	0,35%	0,17%
Numero di recensioni sul totale in cui il topic i-esimo è prevalente	227	682	29	150	58
Percentuale sul totale delle recensioni (3950)	5,74%	17,26%	0,73%	3,79%	1,46%
	Topic 16	**Topic 17**	**Topic 18**	**Topic 19**	**Topic 20**
Prevalenza media del topic i-esimo sul totale delle recensioni(3950)	2,03%	4,41%	0,79%	0,91%	2,44%
Deviazione standard	3,69%	4,99%	2,21%	1,73%	5%
Varianza	0,13%	0,24%	0,04%	0,03%	0,25%
Numero di recensioni sul totale in cui il	54	92	20	10	86

topic i-esimo è prevalente					
Percentuale sul totale delle recensioni (3950)	1,36%	2,32%	0,50%	0,25%	2,17%
	Topic 21	**Topic 22**	**Topic 23**	**Topic 24**	**Topic 25**
Prevalenza media del topic i-esimo sul totale delle recensioni(3950)	9,77%	1,68%	17,93%	1,25%	1,3%
Deviazione standard	8,20%	2,10%	13,84%	2,64%	0,55%
Varianza	0,67%	0,04%	1,91%	0,06%	0,00306%
Numero di recensioni sul totale in cui il topic i-esimo è prevalente	398	14	1338	28	1
Percentuale sul totale delle recensioni (3950)	10,07%	0,35%	33,87%	0,70%	0,02%

Dopo aver effettuato tale analisi si è costruito un istogramma che rappresentasse la media percentuale di prevalenza per ogni topic considerando solo le recensioni per cui il topic i-esimo è prevalente (figura 24).
Tutti i valori di media percentuale di prevalenza sono compresi nell'intervallo [13,25% , 33,27%]. Il topic che possiede la media percentuale di prevalenza più elevata è il topic 23 (Intrattenimento) con θ_{medio}= 33,27% che indica che tale topic non solo è quello con la prevalenza media più elevata sul totale delle recensioni del segmento di mercato Medio Oriente, ma ha anche il valore di prevalenza media più elevato considerando solo le recensioni in cui il topic 23 è prevalente (1338).
Il topic con la prevalenza media più bassa rispetto al numero di recensioni in cui è prevalente è il topic 25 (Problemi riscontrati con la compagnia aerea) con un θ_{medio}= 13,25%, e questo è dovuto al fatto che il topic 25 risulta essere prevalente solo in una recensione sui 3950 documenti con il valore 13,25 che diviso per 1 restituisce il medesimo risultato.
Il topic 7 (Cancellazione volo) che è prevalente in 86 recensioni ha un θ_{medio}= 31,08%, di poco inferiore a quella del topic 23.
Questo dato mostra che nonostante il topic 7 non sia tra i più discussi sul totale delle recensioni (prevalente in 86 recensioni) e abbia un θ_{medio} basso (2 %) ha dei valori di prevalenza elevati all'interno delle 86 recensioni in cui è prevalente e di conseguenza un θ_{medio} su quelle 86 recensioni molto elevato rispetto agli altri topic.

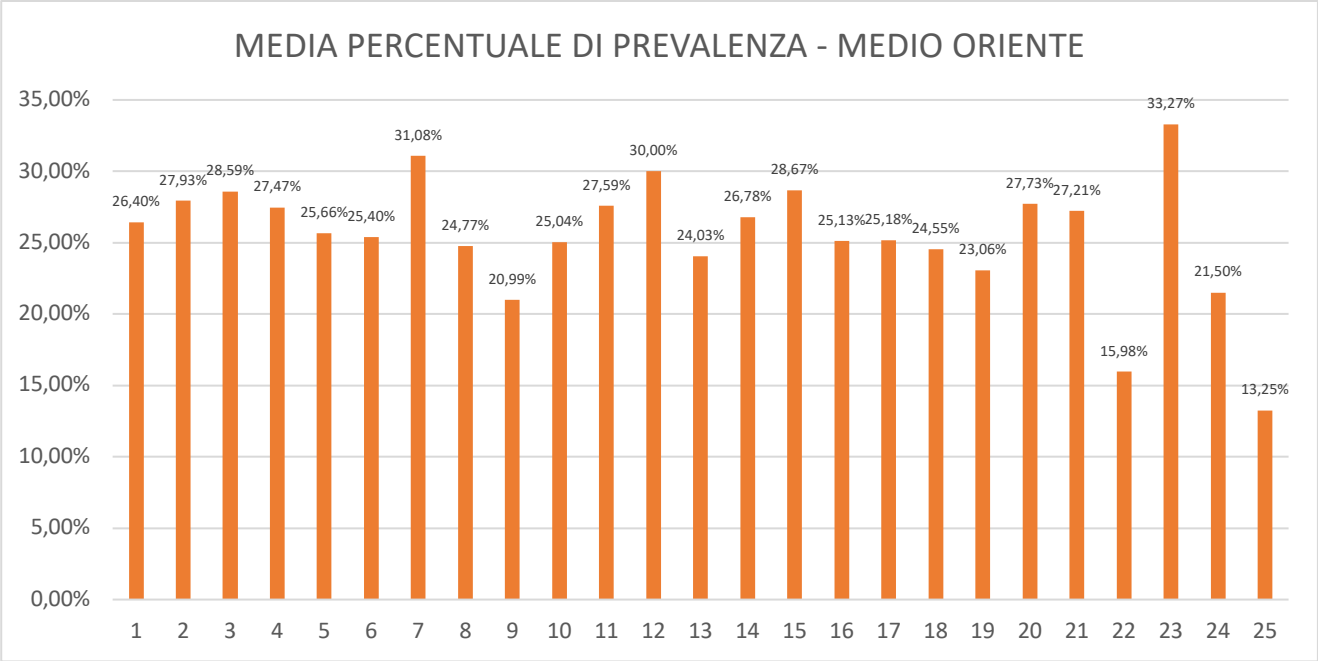


Figura 24. *Grafico media percentuale di prevalenza – Medio Oriente.*

3.2.2. RISULTATI EUROPA

Le compagnie aeree appartenenti a questa sezione sono: Lufthansa , Alitalia, Air France, British Airways e Iberia. Il totale delle recensioni di questa sezione è 9498.
Le recensioni che si riferiscono a Lufthansa dopo la fase di preprocessing sono 2500 (circa il 26,32%), quelle su Alitalia sono 2998 (circa il 31,56%), quelle su Air France sono 1400 (circa il 14,73%), quelle su British Airways sono 1600 (circa il 16,84%) e quelle su Iberia sono 1000 (circa il 10,52).
La Tabella 13 mostra che il topic con la prevalenza media più elevata sul campione di recensioni considerate è il topic 21 (Competenze personale (2)) con θ_{medio}=11,36%, deviazione standard pari a 10,94% , varianza 1,20% e prevalente 1774 recensioni (circa il 18,68% delle recensioni del campione).
Seguono il topic 23 (Intrattenimento) con θ_{medio}=6,60% e prevalente in 794 recensioni e il topic 6 (Comodità viaggio) con θ_{medio}=5,89% e prevalente in 362 recensioni.
Il topic meno discusso è il topic 25 (Problemi riscontrati con la compagnia aerea) con θ_{medio}=1,66% e prevalente in appena 4 recensioni. Tra i topic con la prevalenza più bassa e prevalenti in poche recensioni del campione vi sono il topic 3 (Famiglia) con θ_{medio}=1,84% e prevalente in 108 recensioni e il topic 9 (Snack & bevande) con θ_{medio}=2,09 % e prevalente in 101 recensioni.
In questo campione si può notare che gli argomenti più discussi e prevalenti sono l'intrattenimento, le competenze del personale e la comodità del viaggio mentre gli argomenti meno discussi all'interno delle recensioni sono la presenza di famiglie a bordo e commenti da parte di clienti con una famiglia, problemi riscontrati con la compagnia aerea e il consumo di snack e bevande a bordo durante il volo.

Tabella 13. *Risultati di prevalenza media, dev. standard, varianza, numero di recensioni sul totale in cui il topic i-esimo è prevalente, percentuale sul totale per le recensioni dell'area Europa.*

	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
Prevalenza media del topic i-esimo sul totale delle recensioni (9498)	2,35%	4,87%	1,84%	4,63%	4,01%
Deviazione standard	4,50%	6,26%	3,53%	6,97%	5,75%
Varianza	0,20%	0,39%	0,12%	0,49%	0,33%
Numero di recensioni sul totale in cui il topic i-esimo è prevalente	155	485	108	527	334
Percentuale sul totale delle recensioni (9498)	1,63%	5,11%	1,14%	5,55%	3,52%
	Topic 6	**Topic 7**	**Topic 8**	**Topic 9**	**Topic 10**
Prevalenza media del topic i-esimo sul totale delle recensioni(9498)	5,89%	4,90%	3,79%	2,09%	2,98%
Deviazione standard	5,26%	9,57%	4,70%	3,64%	5,51%
Varianza	0,28%	0,91%	0,22%	0,13%	0,30%
Numero di recensioni sul totale in cui il topic i-esimo è prevalente	362	645	251	101	224
Percentuale sul totale delle recensioni(9498)	2,36%	2,36%	2,36%	2,36%	2,36%
	Topic 11	**Topic 12**	**Topic 13**	**Topic 14**	**Topic 15**
Prevalenza media del topic i-esimo sul totale delle recensioni(9498)	2,44%	5,26%	2,79%	4,57%	3,54%
Deviazione standard	4,05%	6,04%	4,21%	6,66%	7,16%
Varianza	0,16%	0,36%	0,18%	0,44%	0,51%

Numero di recensioni sul totale in cui il topic i-esimo è prevalente	169	397	192	435	373
Percentuale sul totale delle recensioni (9498)	1,78%	4,18%	2,02%	4,58%	3,93%
	Topic 16	**Topic 17**	**Topic 18**	**Topic 19**	**Topic 20**
Prevalenza media del topic i-esimo sul totale delle recensioni(9498)	2,97%	4,50%	2,55%	2,68%	3,59%
Deviazione standard	5,09%	5,97%	5,44%	5,00%	6,99%
Varianza	0,26%	0,36%	0,30%	0,25%	0,49%
Numero di recensioni sul totale in cui il topic i-esimo è prevalente	235	381	244	292	353
Percentuale sul totale delle recensioni (9498)	2,47%	4,01%	2,57%	3,07%	3,72%
	Topic 21	**Topic 22**	**Topic 23**	**Topic 24**	**Topic 25**
Prevalenza media del topic i-esimo sul totale delle recensioni(9498)	11,36%	5,25%	6,60%	2,90%	1,66%
Deviazione standard	10,94%	6,49%	8,26%	5,22%	0,89%
Varianza	1,20%	0,42%	0,68%	0,27%	0,01%
Numero di recensioni sul totale in cui il topic i-esimo è prevalente	1774	427	794	236	4
Percentuale sul totale delle recensioni (9498)	18,68%	4,50%	8,36%	2,48%	0,04%

Successivamente si è costruito un istogramma che rappresentasse la media percentuale di prevalenza per ogni topic considerando solo le recensioni per cui il topic i-esimo è prevalente (Figura 25).

Tutti i valori di media percentuale di prevalenza sono compresi nell'intervallo [23,40% , 34,55%]. Il topic che possiede la media percentuale di prevalenza più elevata è il topic 7 (Cancellazione volo) con θ_{medio}= 34,55%. Il topic con la prevalenza media più bassa rispetto al numero di recensioni in cui è prevalente è il topic 6 (Comodità viaggio) con un θ_{medio}= 23,40%. Il topic 25 (Problemi riscontrati con la compagnia aerea) che nelle precedenti analisi presentava il valore di θ_{medio} più basso o tra i più bassi rispetto al numero di recensioni in cui è prevalente in questa analisi ha un θ_{medio} tra i più elevati e nonostante sia prevalente solo in 4 recensioni, questo indica che i valori di massimo nelle 4 recensioni sono elevati e significa che in queste 4 recensioni il topic 25 è prevalente e il più discusso.

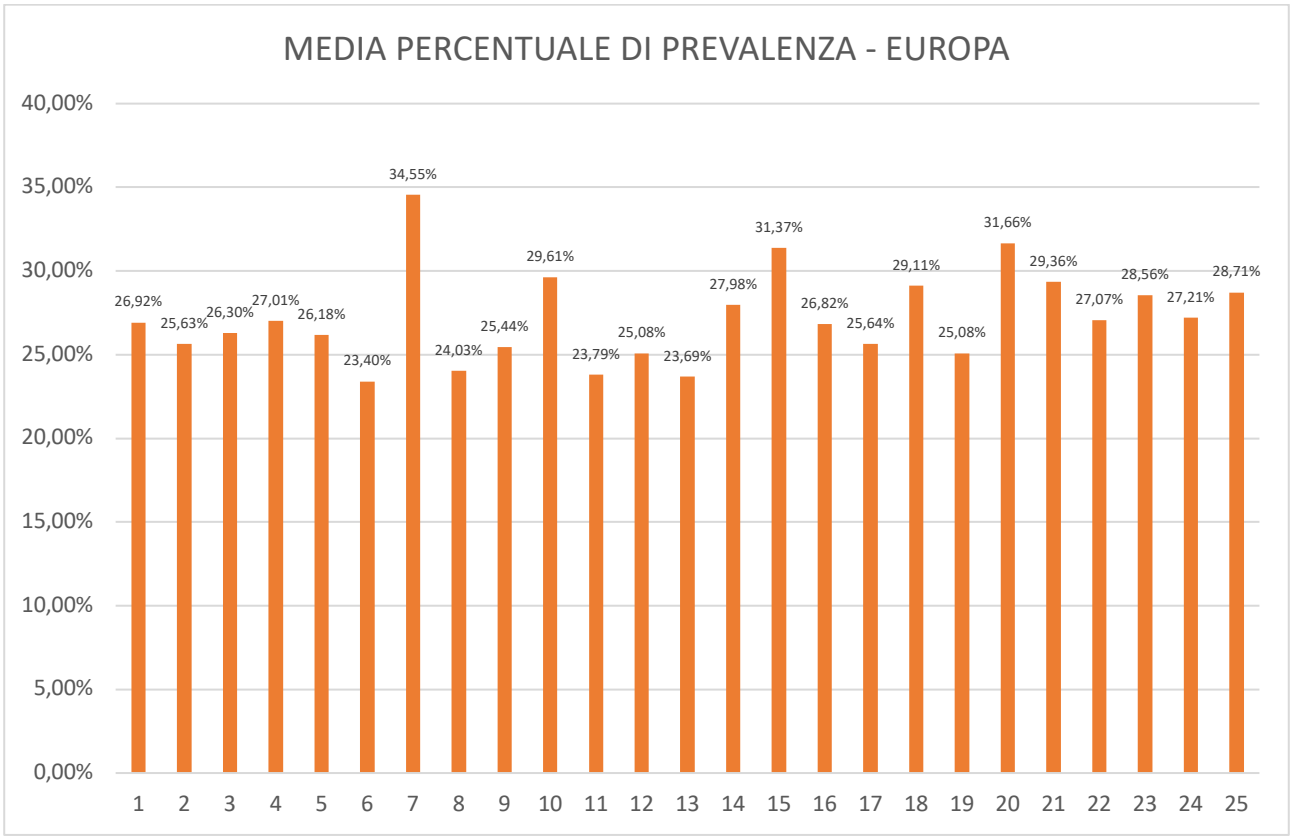


Figura 25. *Grafico media percentuale di prevalenza – Europa.*

3.2.3. RISULTATI EUROPA – LOW COST.

A questa categoria appartengono due compagnie aeree del dataset di riferimento: Ryanair (2999 recensioni dopo la fase di preprocessing) e Easyjet (3999 recensioni dopo la fase di preprocessing). Il totale delle recensioni del campione Europa low-cost è 6998.
La tabella 14 mostra che il topic con la prevalenza media sul totale delle recensioni più elevata è il topic 18 (Bagaglio) con θ_{medio}= 11,21%, deviazione standard pari a 13,81 e varianza pari a 1,91%; il topic 18 è il più discusso in 1239 recensioni (circa il 17,71% del totale delle recensioni del campione Europa – Low Cost).
Seguono il topic 18 il topic 21 (Competenze personale (2)) con θ_{medio}=10,02% e il più discusso in 1034 recensioni (circa il 14,78% del recensioni del campione) e il topic 22 (Anticipo / Ritardo volo) con θ_{medio}= 9,40% e il più discusso in 784 recensioni (circa l'11,20% del totale delle recensioni del campione.
I topic meno discussi sono rappresentati dal topic 4 (Pasti a bordo) con θ_{medio}= 1,13 % e prevalente in 27 recensioni (circa lo 0,39% delle recensioni), il topic 5 (Coincidenza volo) con θ_{medio}=1,07 % e il più discusso in 24 recensioni (circa lo 0,34% delle recensioni) e il topic meno discusso è il topic 23 (Intrattenimento) con θ_{medio}=0,54% e prevalente in una sola recensione (circa 0,01% delle recensioni).
Questa analisi mostra che nel segmento di mercato low cost gli argomenti più discussi e dove viene posta la maggior parte dell'attenzione del cliente nel momento in cui esso rilascia una recensione sono sui bagagli, principalmente sul bagaglio a mano e sul costo aggiuntivo per poter mettere un bagaglio nella stiva, sulle competenze del personale a bordo e sull'eventuale anticipo / ritardo del volo.
I topic che vengono meno discussi dai clienti che volano con voli low cost sono il consumo di pasti a bordo, la coincidenza con un volo da prendere dopo aver fatto scalo in un aeroporto e l'argomento meno discusso dalla clientela è l'intrattenimento a bordo e la presenza e l'utilizzo di extra a bordo.

Tabella 14. *Risultati di prevalenza media, dev. standard, varianza, numero di recensioni sul totale in cui il topic i-esimo è prevalente, percentuale sul totale per le recensioni dell'area Europa Low Cost.*

	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
Prevalenza media del topic i-esimo sul totale delle recensioni (6998)	4,49%	1,62%	2,15%	1,13%	1,07%
Deviazione standard	6,97%	2,49%	3,72%	2,19%	1,90%
Varianza	0,49%	0,06%	0,14%	0,05%	0,04%
Numero di recensioni sul totale in cui il topic i-esimo è prevalente	279	30	86	27	24

Percentuale sul totale delle recensioni (6998)	3,99%	0,43%	1,23%	0,39%	0,34%
	Topic 6	**Topic 7**	**Topic 8**	**Topic 9**	**Topic 10**
Prevalenza media del topic i-esimo sul totale delle recensioni(6998)	4,89%	5,29%	2,70%	2,03%	5,23%
Deviazione standard	4,68%	10,06%	3,50%	3,36%	8,12%
Varianza	0,22%	1,01%	0,12%	0,11%	0,66%
Numero di recensioni sul totale in cui il topic i-esimo è prevalente	162	422	73	56	366
Percentuale sul totale delle recensioni(6998)	2,31%	6,03%	1,04%	0,80%	5,23%
	Topic 11	**Topic 12**	**Topic 13**	**Topic 14**	**Topic 15**
Prevalenza media del topic i-esimo sul totale delle recensioni(6998)	1,24%	1,27%	5,60%	2,28%	5,36%
Deviazione standard	2,38%	1,79%	7,15%	3,65%	8,68%
Varianza	0,06%	0,03%	0,51%	0,13%	0,75%
Numero di recensioni sul totale in cui il topic i-esimo è prevalente	31	10	372	85	365
Percentuale sul totale delle recensioni (6998)	0,44%	0,14%	5,32%	1,21%	5,22%
	Topic 16	**Topic 17**	**Topic 18**	**Topic 19**	**Topic 20**
Prevalenza media del topic i-esimo sul totale delle recensioni(6998)	1,56%	3,17%	11,21%	8,63%	3,37%
Deviazione standard	3,29%	4,88%	13,81%	10,43%	6,47%
Varianza	0,11%	0,24%	1,91%	1,09%	0,42%
Numero di recensioni sul totale in cui il	52	158	1239	894	185

topic i-esimo è prevalente					
Percentuale sul totale delle recensioni (6998)	0,74%	2,26%	17,71%	12,78%	2,64%
	Topic 21	**Topic 22**	**Topic 23**	**Topic 24**	**Topic 25**
Prevalenza media del topic i-esimo sul totale delle recensioni(6998)	10,02%	9,40%	0,54%	4,31%	1,42%
Deviazione standard	11,08%	9,40%	0,65%	7,04%	0,62%
Varianza	1,23%	0,88%	0,004%	0,50%	0,004%
Numero di recensioni sul totale in cui il topic i-esimo è prevalente	1034	784	1	262	1
Percentuale sul totale delle recensioni (6998)	14,78%	11,20%	0,01%	3,74%	0,01%

In seguito è stato possibile creare un istogramma che rappresentasse la media percentuale di prevalenza per ogni topic considerando solo le recensioni per cui il topic i-esimo è prevalente (Figura 26).
Tutti i valori del grafico sono compresi nell'intervallo [19,51% , 38,86%].
Il topic con θ_{medio} più elevato è il topic 7 (Cancellazione volo) che ha un θ_{medio}= 38,86 % seguito dal topic 18 (Bagaglio) con θ_{medio}= 35,88% e dal topic 15 (Chech-in) con θ_{medio}= 34,64%.
Questi sono i topic con la prevalenza media più elevata sul totale delle recensioni in cui sono prevalenti.
I topic con la percentuale di prevalenza media inferiore considerando le recensioni per cui il topc i-esimo è prevalente sono il topic 12 (Qualità dei servizi (2)) con θ_{medio}= 20,01% e il topic 23 (Intrattenimento) con $\theta_{medio=}$19,51%.

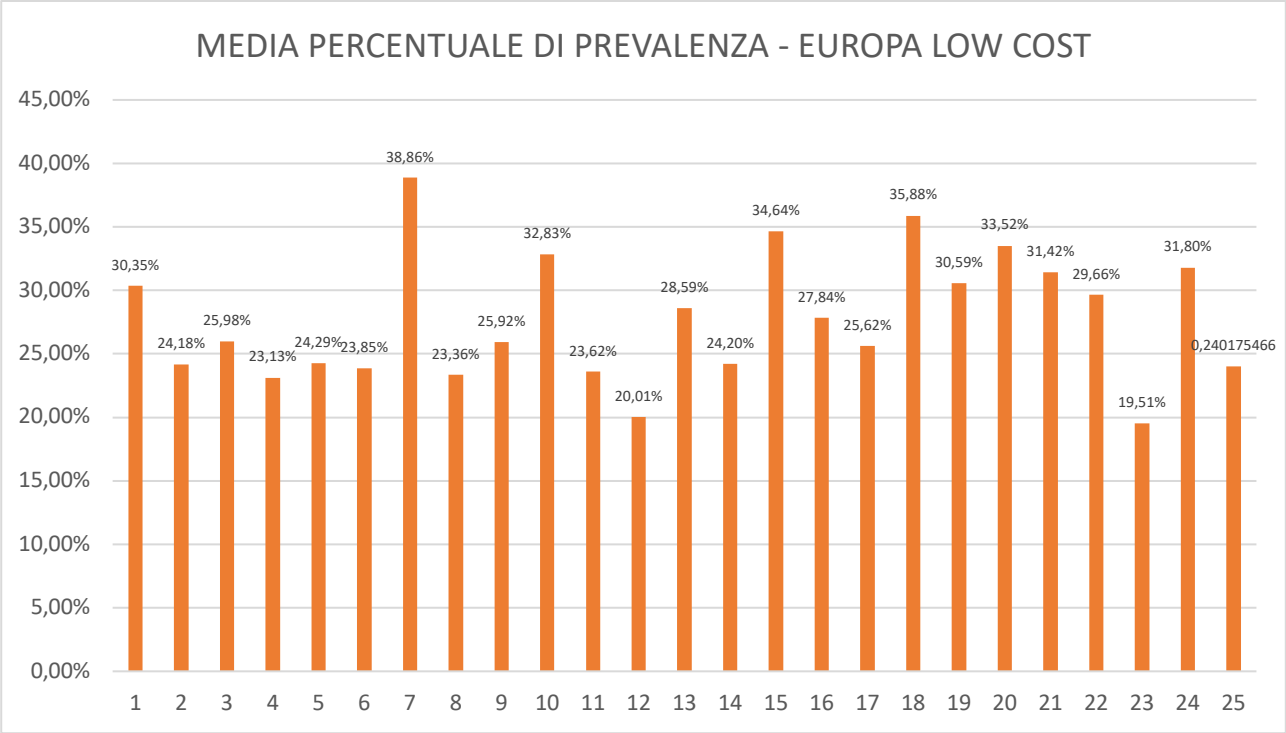


Figura 26. *Grafico media percentuale di prevalenza – Europa Low Cost.*

3.2.4. ANALOGIE E DIFFERENZE TRA IL MODELLO RYANAIR E IL MODELLO EASYJET

Il modello Easyjet è una valida alternativa alla potente rivale Ryanair. La sua impostazione strategica presenta delle peculiarità che in parte la discostano dal rigido schema delle low cost airlines, manifestando una particolare interpretazione del modello che mantiene ad ogni modo i principi fondamentali.
Una prima fondamentale differenza fra i due modelli deriva dalla diversità di scelte per quanto riguarda il network aeroportuale. Entrambe fanno capo alla rete point to point, scelta classica delle compagnie low cost, ma Ryanair aggiunge sempre nuove rotte, mentre Easyjet punta al rafforzamento delle rotte esistenti con l'aggiunta di nuove collegamenti fra i vari aeroporti già appartenenti al network.
La strategia di sviluppo della rete di Easyjet è costituita da due punti fondamentali: il primo è l'aggiunta di nuovi collegamenti tra gli aeroporti operativi, il secondo è l'incremento della frequenza dei voli sulle rotte più importati e profittevoli. Questo modo di operare è diverso dalla strategia di Ryanair, la quale concentra la propria attenzione sulla rapida crescita del numero di aeroporti serviti, al fine di sfruttare le opportunità che le si presentano davanti e per garantire alla clientela un insieme di rotte sempre più variegato e che comincia ad espandersi sempre più anche all'esterno dei confini europei.
Ma la caratteristica che più contraddistingue la diversità della filosofia low cost è sicuramente data dall'utilizzo di scali di primaria importanza, con i conseguenti maggiori costi che ne derivano. EasyJet si presenta sul mercato come una compagnia il cui target di

clientela non viene circoscritto al solo segmento dei viaggi di piacere; l'idea era quella di poter beneficiare della crescente attenzione ai costi di quelle aziende che hanno bisogno di utilizzare il servizio di trasporto aereo per i propri dipendenti. Proprio per venire incontro a queste esigenze, Easyjet ha scelto di volare su aeroporti principali (come Amsterdam, Parigi, Ginevra, Milano, Barcellona), posti ad una maggiore vicinanza dal centro cittadino. Il trattamento che viene riservato ai passeggeri business di Easyjet rimane uguale a quello di tutti gli altri e fedele alla filosofia low cost: pasti a bordo a pagamento, spazi più ristretti, nessuna distinzione fra classi e nessuna discriminazione tariffaria. I prezzi in media sono leggermente più elevati rispetto a Ryanair a causa della maggiore consistenza dei costi aeroportuali, questo comunque non influenza le aspettative dei passeggeri business, disposti a pagare qualcosa in più per volare verso aeroporti più centrali.
Anche la strategia tariffaria rappresenta una differenza tra i due modelli. In Easyjet vi è un diverso orientamento nella massimizzazione del profitto dalle attività di volo. Ryanair mantiene un margine di profitto unitario più basso al fine di continuare ad accrescere il numero di passeggeri, variabile fondamentale per la sua struttura operativa e per mantenere i vantaggi delle economie di scala. Easyjet, al contrario, punta a dei margini più elevati cercando di massimizzare il guadagno per passeggero trasportato. Si tratta di un atteggiamento differente che però non intacca la strategia delle tariffe, semplice e flessibile come quella implementata da Ryanair.
Nonostante queste differenze, entrambe, hanno ottenuto successi e profitti che hanno costretto le più grandi compagnie aeree del mondo a modificare la propria filosofia, ma che in ogni caso non sono più in grado di competere con questi players nel segmento low cost.

3.2.5. RYANAIR VS EASYJET

E' stata compiuta un'analisi su Ryanair e Easyjet considerandole separatamente per poter stabilire tramite la text mining analysis analogie e differenze tra il modello Ryanair e il modello Easyjet.
La prima compagnia analizzata è stata Ryanair. La tabella 15 mostra che il topic con la prevalenza media più elevata sul totale delle recensioni è il topic 19 (Rapporto qualità / prezzo (2)) con un θ_{medio}= 11,07%, deviazione standard pari a 11,84% e varianza pari a 1,40%. Il topic 19 è il più discusso in 551 recensioni (circa il 18,37 % delle recensioni di Ryanair).
Gli altri topic con θ_{medio} più elevato e più discussi all'interno delle recensioni sono il topic 18 (Bagaglio) e il topic 22 (Anticipo / Ritardo volo) rispettivamente con θ_{medio}= 9,39% e θ_{medio}= 9,11% e i più discussi rispettivamente in 407 (circa il 13,57%) e in 345 (circa l'11,50%) delle recensioni di Ryanair.
I tre topic con la prevalenza media più bassa sono il topic 11 (Manutenzione e pulizia) con θ_{medio}= 1,01% , il topic 12 (Qualità servizi (2)) con θ_{medio}= 0,76% e il topic con la prevalenza media inferiore è il topic 23 (Intrattenimento) con θ_{medio}=0,47%.
Il topic 23 (Intrattenimento) e il topic 25 (Problemi riscontrati con la compagnia aerea) sono i due topic che non sono prevalenti in nessuna recensione.
Gli argomenti più discussi rilasciati da clienti che hanno volato con Ryanair sono il rapporto qualità / prezzo, la possibilità di mettere uno o più bagagli nella stiva e il costo che si paga per questo servizio, il bagaglio a mano e la puntualità di partenze e arrivi dei voli.

Tabella 15. *Risultati di prevalenza media, dev. standard, varianza, numero di recensioni sul totale in cui il topic i-esimo è prevalente, percentuale sul totale per le recensioni Ryanair.*

	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
Prevalenza media del topic i-esimo sul totale delle recensioni (2999)	6,86%	1,41%	2,19%	1,47%	1,16%
Deviazione standard	8,89%	2,35%	3,59%	2,67%	2,14%
Varianza	0,79%	0,06%	0,13%	0,07%	0,05%
Numero di recensioni sul totale in cui il topic i-esimo è prevalente	227	12	35	20	12
Percentuale sul totale delle recensioni (2999)	7,57%	0,40%	1,17%	0,67%	0,40%
	Topic 6	**Topic 7**	**Topic 8**	**Topic 9**	**Topic 10**
Prevalenza media del topic i-esimo sul totale delle recensioni(2999)	4,33%	5,00%	2,44%	2,17%	5,40%
Deviazione standard	4,19%	9,34%	3,18%	3,41%	8,54%
Varianza	0,18%	0,87%	0,10%	0,12%	0,73%
Numero di recensioni sul totale in cui il topic i-esimo è prevalente	57	153	22	22	170
Percentuale sul totale delle recensioni(2999)	1,90%	5,10%	0,73%	0,73%	5,67%
	Topic 11	**Topic 12**	**Topic 13**	**Topic 14**	**Topic 15**
Prevalenza media del topic i-esimo sul totale delle recensioni(2999)	1,01%	0,76%	6,66%	1,86%	5,62%
Deviazione standard	1,85%	1,11%	7,95%	3,05%	8,83%
Varianza	0,03%	0,01%	0,63%	0,09%	0,78%
Numero di recensioni sul	9	2	205	26	168

totale in cui il topic i-esimo è prevalente					
Percentuale sul totale delle recensioni (2999)	0,30%	0,07%	6,84%	0,87%	5,60%
	Topic 16	**Topic 17**	**Topic 18**	**Topic 19**	**Topic 20**
Prevalenza media del topic i-esimo sul totale delle recensioni(2999)	1,63%	3,19%	9,39%	11,07%	3,93%
Deviazione standard	3,18%	4,98%	10,96%	11,84%	7,44%
Varianza	0,10%	0,25%	1,20%	1,40%	0,55%
Numero di recensioni sul totale in cui il topic i-esimo è prevalente	20	68	407	551	106
Percentuale sul totale delle recensioni (2999)	0,67%	2,27%	13,57%	18,37%	3,53%
	Topic 21	**Topic 22**	**Topic 23**	**Topic 24**	**Topic 25**
Prevalenza media del topic i-esimo sul totale delle recensioni(2999)	6,74%	9,11%	0,47%	4,74%	1,40%
Deviazione standard	8,07%	9,70%	0,51%	7,29%	0,53%
Varianza	0,65%	0,94%	0,00%	0,53%	0,00%
Numero di recensioni sul totale in cui il topic i-esimo è prevalente	222	345	0	140	0
Percentuale sul totale delle recensioni (2999)	7,40%	11,50%	0,00%	4,67%	0,00%

In seguito è stato possibile fare un grafico che rappresentasse la media percentuale di prevalenza per ogni topic considerando solo le recensioni per cui il topic i-esimo è prevalente (Figura 27).
Si è notato che tutti i valori dell'istogramma sono compresi nell'intervallo [0% , 38,95%].
Il topic 23 e il topic 25 hanno valore zero perché, come sottolineato in precedenza, non sono i più discussi in alcuna recensione.

Il topic con il θ_{medio} più elevato è il topic 7 (Cancellazione volo) con θ_{medio}= 38,95% seguito dal topic 20 (Prenotazione volo on-line) e dal topic 15 (Check-in) rispettivamente con θ_{medio}= 34,68% e θ_{medio}= 34,59%.

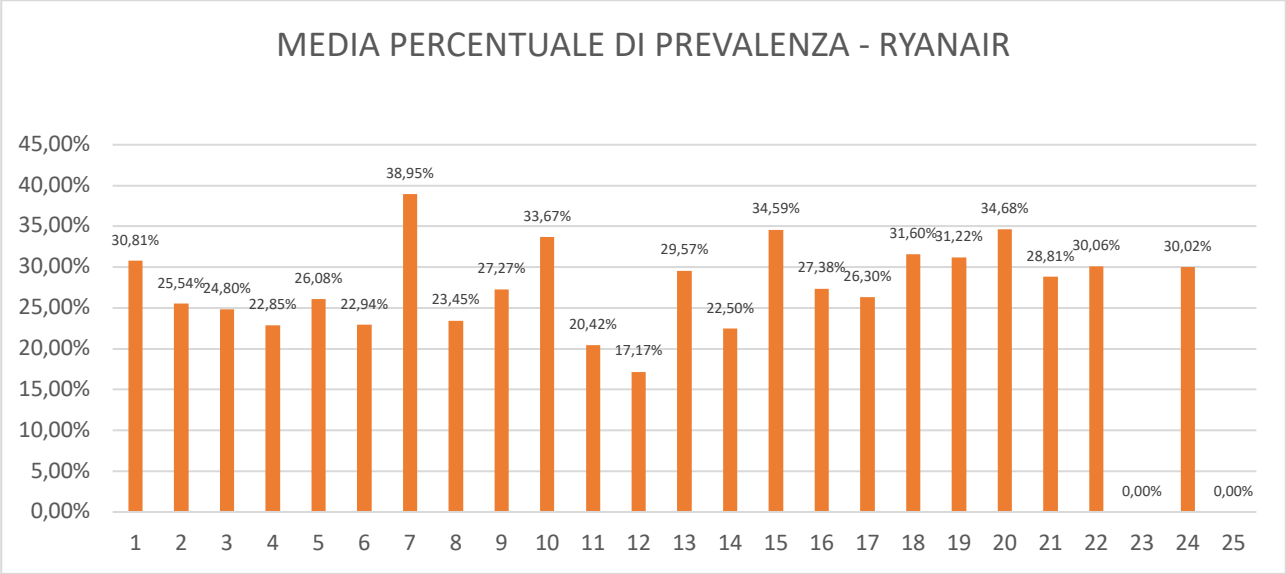


Figura 27. *Grafico media percentuale di prevalenza – Ryanair.*

Successivamente sono state analizzate le 3999 recensioni riferite a Easyjet.
Nella tabella 16 si può notare che nelle recensioni rilasciate dai clienti Easyjet l'argomento con la prevalenza media più elevata è il topic 18 (Bagaglio) con θ_{medio}= 12,58% e il più discusso in 832 recensioni (circa il 20,81% del campione Easyjet).
Gli altri argomenti con θ_{medio} più elevato e più discussi dai clienti di Easyjet sono il topic 21 (Competenze personale (2) e il topic 22 (Anticipo / ritardo volo) rispettivamente con θ_{medio}= 12,49% e θ_{medio}= 9,63%. Il topic 21 è il più discusso in 812 recensioni (circa il 20,31% del campione) mentre il topic 22 è il più discusso dai clienti in 439 recensioni (circa il 10,98% del campione).
Gli argomenti meno discussi dai clienti Easyjet sono il topic 5 (Coincidenza volo) con θ_{medio}= 1% e il più discusso in 12 recensioni (circa lo 0,30% del campione Easyjet) e il topic 4 (Pasti a bordo) con θ_{medio}= 0,88% e prevalente in 7 recensioni (circa lo 0,18%).
L'argomento con θ_{medio} inferiore e il meno prevalente è il topic 23 (Intrattenimento) con θ_{medio}= 0,59% e il più discusso in 1 recensione (circa lo 0,03 %).

Tabella 16. *Risultati di prevalenza media, dev. standard, varianza, numero di recensioni sul totale in cui il topic i-esimo è prevalente, percentuale sul totale per le recensioni Easyjet.*

	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
Prevalenza media del topic i-esimo sul totale delle recensioni (3999)	2,72%	1,77%	2,13%	0,88%	1,00%
Deviazione standard	4,29%	2,57%	3,82%	1,69%	1,70%
Varianza	0,18%	0,07%	0,15%	0,03%	0,03%
Numero di recensioni sul totale in cui il topic i-esimo è prevalente	52	18	51	7	12
Percentuale sul totale delle recensioni (3999)	1,30%	0,45%	1,28%	0,18%	0,30%
	Topic 6	**Topic 7**	**Topic 8**	**Topic 9**	**Topic 10**
Prevalenza media del topic i-esimo sul totale delle recensioni(3999)	5,31%	5,50%	2,90%	1,93%	5,09%
Deviazione standard	4,98%	10,56%	3,71%	3,32%	7,79%
Varianza	0,25%	1,12%	0,14%	0,11%	0,61%
Numero di recensioni sul totale in cui il topic i-esimo è prevalente	105	269	51	34	196
Percentuale sul totale delle recensioni(3999)	2,63%	6,73%	1,28%	0,85%	4,90%
	Topic 11	**Topic 12**	**Topic 13**	**Topic 14**	**Topic 15**
Prevalenza media del topic i-esimo sul totale delle recensioni(3999)	1,42%	1,66%	4,81%	2,59%	5,17%
Deviazione standard	2,70%	2,09%	6,37%	4,01%	8,56%
Varianza	0,07%	0,04%	0,41%	0,16%	0,73%
Numero di recensioni sul	22	8	167	59	197

totale in cui il topic i-esimo è prevalente					
Percentuale sul totale delle recensioni (3999)	0,55%	0,20%	4,18%	1,48%	4,93%
	Topic 16	**Topic 17**	**Topic 18**	**Topic 19**	**Topic 20**
Prevalenza media del topic i-esimo sul totale delle recensioni(3999)	1,51%	3,16%	12,58%	6,81%	2,95%
Deviazione standard	3,37%	4,81%	15,47%	8,80%	5,59%
Varianza	0,11%	0,23%	2,39%	0,77%	0,31%
Numero di recensioni sul totale in cui il topic i-esimo è prevalente	32	90	832	343	79
Percentuale sul totale delle recensioni (3999)	0,80%	2,25%	20,81%	8,58%	1,98%
	Topic 21	**Topic 22**	**Topic 23**	**Topic 24**	**Topic 25**
Prevalenza media del topic i-esimo sul totale delle recensioni(3999)	12,49%	9,63%	0,59%	3,98%	1,44%
Deviazione standard	12,32%	9,16%	0,73%	6,83%	0,68%
Varianza	1,52%	0,84%	0,01%	0,47%	0,00%
Numero di recensioni sul totale in cui il topic i-esimo è prevalente	812	439	1	122	1
Percentuale sul totale delle recensioni (3999)	20,31%	10,98%	0,03%	3,05%	0,03%

In seguito è stato possibile creare un grafico che rappresentasse la media percentuale di prevalenza per ogni topic considerando solo le recensioni per cui il topic i-esimo è prevalente (Figura 28).
Si è notato che tutti i valori dell'istogramma sono compresi nell'intervallo [19,51% , 38,81%].

Il topic con il θ_{medio} più elevato è il topic 7 (Cancellazione volo) con θ_{medio}=38,81% seguito dal topic 18 (Bagaglio) e dal topic 15 (Check-in) rispettivamente con θ_{medio}= 37,98% e θ_{medio}= 34,69%.
I topic con θ_{medio} più basso considerando le recensioni in cui sono prevalenti sono il topic 5 (Coincidenza volo) con θ_{medio}= 22,50%, il topic 12 (Qualità servizi (2)) con θ_{medio}= 20,72% e il topic 23 (Intrattenimento) con θ_{medio}= 19,51%.

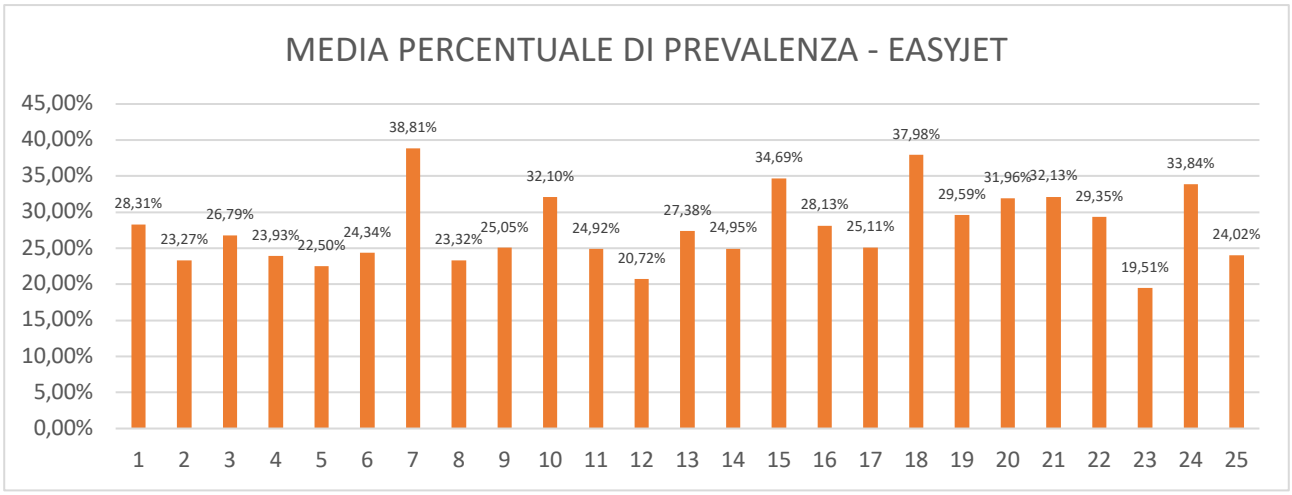


Figura 28. *Grafico media percentuale di prevalenza – Easyjet.*

In conclusione, entrambe le compagnie operano nel segmento low-cost, e dalla Text Mining Analysis effettuata sulle due compagnie separatamente è emerso che per entrambe le compagnie due tra gli argomenti più discussi dai clienti sono il topic 18 (Bagaglio) e il topic 22 (Anticipo / ritardo volo), i punti fondamentali su cui vertono la maggior parte delle recensioni dei clienti di queste due compagnie sono la possibilità di imbarcare un bagaglio e il costo che ne consegue, le dimensioni del bagaglio che è possibile portare a bordo dell'aereo, il peso del bagaglio a mano, le eventuali aggiunte di prezzo per un peso del bagaglio superiore e la puntualità dei voli.
L'argomento che invece non viene trattato dai clienti che volano con compagnie low cost è il topic (23) l'intrattenimento a bordo. I clienti che volano con queste compagnie aeree le scelgono per il rapporto qualità / prezzo che viene mediamente discusso nelle recensioni e la possibilità di poter arrivare negli aeroporti principali e quindi non sono interessati all'intrattenimento a bordo (film, riviste, musica, ecc) o extra (snack, bevande, pasti, ecc.) che sono a pagamento e non rappresentano un punto di forza della strategia delle compagnie low cost mentre al contrario sono presenti e tra gli argomenti più discussi e alla base della strategia di altre compagnie aeree come i "super connectors" (Etihad Airways, Qatar Airways e Emirates).

3.3. CORRELAZIONI TRA I TOPIC.

Terminata la fase di analisi e presentazione dei risultati è stato possibile osservare se vi è correlazione tra i topic o meno.
E' stata utilizzata la funzione topicCorr del pacchetto (libreria) STM in R ed è stato possibile determinare le correlazioni tra i vari topic se esse sono presenti e se vi è la correlazione tra due topic quanto essi sono correlati tra loro.
La figura 29 mostra tutte le correlazioni che vi sono tra i vari topic. Utilizzando la funzione "topicCorr" in R è possibile sapere ciascun topic con quanti topic è correlato ed è rappresentato dal numero di linee uscenti da ciascun nodo (ogni nodo rappresenta un topic).

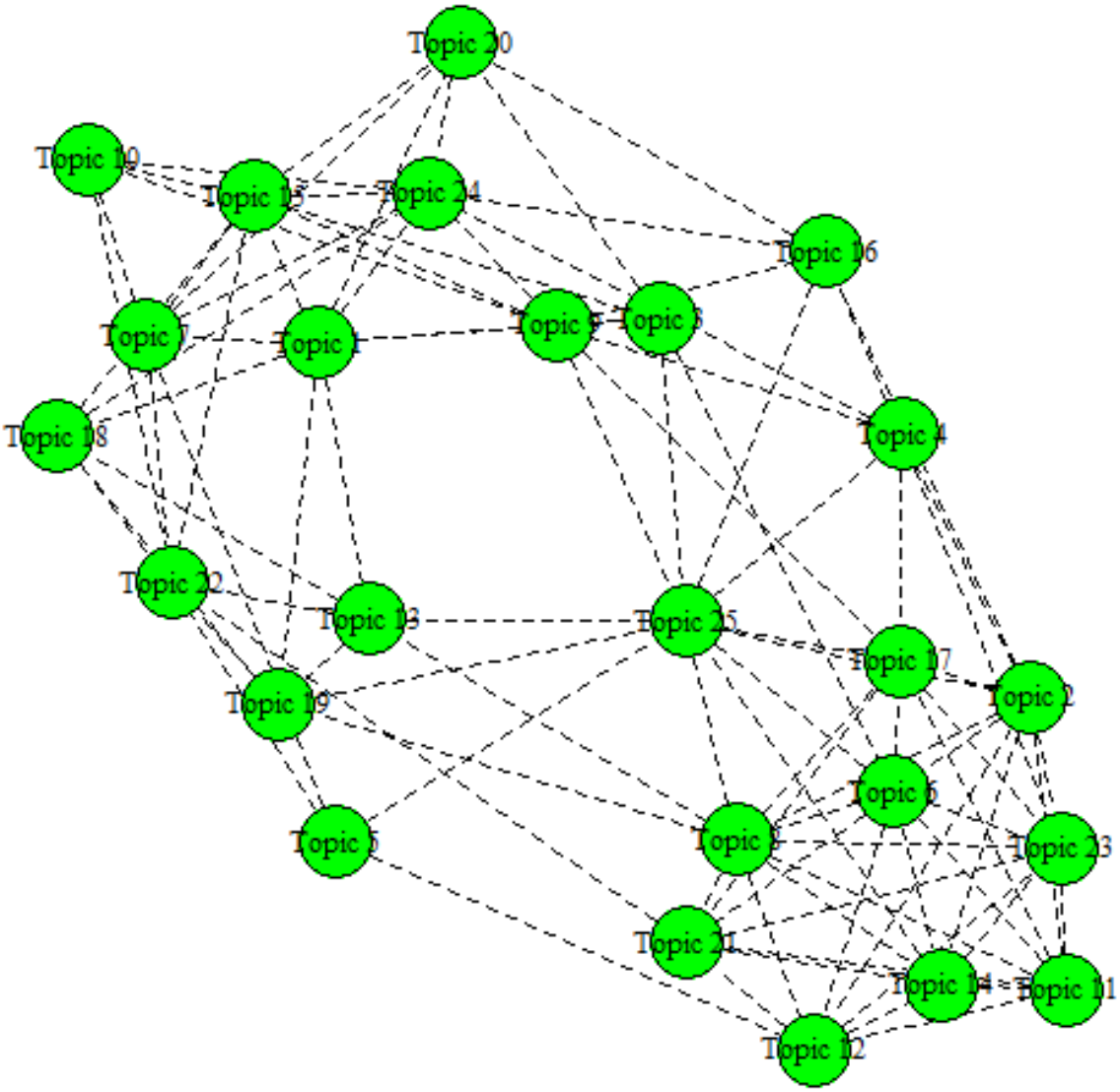


Figura 29. *Grafico correlazioni tra i topic.*

La funzione topicCorr permette di ottenere una matrice simmetrica 25 x 25 (tabella 17) che permette di sapere se un topic è correlato oppure no con un altro topic.
All'interno di ogni casella vi è 1 o 0. Vi è un 1 se vi è correlazione tra i due topic altrimenti vi è uno 0.
Sulla diagonale maggiore vi sono tutti 1 poiché ogni casella della diagonale maggiore indica la correlazione del topic i-esimo con se stesso.
Il topic che ha il maggior numero di correlazioni con gli altri topic è il topic 25 (Problemi con la compagnia aerea) che è correlato con altri 12 topic. Il topic meno correlato con gli altri è il topic 5 (Coincidenza volo) che è correlato solo con altri 4 topic.

Tabella 17. *Matrice simmetrica 25 x 25 di correlazione tra i topic.*

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**
1	1	0	1	0	0	0	1	0	1	0	0	0	1	0	1	0	0	1	1	1	0	0	0	1	0
2	0	1	0	1	0	1	0	1	0	0	1	1	0	1	0	1	1	0	0	0	0	0	1	0	1
3	1	0	1	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1
4	0	1	0	1	0	0	0	0	1	0	0	0	0	0	0	1	1	0	0	0	0	0	1	1	1
5	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1
6	0	1	1	0	0	1	0	1	0	0	1	1	0	1	0	0	1	0	0	0	1	0	1	0	1
7	1	0	0	0	1	0	1	0	0	1	0	0	0	0	1	0	0	0	0	1	0	1	0	1	0
8	0	1	0	0	0	1	0	1	0	0	1	1	1	1	0	0	1	0	1	0	1	0	1	0	1
9	1	0	1	1	0	0	0	0	1	1	0	0	0	0	1	1	1	0	0	0	0	0	0	1	1
10	0	0	0	0	0	0	1	0	1	1	0	0	0	0	1	0	0	0	0	0	0	1	0	1	0
11	0	1	0	0	0	1	0	1	0	0	1	1	0	1	0	0	1	0	0	0	1	0	1	0	0
12	0	1	0	0	1	1	0	1	0	0	1	1	0	1	0	0	0	0	0	0	1	0	1	0	0
13	1	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	1	0	0	1	0	0	1
14	0	1	0	0	0	1	0	1	0	0	1	1	0	1	0	0	0	0	0	0	1	0	1	0	1
15	1	0	1	0	0	0	1	0	1	1	0	0	0	0	1	0	0	1	0	1	0	1	0	1	0
16	0	1	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	1
17	0	1	0	1	0	1	0	1	1	0	1	0	0	0	0	0	1	0	0	0	1	0	1	0	1
18	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	1	0	0	1	0	1	0
19	1	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	1	0	0	1	0	0	1
20	1	0	1	0	0	0	1	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	1	0
21	0	0	0	0	0	1	0	1	0	0	1	1	0	1	0	0	1	0	0	0	1	1	1	0	0
22	0	0	0	0	1	0	1	0	0	1	0	0	1	0	1	0	0	1	1	0	1	1	0	0	0
23	0	1	0	1	0	1	0	1	0	0	1	1	0	1	0	0	1	0	0	0	1	0	1	0	0
24	1	0	0	1	0	0	1	0	1	1	0	0	0	0	1	1	0	1	0	1	0	0	0	1	0
25	0	1	1	1	1	1	0	1	1	0	0	0	1	1	0	1	1	0	1	0	0	0	0	0	1

La funzione topicCorr inoltre fornisce un'ulteriore matrice simmetrica 25 x 25 (tabella 18) in cui all'interno di ciascuna casella della matrice vi è l'indice di correlazione positivo tra il topic i-esimo (con i= 1,…..,25) e il topic j-esimo (con j= 1,…..,25).
Sulla diagonale maggiore vi sono tutti 1 poiché ogni casella della diagonale maggiore indica la correlazione del topic i-esimo con se stesso.

Tabella 18. *Matrice simmetrica 25 x 25 di correlazione positiva tra i topic.*

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1	1,000	0,000	0,029	0,000	0,000	0,000	0,045	0,000	0,070	0,000	0,000	0,000	0,046	0,000	0,102	0,000	0,000	0,160	0,130	0,152	0,000	0,000	0,000	0,119	0,000
2	0,000	1,000	0,000	0,084	0,000	0,069	0,000	0,105	0,000	0,000	0,087	0,131	0,000	0,089	0,000	0,026	0,063	0,000	0,000	0,000	0,000	0,000	0,081	0,000	0,081
3	0,029	0,000	1,000	0,000	0,000	0,013	0,000	0,000	0,029	0,000	0,000	0,000	0,000	0,000	0,023	0,000	0,000	0,000	0,000	0,012	0,000	0,000	0,000	0,000	0,031
4	0,000	0,084	0,000	1,000	0,000	0,000	0,000	0,000	0,109	0,000	0,000	0,000	0,000	0,000	0,000	0,101	0,147	0,000	0,000	0,000	0,000	0,000	0,076	0,033	0,079
5	0,000	0,000	0,000	0,000	1,000	0,000	0,106	0,000	0,000	0,000	0,000	0,028	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,011	0,000	0,000	0,061
6	0,000	0,069	0,013	0,000	0,000	1,000	0,000	0,075	0,000	0,000	0,062	0,156	0,000	0,110	0,000	0,000	0,023	0,000	0,000	0,000	0,156	0,000	0,118	0,000	0,090
7	0,045	0,000	0,000	0,000	0,106	0,000	1,000	0,000	0,000	0,032	0,000	0,000	0,000	0,000	0,192	0,000	0,000	0,000	0,000	0,180	0,000	0,120	0,000	0,114	0,000
8	0,000	0,105	0,000	0,000	0,000	0,075	0,000	1,000	0,000	0,000	0,121	0,161	0,051	0,131	0,000	0,000	0,048	0,000	0,020	0,000	0,095	0,000	0,100	0,000	0,073
9	0,070	0,000	0,029	0,109	0,000	0,000	0,000	0,000	1,000	0,024	0,000	0,000	0,000	0,000	0,061	0,045	0,039	0,000	0,000	0,000	0,000	0,000	0,000	0,063	0,083
10	0,000	0,000	0,000	0,000	0,000	0,000	0,032	0,000	0,024	1,000	0,000	0,000	0,000	0,000	0,059	0,000	0,000	0,000	0,000	0,000	0,000	0,227	0,000	0,026	0,000
11	0,000	0,087	0,000	0,000	0,000	0,062	0,000	0,121	0,000	0,000	1,000	0,176	0,000	0,050	0,000	0,000	0,047	0,000	0,000	0,000	0,024	0,000	0,183	0,000	0,000
12	0,000	0,131	0,000	0,000	0,028	0,156	0,000	0,161	0,000	0,000	0,176	1,000	0,000	0,174	0,000	0,000	0,000	0,000	0,000	0,000	0,074	0,000	0,294	0,000	0,000
13	0,046	0,000	0,000	0,000	0,000	0,000	0,000	0,051	0,000	0,000	0,000	0,000	1,000	0,000	0,000	0,000	0,000	0,050	0,226	0,000	0,000	0,033	0,000	0,000	0,074
14	0,000	0,089	0,000	0,000	0,000	0,110	0,000	0,131	0,000	0,000	0,050	0,174	0,000	1,000	0,000	0,000	0,000	0,000	0,000	0,000	0,039	0,000	0,013	0,000	0,049
15	0,102	0,000	0,023	0,000	0,000	0,000	0,192	0,000	0,061	0,059	0,000	0,000	0,000	0,000	1,000	0,000	0,000	0,147	0,000	0,154	0,000	0,065	0,000	0,156	0,000
16	0,000	0,026	0,000	0,101	0,000	0,000	0,000	0,000	0,045	0,000	0,000	0,000	0,000	0,000	0,000	1,000	0,000	0,000	0,000	0,011	0,000	0,000	0,000	0,016	0,073
17	0,000	0,063	0,000	0,147	0,000	0,023	0,000	0,048	0,039	0,000	0,047	0,000	0,000	0,000	0,000	0,000	1,000	0,000	0,000	0,000	0,049	0,000	0,124	0,000	0,058
18	0,160	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,050	0,000	0,147	0,000	0,000	1,000	0,097	0,000	0,000	0,025	0,000	0,106	0,000
19	0,130	0,000	0,000	0,000	0,000	0,000	0,000	0,020	0,000	0,000	0,000	0,000	0,226	0,000	0,000	0,000	0,000	0,097	1,000	0,000	0,000	0,014	0,000	0,000	0,052
20	0,152	0,000	0,012	0,000	0,000	0,000	0,180	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,154	0,011	0,000	0,000	0,000	1,000	0,000	0,000	0,000	0,072	0,000
21	0,000	0,000	0,000	0,000	0,000	0,156	0,000	0,095	0,000	0,000	0,024	0,074	0,000	0,039	0,000	0,000	0,049	0,000	0,000	0,000	1,000	0,031	0,044	0,000	0,000
22	0,000	0,000	0,000	0,000	0,011	0,000	0,120	0,000	0,000	0,227	0,000	0,000	0,033	0,000	0,065	0,000	0,000	0,025	0,014	0,000	0,031	1,000	0,000	0,000	0,000
23	0,000	0,081	0,000	0,076	0,000	0,118	0,000	0,100	0,000	0,000	0,183	0,294	0,000	0,013	0,000	0,000	0,124	0,000	0,000	0,000	0,044	0,000	1,000	0,000	0,000
24	0,119	0,000	0,000	0,033	0,000	0,000	0,114	0,000	0,063	0,026	0,000	0,000	0,000	0,000	0,156	0,016	0,000	0,106	0,000	0,072	0,000	0,000	0,000	1,000	0,000
25	0,000	0,081	0,031	0,079	0,061	0,090	0,000	0,073	0,083	0,000	0,000	0,000	0,074	0,049	0,000	0,073	0,058	0,000	0,052	0,000	0,000	0,000	0,000	0,000	1,000

3.3.1. CLUSTERING.

La funzione "topicCorr" permette di stabilire delle famiglie (cluster) di topic. Modificando all'interno della funzione "topicCorr" il method e ponendo method = "simple" e modificando il cutoff (soglia al di sotto della quale le correlazioni sono troncate a zero).
Il cutoff viene impostato dall'analista ed è una soglia che tronca a zero tutti gli indici di correlazione positiva inferiore alla soglia impostata.
Il cutoff impostato nell'analisi per ottenere delle famiglie di topic è cutoff = 0,15 e quindi la funzione topicCorr tronca a zero tutti i valori di correlazione positiva inferiori a 0,15 creando un grafico (figura 30) in cui sono presenti solo le correlazioni tra i topic che hanno un indice di correlazione positiva superiore a 0,15 e in questo modo si creano delle famiglie di topic a cui è possibile assegnare un nome in base agli argomenti trattati dai topic che fanno parte di ciascuna famiglia.
La figura 30 mostra che, impostando il cutoff = 0,15 vi sono 4 famiglie di topic e 8 topic singoli.
La famiglia "Prenotazione / cancellazione volo" è formata dai 6 topic 1, 7, 15, 18, 20, 24 che trattano tutti argomenti riguardanti la fase di prenotazione o cancellazione di un volo e per tale motivo sono correlati tra loro.
La seconda famiglia di topic "Interazione con il cliente" è formata da 7 topic: 6, 8, 11, 12, 14, 21,23. Questi topic sono accomunati tra loro dal fatto che tutti trattano la qualità dei servizi forniti a bordo dal personale ai clienti, il rapporto di empatia tra il cliente e il personale,

l'efficienza del personale nell'adempiere alle richieste dei clienti e la comodità del posto durante il viaggio.
La terza famiglia "Rapporto qualità / prezzo" è formata dal topic 13 e dal topic 19, i quali trattano entrambi il rapporto qualità / prezzo e quindi sono fortemente correlati.
L'ultima famiglia "Orario volo" ingloba i topic 10 e 22 che trattano entrambi l'orario dei voli, la puntualità di arrivi e partenze degli aerei, la fase di decollo, la fase di atterraggio, le comunicazioni fornite dal pilota dell'aereo riguardo all'ora di arrivo nell'aeroporto di destinazione, le condizioni della pista, ecc.
Rimangono 8 topic spaiati che non sono collegati tra loro e non appartengono a nessuna famiglia che sono i topic 2, 3, 4, 5, 9, 16, 17 e 25. Questi topic trattano degli argomenti che con cutoff = 0,15 non sono collegati alle altre famiglie di topic e nemmeno tra loro.

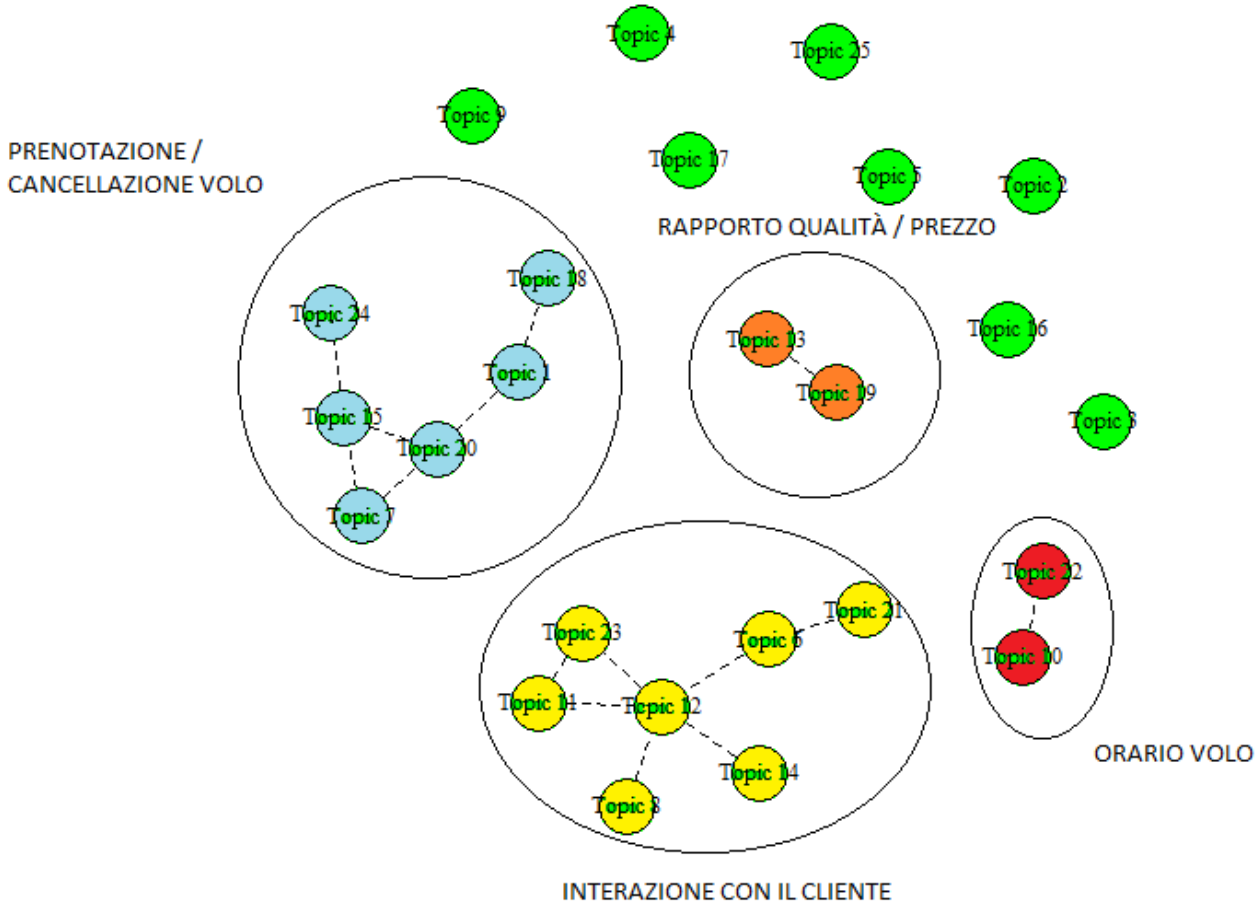


Figura 30. *Clustering.*

Applicando la funzione "topicCorr", oltre alla figura che mostra i vari cluster di topic, è possibile osservare due matrici simmetriche.
La prima matrice (tabella 19) indica se vi è correlazione oppure no tra un topic e un altro secondo le specifiche method= "simple" e cutoff = 0,15 impostate nella funzione "topicCorr" e ciò è già osservabile nella figura precedente (figura 30), la prima matrice è una rappresentazione della figura 30 e all'interno di ciascuna casella della matrice vi è 1 se tra il topic i-esimo e il topic j-esimo vi è correlazione, 0 altrimenti.
Sulla diagonale maggiore vi sono tutti 1 poiché ogni casella della diagonale maggiore indica la correlazione del topic i-esimo con se stesso.

Il topic con il maggior numero di correlazioni è il topic 12 appartenente al cluster “Interazione con il cliente” che è correlato con altri 5 topic all’interno del cluster.
I topic meno correlati sono i topic spaiati, che non appartengono a nessuna famiglia e non hanno alcuna correlazione tra di loro o con i topic appartenenti ad una famiglia e l’unica correlazione che possiedono è quella con se stessi presente sulla diagonale maggiore.

Tabella 19. *Matrice simmetrica 25 x 25 di correlazione tra i topic che appartengono alla stessa famiglia.*

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0
2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0
7	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0
8	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
11	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0
12	0	0	0	0	0	1	0	1	0	0	1	1	0	1	0	0	0	0	0	0	0	0	1	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
18	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0
20	1	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0
21	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
22	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
23	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

La funzione “topicCorr” permette di ottenere una seconda matrice simmetrica 25 x 25 (tabella 20) in cui all’interno di ciascuna casella della matrice vi è l’indice di correlazione positivo tra il topic i-esimo (con i=1,…..,25) e il topic j-esimo (con j=1,…..,25).
Sulla diagonale maggiore vi sono tutti 1 poiché ogni casella della diagonale maggiore indica la correlazione del topic i-esimo con se stesso.

Tabella 20. *Matrice simmetrica 25 x 25 di correlazione positiva tra i topic che appartengono alla stessa famiglia.*

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1	1,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,160	0,000	0,152	0,000	0,000	0,000	0,000	0,000
2	0,000	1,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
3	0,000	0,000	1,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
4	0,000	0,000	0,000	1,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
5	0,000	0,000	0,000	0,000	1,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
6	0,000	0,000	0,000	0,000	0,000	1,000	0,000	0,000	0,000	0,000	0,000	0,156	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,156	0,000	0,000	0,000	0,000
7	0,000	0,000	0,000	0,000	0,000	0,000	1,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,192	0,000	0,000	0,000	0,000	0,180	0,000	0,000	0,000	0,000	0,000
8	0,000	0,000	0,000	0,000	0,000	0,000	0,000	1,000	0,000	0,000	0,000	0,161	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
9	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	1,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
10	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	1,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,227	0,000	0,000	0,000
11	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	1,000	0,176	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,183	0,000	0,000
12	0,000	0,000	0,000	0,000	0,000	0,156	0,000	0,161	0,000	0,000	0,176	1,000	0,000	0,174	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,294	0,000	0,000
13	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	1,000	0,000	0,000	0,000	0,000	0,000	0,226	0,000	0,000	0,000	0,000	0,000	0,000
14	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,174	0,000	1,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
15	0,000	0,000	0,000	0,000	0,000	0,000	0,192	0,000	0,000	0,000	0,000	0,000	0,000	0,000	1,000	0,000	0,000	0,000	0,000	0,154	0,000	0,000	0,000	0,156	0,000
16	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	1,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
17	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	1,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
18	0,160	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	1,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
19	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,226	0,000	0,000	0,000	0,000	0,000	1,000	0,000	0,000	0,000	0,000	0,000	0,000
20	0,152	0,000	0,000	0,000	0,000	0,000	0,180	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,154	0,000	0,000	0,000	0,000	1,000	0,000	0,000	0,000	0,000	0,000
21	0,000	0,000	0,000	0,000	0,000	0,156	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	1,000	0,000	0,000	0,000	0,000
22	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,227	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	1,000	0,000	0,000	0,000
23	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,183	0,294	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	1,000	0,000	0,000
24	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,156	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	1,000	0,000
25	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	1,000

CAPITOLO 4

4.1. ANALISI TEMPORALE DEI TOPIC.

L'analisi temporale di ciascun topic ha la funzione di mostrare all'analista in che periodi il topic è stato maggiormente discusso dagli users e in quali è stato meno discusso e se vi sono dei trend o stagionalità.
Per effettuare quest'analisi per ciascun topic i-esimo sono state considerate tutte le 20446 percentuali di prevalenza θ del topic i-esimo all'interno delle recensioni corrispondenti e la data in cui è stato effettuato il volo dall'utente.
Le date e le relative percentuali di prevalenza sono state raccolte in un file EXCEL e ordinate in base alla data da gennaio 2015 a marzo 2020.
Per ciascun mese di ciascun anno si è fatta la media di tutte le percentuali di prevalenza ottenendo un θ_{medio} relativo al mese corrispondente e in questo modo è stato possibile determinare per ciascun topic un grafico che mostra l'andamento del topic nel tempo da gennaio 2015 a marzo 2020.
Sono state escluse dal campione utilizzato per effettuare l'analisi temporale 2146 recensioni poiché non è stata inserita dall'utente che ha rilasciato la recensioni in rete la data del volo e quindi non potevano essere utilizzate per l'analisi e hanno la scritta "NULL" al posto della data.
All'interno dell'analisi non sono presenti recensioni riferite a febbraio 2015 poiché non sono presenti nel dataset iniziale recensioni di voli compiuti a febbraio 2020.
Inoltre le recensioni riferite a gennaio 2015 sono solamente due e quindi il valore della media può rappresentare un punto di massimo o di minimo perché la media è fatta solo su due valori ma essendo solo due le recensioni riferite a gennaio 2015 il valore della media non è molto rilevante ma è stato riportato comunque nei grafici.
La figura 31 indica che nella prima colonna vi sono le date dei voli ordinate dalla più lontana alla più recente, nella seconda colonna le relative percentuali di prevalenza, nella terza i mesi da gennaio 2015 a marzo 2020 e nella quarta e ultima colonna i valori di θ_{medio} corrispondenti.

gen-15	0,0225825	gen-15	0,020511
gen-15	0,0184395	mar-15	0,0115787
mar-15	0,0082577	apr-15	0,0122428
mar-15	0,0085354	mag-15	0,0124858
mar-15	0,0109118	giu-15	0,0148304
mar-15	0,0195474	lug-15	0,0144807
mar-15	0,011206	ago-15	0,0145254
mar-15	0,0084433	set-15	0,0150617
mar-15	0,014149	ott-15	0,0137808
apr-15	0,008923	nov-15	0,0138264
apr-15	0,0104006	dic-15	0,0143163
apr-15	0,0091847	gen-16	0,0142163
apr-15	0,0100887	feb-16	0,0133923
apr-15	0,0133712	mar-16	0,0133519
apr-15	0,0119512	apr-16	0,0147897
apr-15	0,0148566	mag-16	0,0148824
apr-15	0,0116139	giu-16	0,0150389
apr-15	0,0149336	lug-16	0,0210763
apr-15	0,0121159	ago-16	0,0154631
apr-15	0,0120872	set-16	0,0147635
apr-15	0,0086574	ott-16	0,0149921
apr-15	0,0086715	nov-16	0,0155093
apr-15	0,00981	dic-16	0,0151879
apr-15	0,0195318	gen-17	0,0147732
apr-15	0,0196878	feb-17	0,015308
mag-15	0,0130103	mar-17	0,0144283
mag-15	0,0083305	apr-17	0,0153729
mag-15	0,0082683	mag-17	0,0153353

Figura 31. *Esempio di dati utilizzati per ottenere il grafico dell'andamento temporale.*

Il topic 1 (Prenotazione volo) mostra un trend positivo (linea tratteggiata rossa) come è possibile osservare in figura 32.
Il punto di minimo si ha in maggio 2015 con un θ_{medio}= 0,63% ed il periodo in cui è stato meno discusso mentre i periodi in cui la prenotazione del volo è stata più discussa dai clienti sono stati settembre 2017, dicembre 2017, novembre 2018, settembre 2019, ottobre 2019,novembre 2019 e il punto di massimo in cui il topic 1 è stato maggiormente discusso è stato a marzo 2020 e ciò è dovuto a valori più elevati di θ, che hanno fatto aumentare il valore di θ_{medio} e anche all'emergenza COVID 19 che ha fatto si che i clienti trattassero maggiormente tale argomento nelle recensioni di marzo 2020 e questo fattore è maggiormente evidente nel grafico del topic 20 (Prenotazione volo (2)).

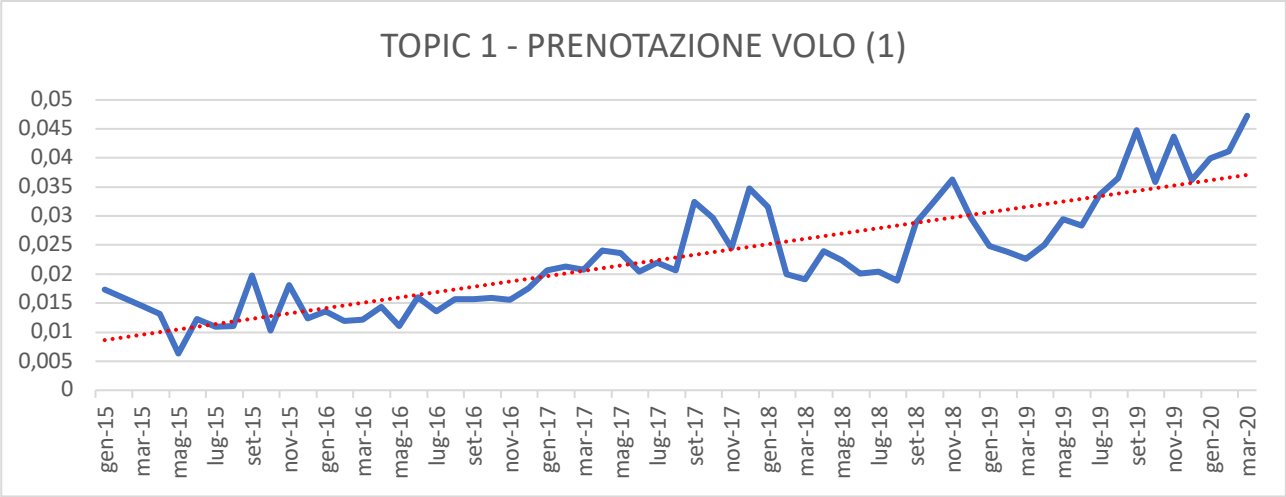


Figura 32. *Andamento temporale topic 1.*

Il topic 2 (Tipo di classe) non mostra un particolare tipo di trend o stagionalità, ha un andamento in cui si alternano valori inferiori e superiori alla media.
Nel 2015, in generale, l'argomento riguardante la classe in cui il cliente ha compiuto il volo era molto discusso e anche nel 2017 il topic 2 è stato molto discusso mentre dalla fine del 2017 a marzo 2020 il tipo di classe è stato sempre meno discusso all'interno delle recensioni delle 10 compagnie aeree del campione fino a culminare nel punto di minimo assoluto di marzo 2020 (Figura 33).

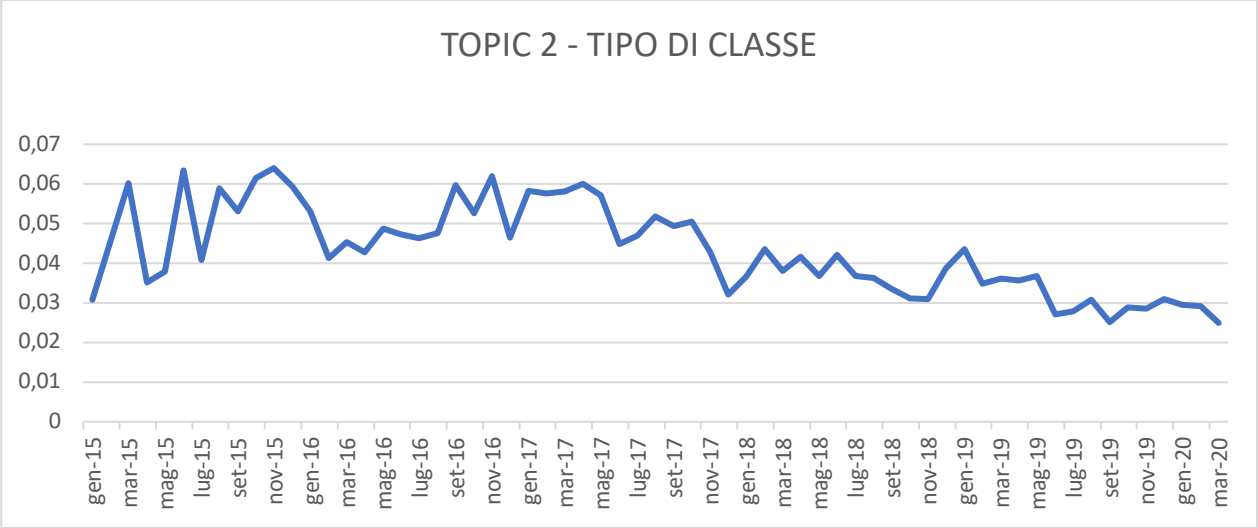


Figura 33. *Andamento temporale topic 2.*

La figura 34 illustra che il topic 3 (Famiglia) ha un andamento pressochè costante da gennaio 2015 a marzo 2020, infatti, tutti i valori di θ_{medio} sono compresi all'interno dell'intervallo [0,01 , 0,03] eccetto i valori di agosto 2015 (θ_{medio}= 3,33%) di poco fuori dall'intervallo e il valore di gennaio 2015 (5,06 %) che rappresenta il punto di massimo, ma ciò è dovuto al fatto che le recensioni di gennaio 2015 sono solo due e hanno valori elevati.

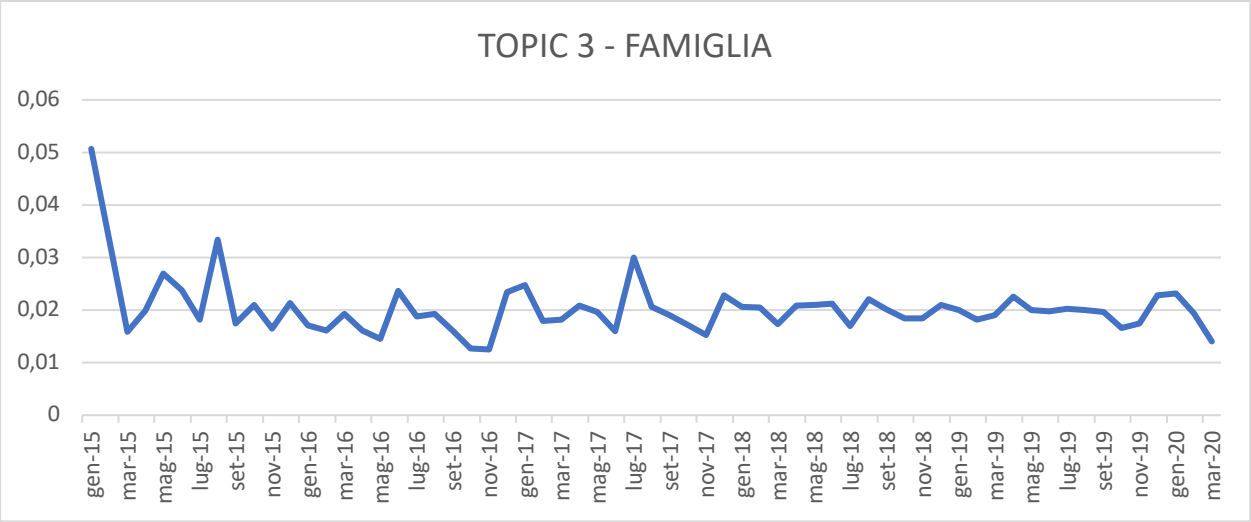


Figura 34. Andamento temporale topic 3.

Il grafico (figura 35) mostra l'andamento del topic 4 (Pasti a bordo). L'argomento è stato poco discusso nel 2015 e si ha il punto di minimo a marzo 2015 (θ_{medio}= 0,93%). Nel 2016 è stato maggiormente discusso fino a raggiungere il punto di massimo in gennaio 2017 e in seguito tutti i valori da aprile 2017 a marzo 2020, eccetto il valore di gennaio 2019 (θ_{medio}= 4,26%), sono compresi nell'intervallo [0,02 , 0,04].

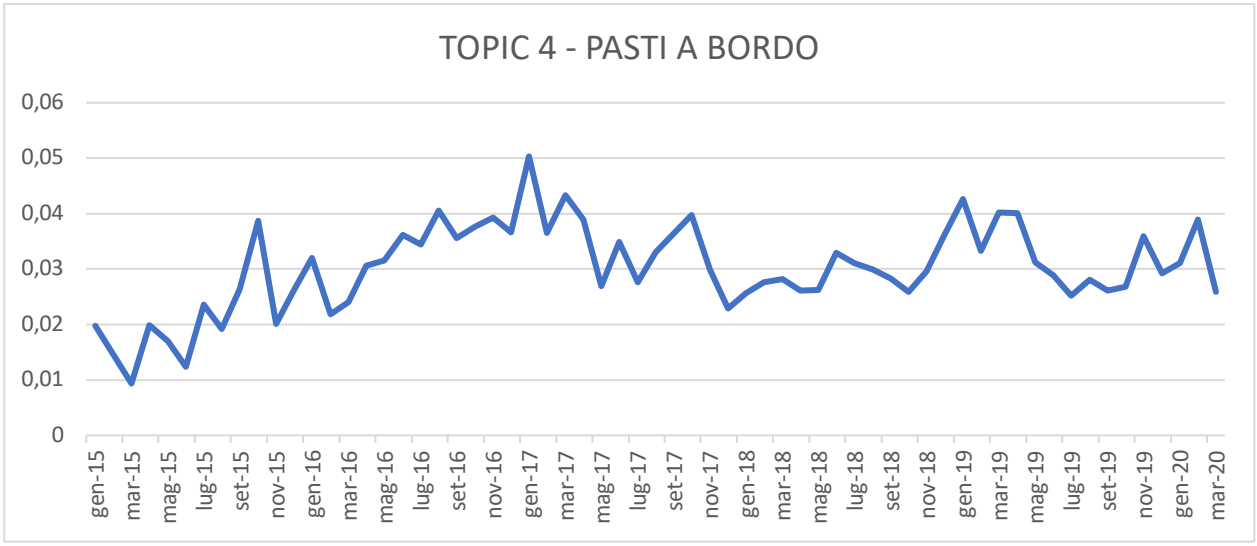


Figura 35. *Andamento temporale topic 4.*

Il topic 5 (Coincidenza volo), escluso il valore iniziale (gennaio 2015) presenta un andamento costante e non vi sono né trend né stagionalità (Figura 36).

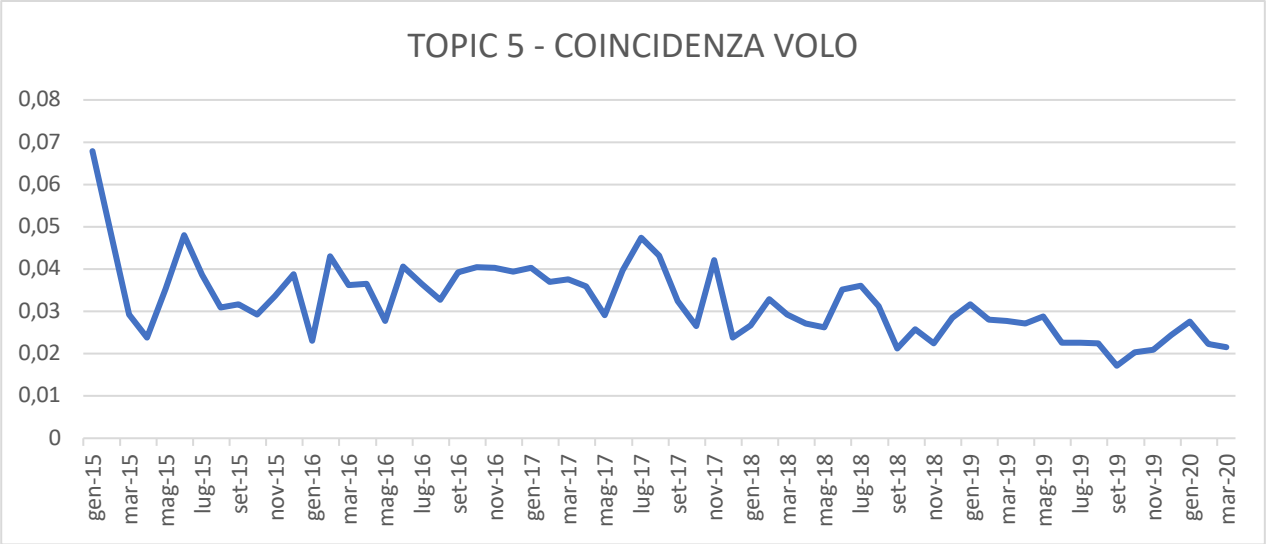


Figura 36. *Andamento temporale topic 5.*

Il topic 6 (Comodità viaggio) presenta un trend decrescente (figura 37). Questo argomento ha un punto di massimo a marzo 2015 con θ_{medio}= 14,05%, infatti la comodità del viaggio è un argomento largamente trattato nelle recensioni nel 2015 mentre dal 2016 diminuiscono mantenendo un andamento costante fino al punto di minimo che si ha a marzo 2020 (θ_{medio}= 3,9%).

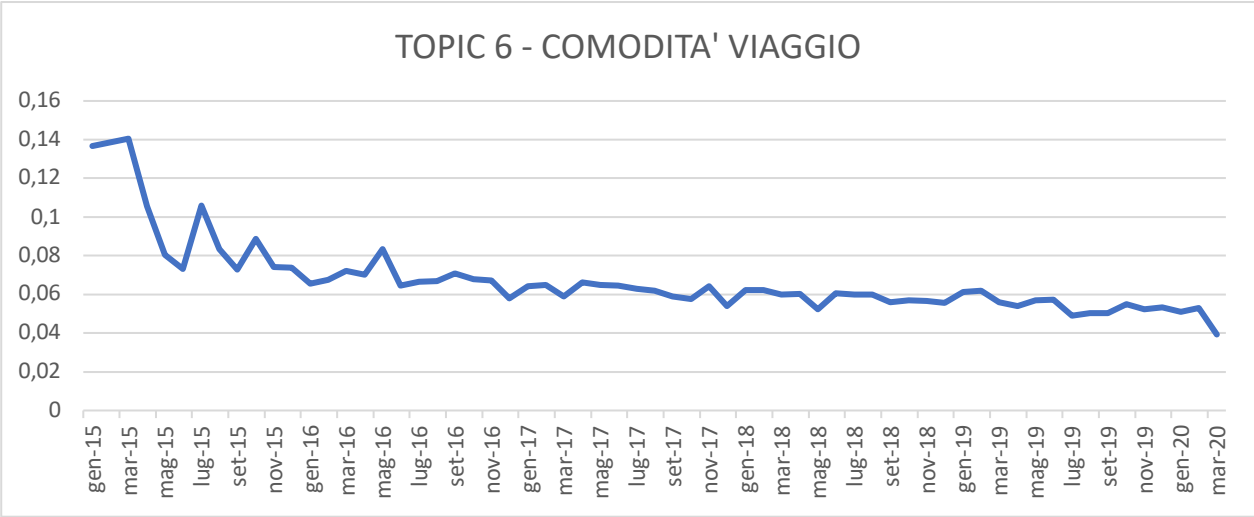


Figura 37. *Andamento temporale topic 6.*

La figura 38 mostra che il topic 7 (Cancellazione volo) ha un trend crescente (linea tratteggiata rossa) da gennaio 2015 a marzo 2020 e stagionalità in maggio, giugno e luglio di ciascun anno.
Il grafico mostra un trend crescente infatti i valori di ciascun anno sono superiori a quelli dell'anno precedente e si può osservare che in ciascun anno nei mesi di maggio, giugno, luglio e dicembre l'argomento cancellazione volo viene notevolmente discusso all'interno delle recensioni.
Il topic mostra dei picchi di massimo relativo rispetto a ciascun anno in giugno 2015 (θ_{medio}= 3,55%), in dicembre 2016 (θ_{medio}= 5,97%), in dicembre 2017 (θ_{medio}= 5,65 %), giugno 2018 (θ_{medio}= 7,99%) e luglio 2019 (θ_{medio}= 6,78%).

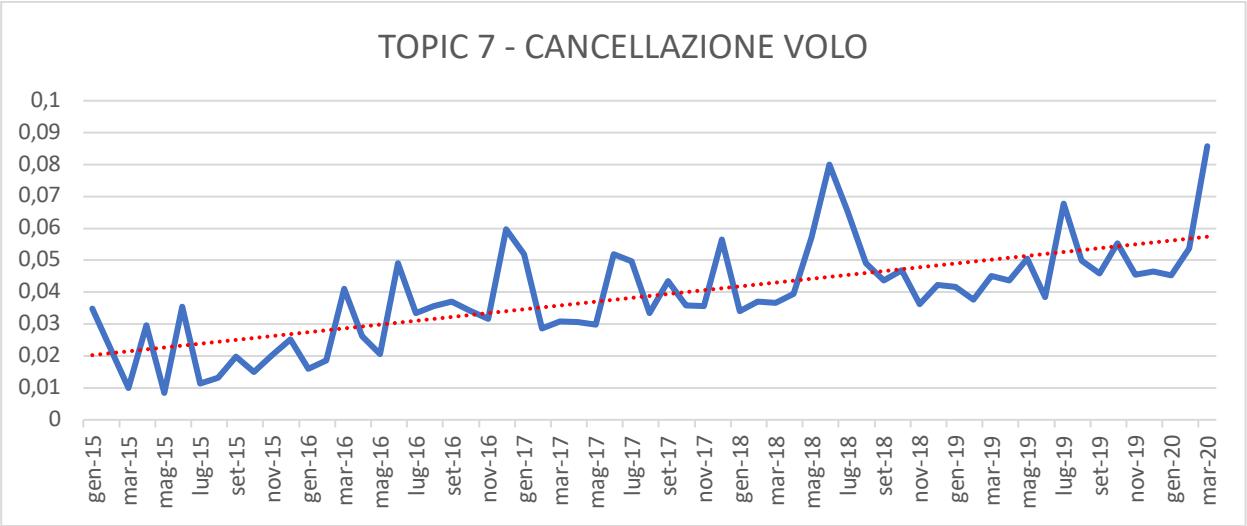


Figura 38. *Andamento temporale topic 7.*

Il grafico (figura 39) espone l'andamento del topic 8 (Qualità servizi (1)) e si può notare che vi è un trend leggermente decrescente da gennaio 2017 a marzo 2020 mentre da gennaio 2015 a dicembre 2016 l'andamento temporale del topic prosegue in modo alternato.

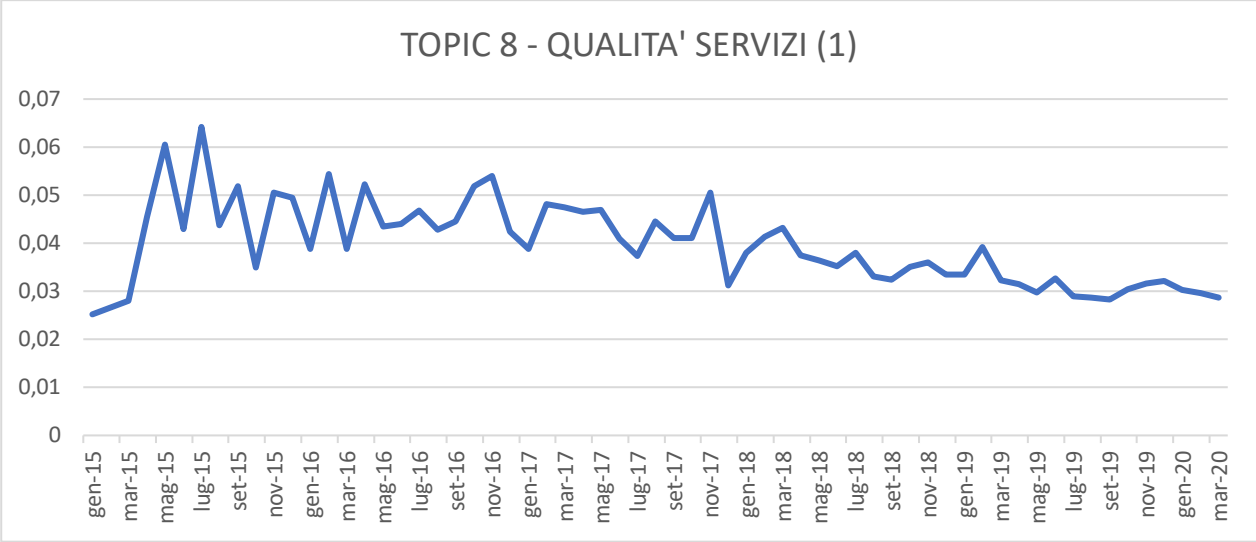


Figura 39. *Andamento temporale topic 8.*

Il topic 9 (Snack & bevande) ha un andamento alternato nel 2015 e nel 2016 con picchi in marzo 2015, giugno 2015 e luglio 2016, mentre presenta un andamento costante leggermente crescente dal 2017 al 2020 (figura 40).

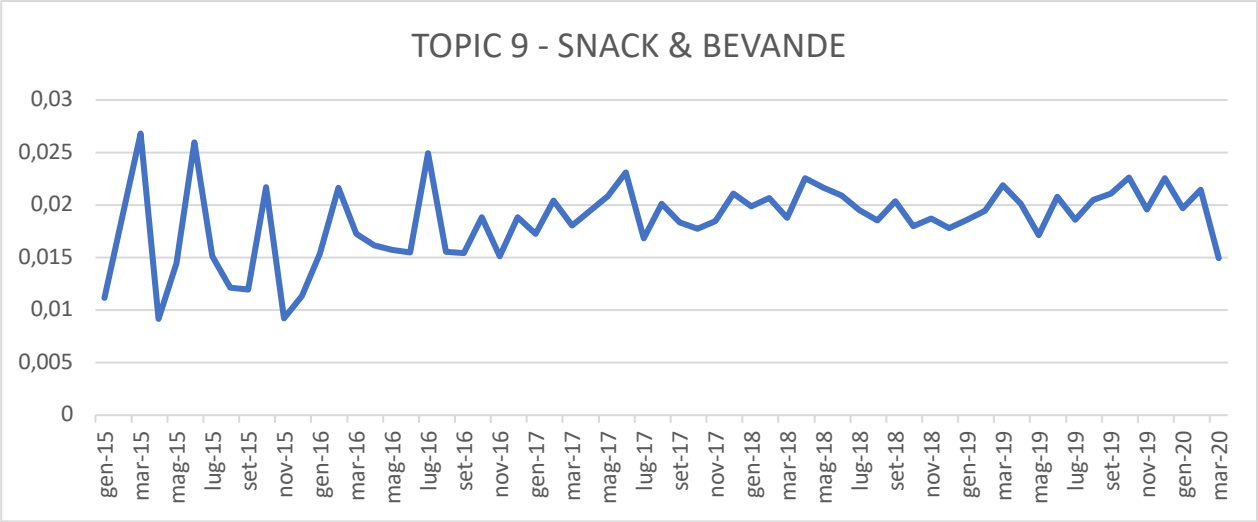


Figura 40. Andamento temporale topic 9.

Il topic 10 (Decollo / atterraggio) ha un trend crescente (figura 41) come mostra la linea di tendenza nel grafico (linea rossa tratteggiata) e una stagionalità nei mesi di novembre, dicembre e gennaio).
Nel 2015 si ha un punto di massimo relativo in marzo 2015 (θ_{medio}= 3,45%). Nel 2016 i valori sono più elevati tra novembre 2016 e gennaio 2017 così come per l'anno 2017 i valori sono più elevati tra novembre 2017 e gennaio 2018.
Nel 2018 i valori crescono maggiormente tra ottobre e dicembre mentre nei tre mesi del 2020 per cui si hanno recensioni i valori sono più elevati a gennaio 2020.
Questo andamento mostra che gli users discutono maggiormente all'interno delle recensioni che rilasciano l'esperienza del decollo e dell'atterraggio dell'aereo soprattutto nei mesi invernali, in cui per molti vi sono le vacanze natalizie e molte persone si spostano in aereo.

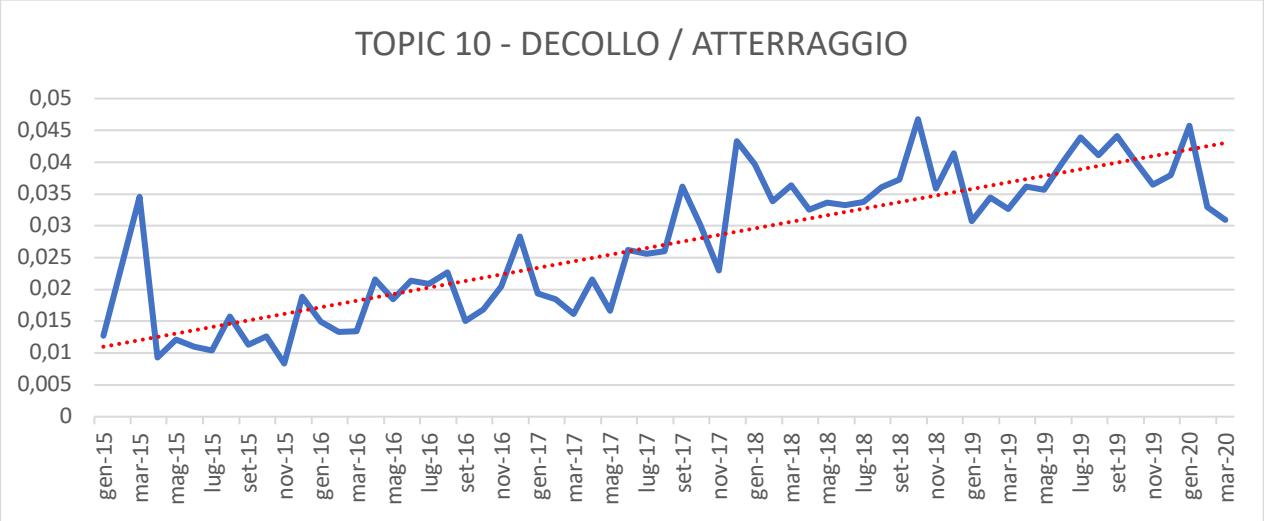


Figura 41. Andamento temporale topic 10.

La figura 42 illustra che il topic 11 (Manutenzione e pulizia) ha un trend fortemente negativo (linea tratteggiata rossa) mentre non vi sono stagionalità.
L'argomento riguardante la manutenzione e la pulizia dei locali dell'aereo (posto a sedere, servizi igienici, corridoio tra i sedili, la moquette per terra, i finestrini, il tavolino, ecc.) è stato

molto discusso nel 2015 e il punto di massimo del grafico si ha in giugno 2015 con un θ_{medio}= 7,90%.
Nel 2016 il topic 11 è stato molto discusso nelle recensioni ma con valori inferiori al 2015 con un punto di massimo relativo in febbraio 2016. Dal 2017 al 2020 i valori sono stati tutti costanti con andamento decrescente senza particolari valori di picco superiore o inferiore.

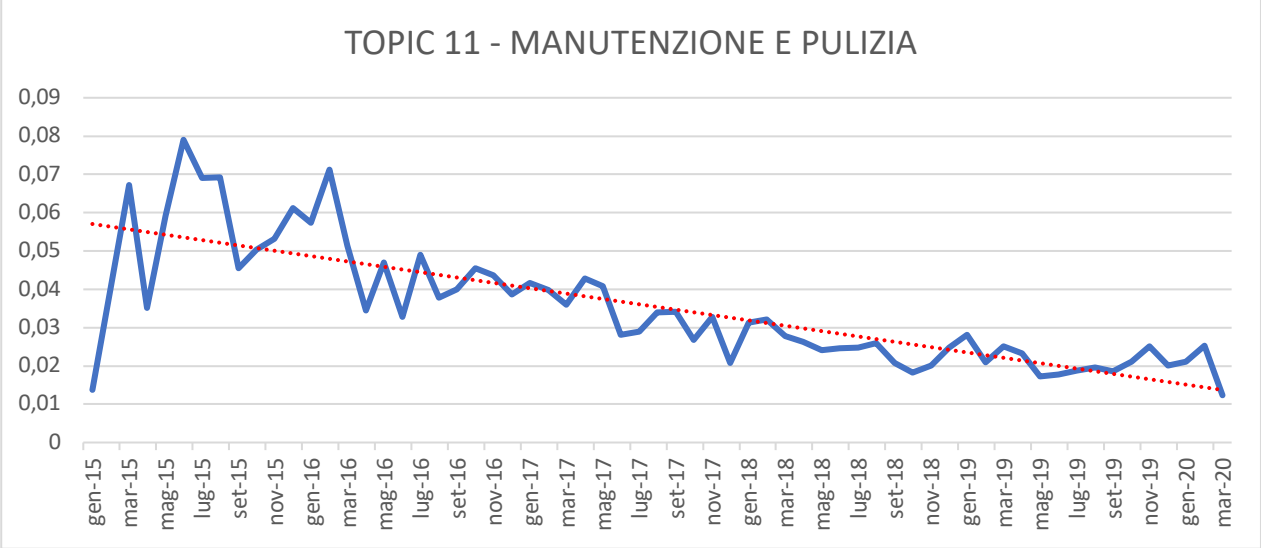


Figura 42. *Andamento temporale topic 11.*

Il topic 12 (Qualità servizi (2)) presenta un trend negativo (linea tratteggiata rossa), infatti i valori nei mesi di ciascun anno sono inferiori ai rispettivi valori dei mesi dell'anno precedente e non vi sono stagionalità.
Se non vi fosse il valore di minimo assoluto di gennaio 2015 il trend sarebbe marcatamente più negativo e ciò sarebbe documentato da una maggiore pendenza della linea di tendenza lineare.

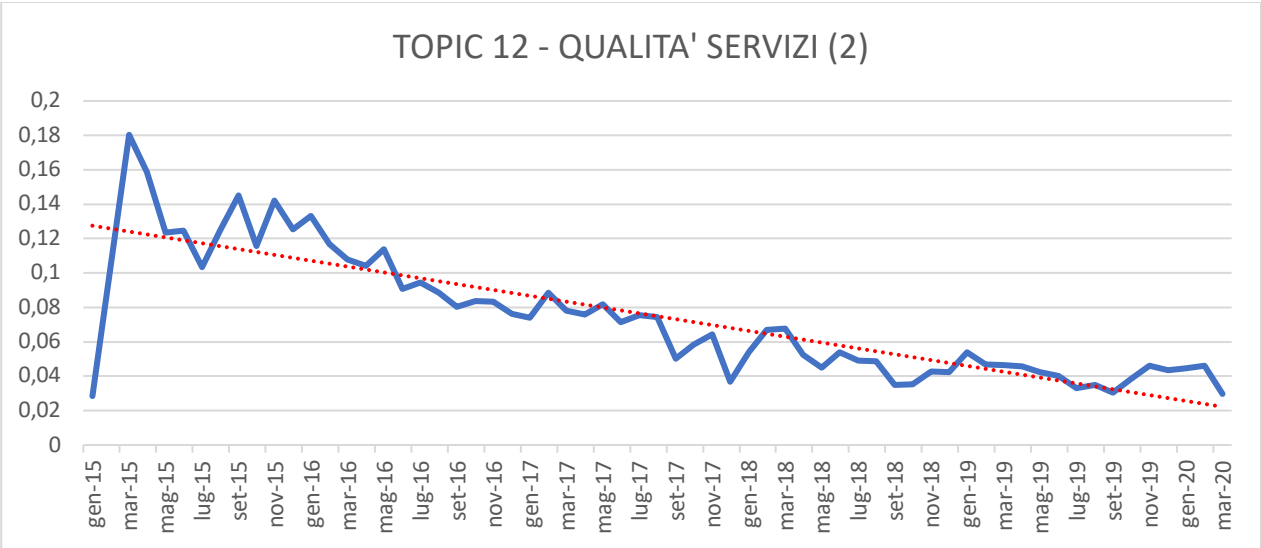


Figura 43. *Andamento temporale topic 12.*

Il topic 13 (Rapporto qualità / prezzo) presenta un trend positivo (linea tratteggiata rossa). L'argomento è stato maggiormente discusso nelle recensioni rilasciate dagli users negli ultimi anni piuttosto che negli anni precedenti (Figura 44).
Il picco di massima discussione del rapporto qualità/prezzo di un volo si è riscontrato in ottobre 2018 mentre il punto di minimo assoluto è rappresentato da marzo 2015.
All'interno del grafico non è stata riscontrata nessuna stagionalità.

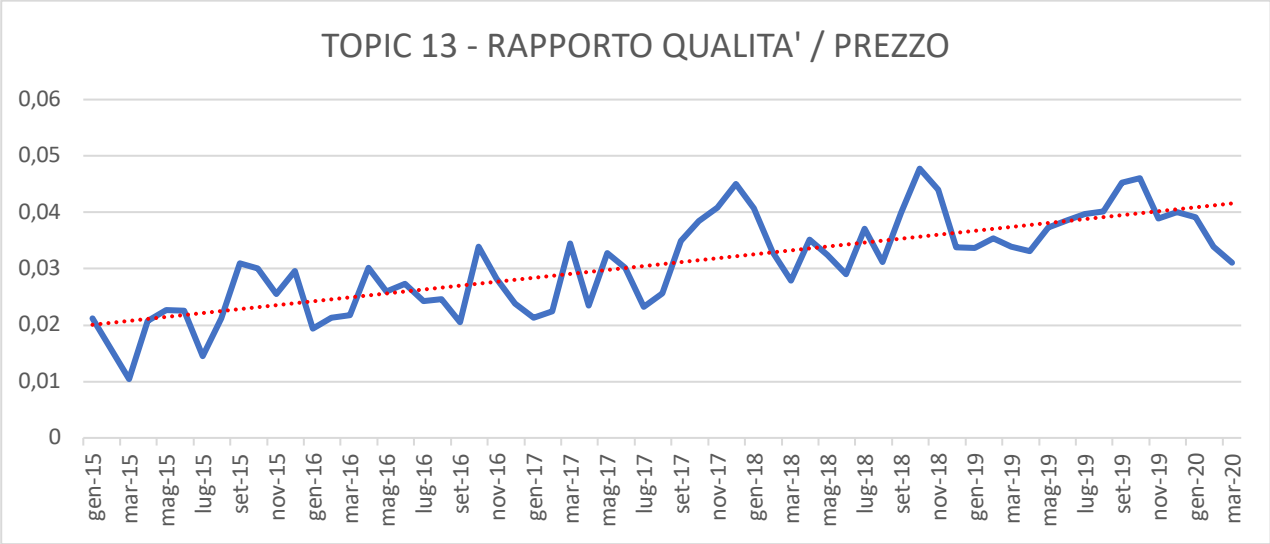


Figura 44. *Andamento temporale topic 13.*

La figura 45 mostra l'andamento del topic 14 (Competenze personale (1)) ed è possibile notare che tale argomento non presenta né trend né stagionalità.
Questo argomento è discusso in maniera costante nel tempo e tutti i valori esclusi alcuni outlier (gennaio 2015, aprile 2015, dicembre 2017, gennaio 2020, marzo 2020) sono compresi nell'intervallo [0,03 ; 0,06].
I clienti delle 10 compagnie aeree prese in considerazione nel campione di riferimento hanno discusso sempre in maniera costante le competenze del personale della compagnia con cui hanno volato all'interno delle recensioni rilasciate on line.

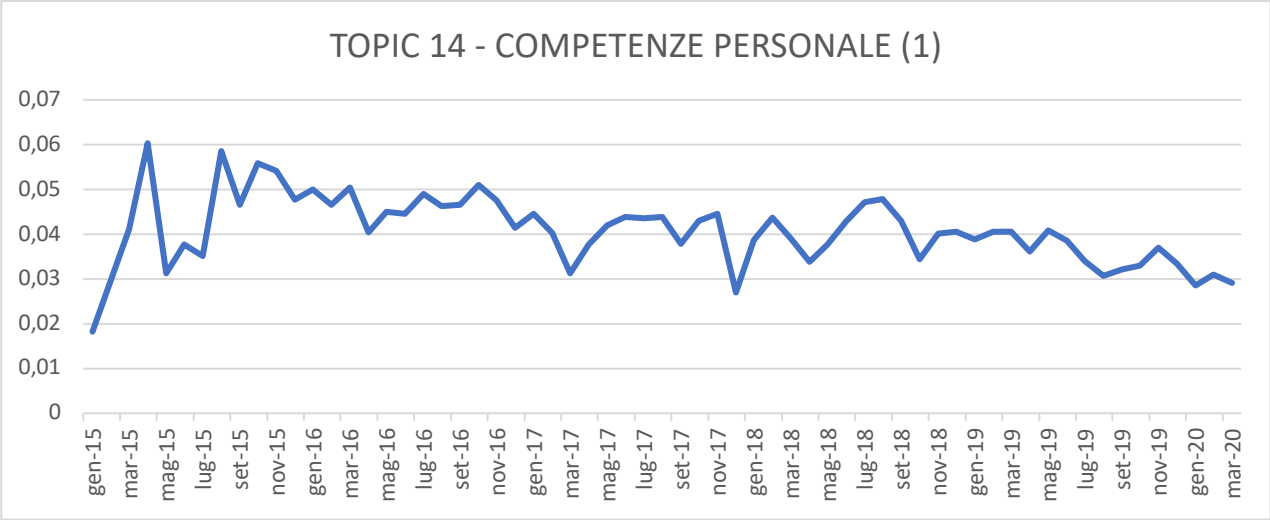


Figura 45. *Andamento temporale topic 14.*

L'andamento del topic 15 (Check-in) rappresentato in figura 46 mostra un trend positivo da gennaio 2015 a marzo 2020 e ciò indica che il check-in è stato un argomento discusso maggiormente con il passare degli anni.
Il periodo in cui è stato meno discusso è stato maggio 2015 (θ_{medio}= 0,56%) in cui si ha il punto di minimo assoluto del grafico, mentre il punto di massimo si ha in marzo 2020 (θ_{medio}= 6,45%). Il topic 15 non presenta stagionalità.

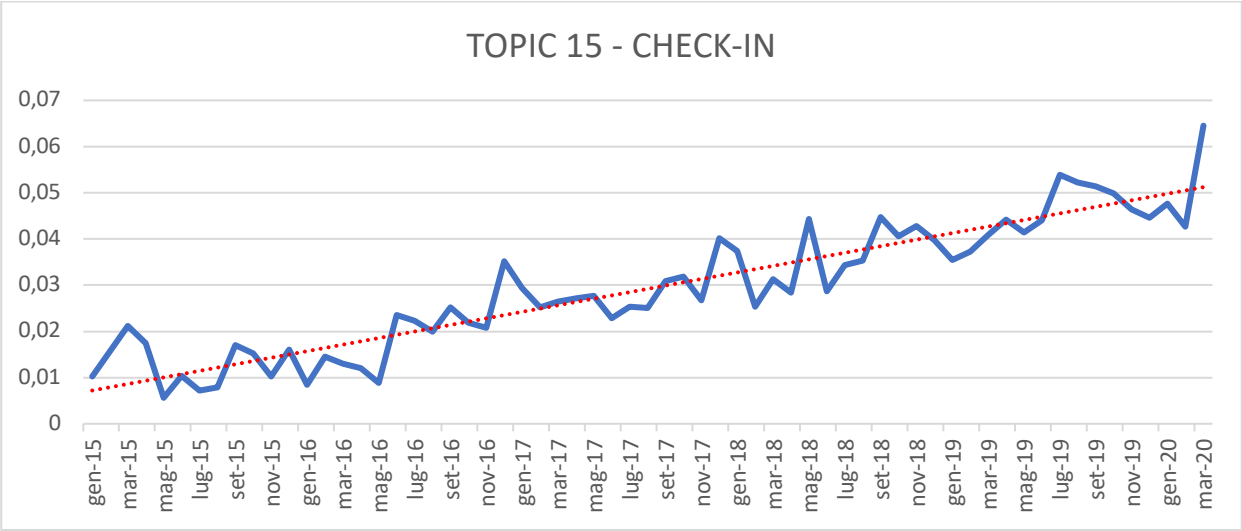


Figura 46. *Andamento temporale topic 15.*

Il topic 16 (Lingua), ossia le lingue parlate a bordo dal comandante e dal personale durante il volo, è un argomento discusso negli UGC in maniera molto incostante e alternata dai clienti delle compagnie aeree (Figura 47).
Il topic 16 non presenta né trend né stagionalità, ma solo un andamento incostante e questo indica che l'argomento è trattato in maniera sporadica nelle recensioni e ciò è giustificato anche dal range molto basso di θ_{medio} nel grafico, infatti tutti i valori sono compresi nell'intervallo [0,01 , 0,035].
L'argomento "lingua" è stato maggiormente discusso nel 2016 e 2017 raggiungendo il picco nelle recensioni di maggio 2017 (θ_{medio}= 3,49%), mentre il punto di minimo del grafico è rappresentato da marzo 2015 (θ_{medio}=1,05%).

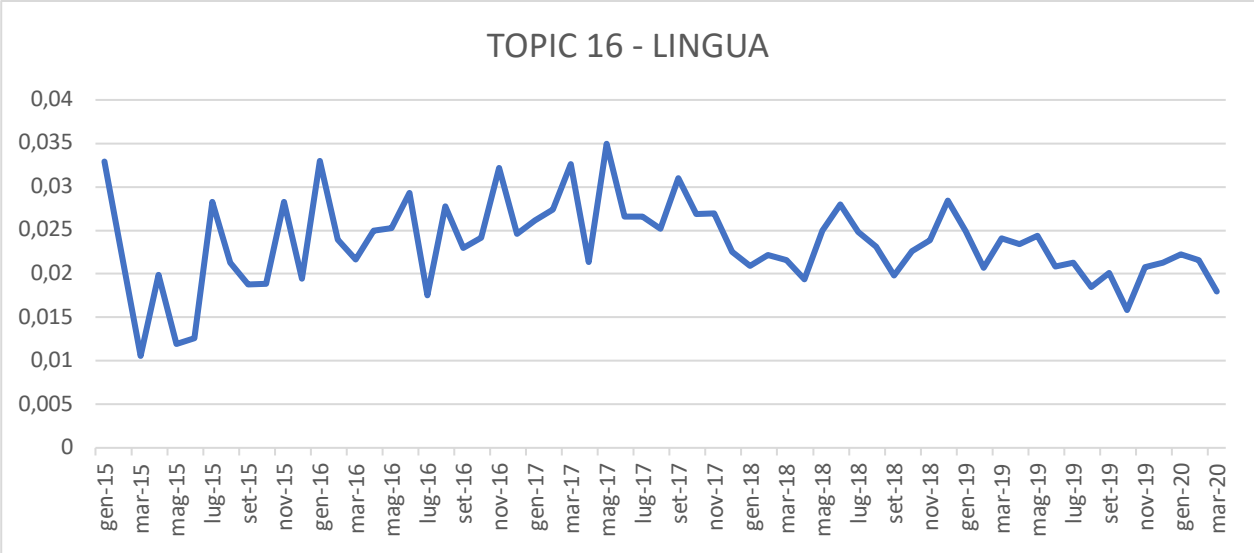


Figura 47. *Andamento temporale topic 16.*

La figura 48 mostra l'andamento del topic 17 (Posto a bordo). Si riferiscono a questo topic le recensioni in cui il cliente spiega la comodità del posto in cui era seduto durante il volo (spazio per le gambe, posto spazioso, sedile pieghevole, tavolino, possibilità di allungarsi, ecc.).
L'argomento "Posto a bordo" ha un andamento dei valori dal 2015 al 2020 costante e ciò è rappresentato dalla linea di tendenza lineare (linea tratteggiata rossa) eccetto che nel 2015 in cui è presente un andamento alternato dei valori ma ciò è dovuto anche al minor numero di recensioni per l'anno 2015 (2 recensioni per gennaio 2015, 0 per febbraio 2015, 7 per marzo 2015, 16 per aprile 2015, 19 per maggio 2015, ecc.), che in totale sono 344 (circa l'1,8% delle recensioni escludendo le recensioni che non hanno data ma il valore NULL al posto della data).
Quindi l'andamento del topic 17 nel tempo non presenta né trend né stagionalità, ma è un argomento che viene trattato abbastanza nelle recensioni e ha valori di prevalenza media elevata che superano anche il 7% in luglio 2015 (θ_{medio}= 7,29%), mentre i valori di θ_{medio} da gennaio 2016 a marzo 2020 sono compresi nell'intervallo [0,02 , 0,06].

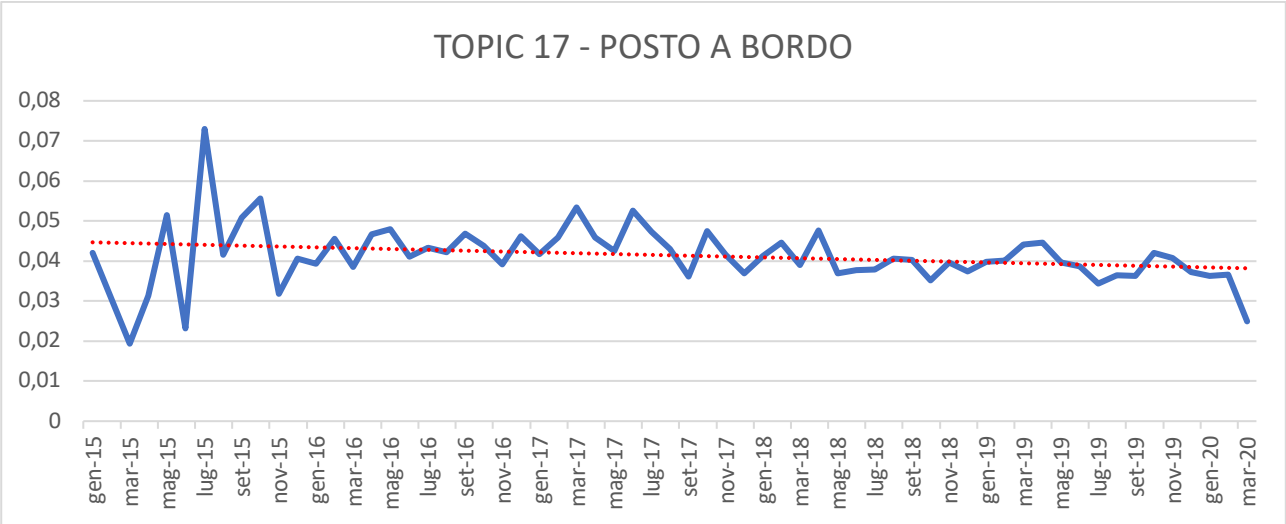


Figura 48. *Andamento temporale topic 17.*

La figura 49 illustra l'andamento del topic 18 (Bagaglio).
È possibile suddividere il grafico raffigurante l'andamento del topic 18 in due parti attraverso una linea rossa verticale: la prima parte va da gennaio 2015 a giugno 2017 e la seconda parte che va da luglio 2017 a marzo 2020.
Nella prima parte vi è un andamento dei dati costante e tutti i valori di θ_{medio} sono racchiusi nell'intervallo [0 , 0,02] e sono valori molto bassi e ciò sta a indicare da gennaio 2015 a giugno 2017 l'argomento riguardante la possibilità di mettere bagagli nella stiva e il costo per tale servizio, il bagaglio a mano, le dimensioni dei bagagli, il peso, ecc. sono stati discussi in maniera minima.
Nella seconda parte del grafico i valori sono nettamente più elevati e compresi nell'intervallo [0,02 , 0,08] e non vi è nessun trend o stagionalità.
Nella seconda parte i dati presentano un andamento alternato e si susseguono punti di massimo e minimo relativo.

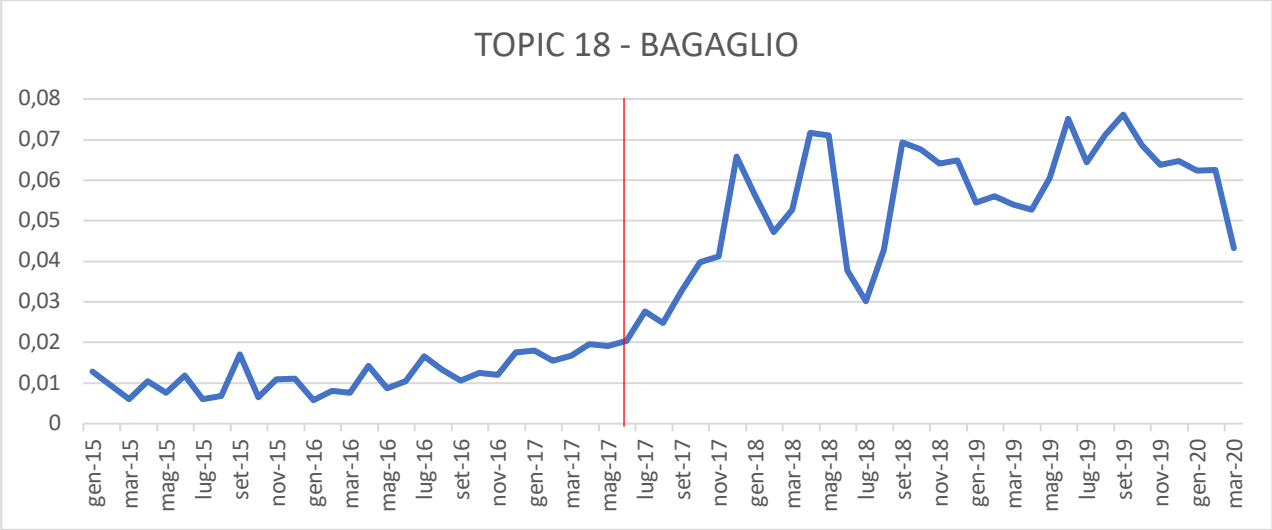


Figura 49. *Andamento temporale topic 18.*

Il grafico (figura 50) raffigura l'andamento del topic 19 (Rapporto qualità / prezzo (2)) e si può notare che eccetto per il valore outlier di gennaio 2015 vi è un trend leggermente crescente dal 2015 al 2020 con il picco in dicembre 2017 (θ_{medio}= 6,73%) e il minimo in marzo 2015 (θ_{medio}=0,52%). Non si può riscontrare una stagionalità marcata anche se è possibile notare un leggero aumento dei valori nei mesi invernali.

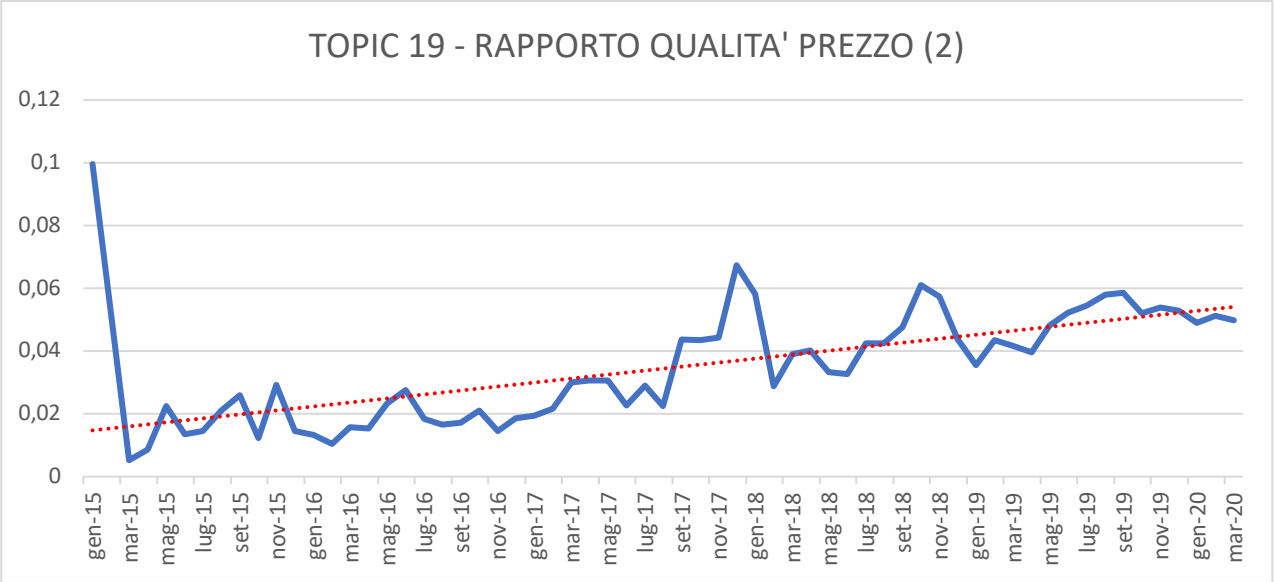


Figura 50. *Andamento temporale topic 19.*

Il topic 20 (Prenotazione volo (2)) presenta un andamento nel tempo costante, infatti, il 77,42% (48 valori sui 62 totali) dei valori di θ_{medio} sono compresi nell'intervallo [0,02 , 0,04] mentre il 22,58% (14 valori sui 64 totali) dei valori di θ_{medio} si trovano al di fuori dell'intervallo. In particolare si può notare la presenza di un outlier che rappresenta il punto di picco del grafico, ossia il valore di marzo 2020 (θ_{medio}= 14,45%). Ciò è dovuto alla presenza di alcuni valori di θ elevati che hanno innalzato il valore di θ_{medio} di marzo 2020 e ad un numero basso di recensioni riferiti a questo periodo (91 recensioni, circa lo 0,49% delle recensioni del campione di recensioni considerate per l'analisi temporale).
Questo argomento presenta un picco in marzo 2020 e la presenza di tale outlier è giustificata anche dal fatto che la prenotazione / cancellazione di un volo è stata notevolmente discussa in rete a causa dell'emergenza sanitaria COVID-19.
Infatti i clienti delle varie compagnie aeree hanno rilasciato molti commenti nelle recensioni riferiti a tale argomento a causa dell'emergenza COVID-19 e questo ha aumentato di molto i valori di prevalenza di tale topic nelle recensioni e di conseguenza è aumentato notevolmente il valore di θ_{medio} nel mese di marzo 2020.
Si riportano alcuni esempi di recensioni di marzo 2020:

- Bravissimi!!! Emergenza Corona Virus. Tutte le compagnie aeree hanno dato uno stop ai voli . Alitalia sta riportando tutti gli italiani in patria! Sono una dei tanti... con un biglietto per il 26 marzo e l'Argentina che aveva dichiarato lo stop ai voli da e' per l'Italia, ieri sono corsa in aeroporto per cercare di tornare a casa! (ero partita prima che l'emergenza scoppiasse..!!) dopo alcune ore veramente terribili in aeroporto in cui ho creduto di dover essere rimpatriata dall'unita di crisi, finalmente all'ultimo secondo un minuto prima della partenza sono stata imbarcata sul volo diretto per Roma!!! Non so descrivere quanto sia stato piacevole e quanto io abbia apprezzato l'accoglienza ospitale e calorosa ricevuta a bordo! Il personale di volo attento e cordiale, ha svolto il servizio con grande professionalita' ma anche con fare amichevole, proprio di noi italiani, che e' quello che mi ha spinto a fare questa recensione. Appena salita a bordo mi sono sentita finalmente a casa ed in luogo sicuro!! Grazie Alitalia, grazie a tutto il personale che si sta adoperando per riportarci a casa!!! Ricordiamoci di questo quando tutto sara' passato, e finalmente potremo ricominciare a spostarci senza paura!!! (recensione: Compagnia: Alitalia-autore:

Emanuela DNP – data: marzo 2020 – da: Buenos Aires a: Napoli – tipo di volo: internazionale – classe: Business – rating: 5).

- Volo puntuale, l'ultimo volo Alitalia per Milano prima della chiusura totale causa COVID. Veivolo pulitissimo, personale gentile e sempre disponibile . abbiamo viaggiato nei posto a due senza nessuno davanti, comodi. Pasti discreti. Il volo e' arrivato puntuale a destinazione e senza disagi nonostante fosse pieno (recensione: Compagnia: Alitalia – autore: tittiecarlo – data: marzo 2020 – da: Malè a: Milano – tipo di volo: internazionale – classe: Economy – rating: 5).

- Voglio ricordare questo volo, perche' non ho volato. Causa l'epidemia in Italia di Corona Virus la compagnia mi ha avvisato tramite sms della cancellazione. Devo dire che sono stati molto solleciti e professionali. Ho potuto decidere per il rimborso in piena autonomia, mediante alcuni semplici passaggi sull'applicazione di Ryanair. Il rimborso, e' stato scritto, avverra' mediante accredito su PayPal. In sette giorni lavorativi. Un bel trattamento da parte di questa low-cost che in questo periodo e' oberata di lavoro. Quando uno staff merita, bisogna parlarne bene (recensione: Compagnia: Ryanair – autore: Jacopo Masi– data: marzo 2020 – da: Roma a: Barcellona – tipo di volo: Europa – classe: Economy – rating: 5).

Figura 51. *Andamento temporale topic 20.*

L'andamento del topic 21 (Competenze personale (2)) è costante e ciò lo si può notare nel grafico (Figura 52) dalla linea rossa tratteggiata.
L'88,70 % dei valori è compreso nell'intervallo [0,08 , 0,12], solo 7 valori su 62 totali sono al di fuori di questo intervallo.
Non sono presenti stagionalità all'interno del grafico.
I θ_{medio} hanno valori elevati e ciò sta ad indicare che le competenze del personale di bordo della compagnia sono un argomento sempre molto discusso all'interno delle recensioni in tutti i mesi da gennaio 2015 a marzo 2020.
Il punto di massimo si ha in maggio 2015 (θ_{medio}= 13,40%) mentre il punto di minimo del grafico si ha a marzo 2020 (θ_{medio}= 7,35%).

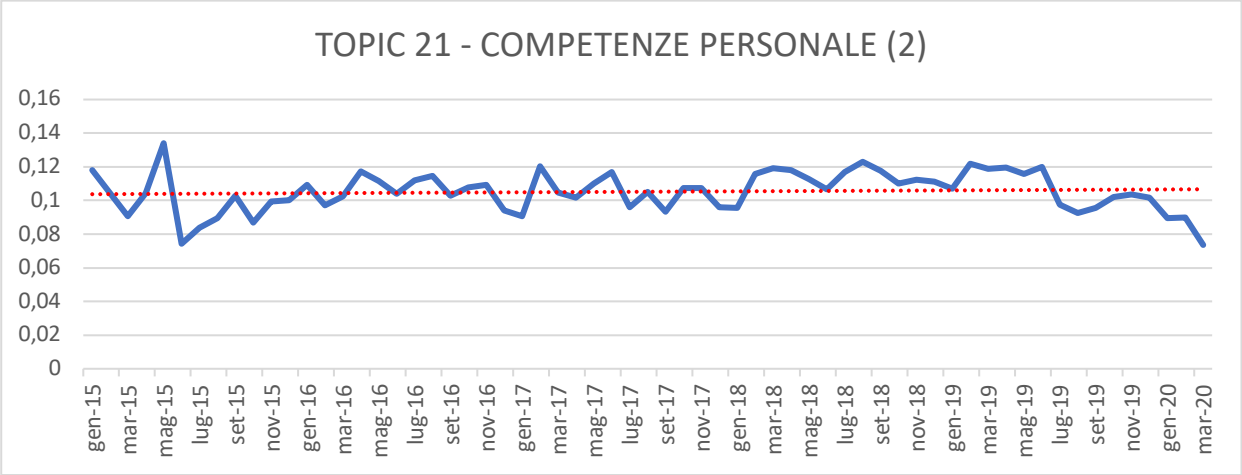


Figura 52. *Andamento temporale topic 21.*

La figura 53 presenta l'andamento nel tempo del topic 22 (Anticipo / ritardo volo) ed è possibile osservare un trend crescente da gennaio 2015 a marzo 2020. Non sono riscontrabili stagionalità.
I valori di θ_{medio} dal 2018 al 2020 sono nettamente superiori a quelli degli anni precedenti e il punto di picco del grafico è rappresentato da luglio 2019 (θ_{medio}=8,32%).
Il punto di minimo del grafico è rappresentato da maggio 2015 (θ_{medio}= 1,17%) ed è il periodo in cui il topic 22 è stato meno discusso nelle recensioni del dataset.

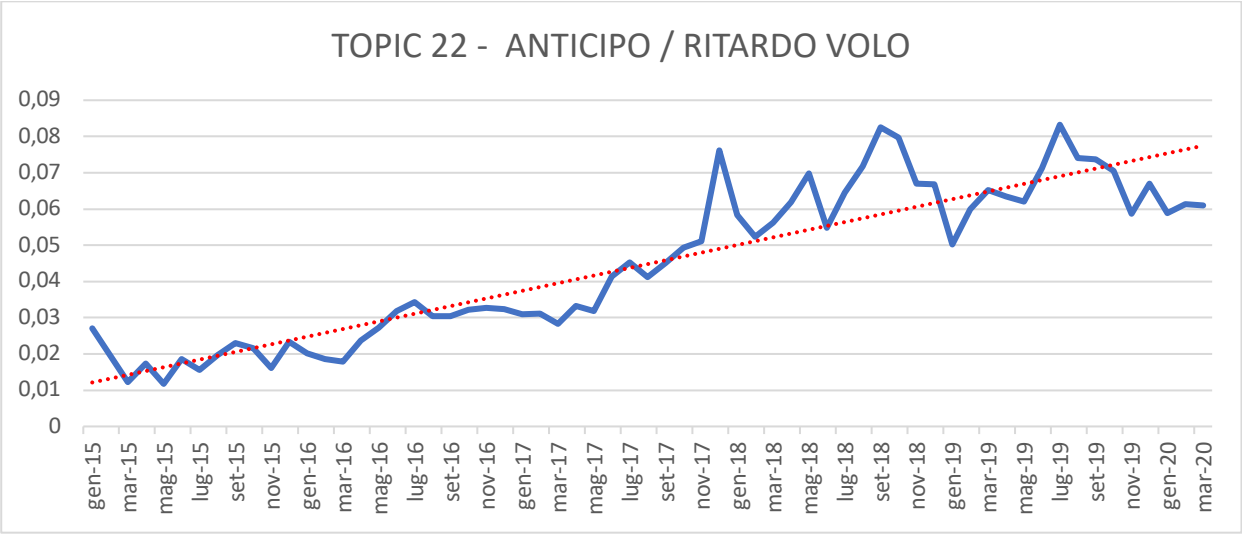


Figura 53. *Andamento temporale topic 22.*

Il topic 23 (Intrattenimento) presenta un trend decrescente (figura 54) nel tempo ed è osservabile dall'andamento della linea di tendenza lineare (linea rossa tratteggiata).
Il topic 23, come è stato discusso nell'analisi precedente sui topic, è stato l'argomento più discusso nelle recensioni riferite ai "superconnectors" (Etihad Airways, Qatar Airways ed Emirates).

Il grafico non presenta alcuna stagionalità.
Il picco si ha in maggio 2015 (θ_{medio}= 19,00%) mentre il punto di minimo si ha in marzo 2020 (θ_{medio}= 2,62%).

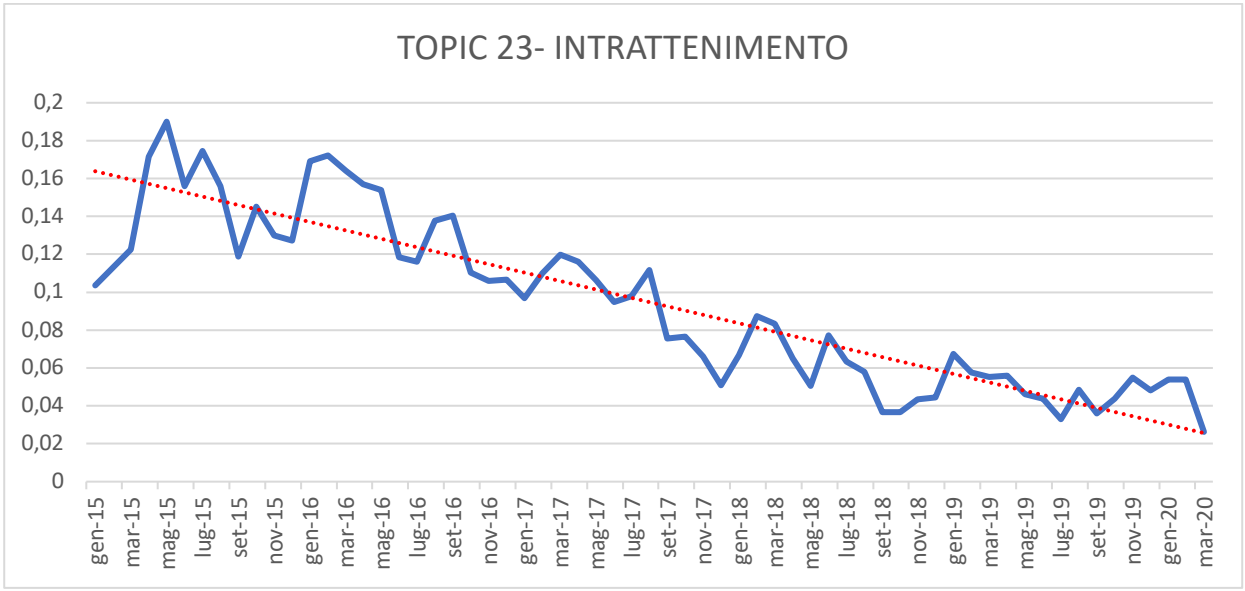


Figura 54. *Andamento temporale topic 23.*

La figura 55 mostra l'andamento del topic 24 (Lamentele dei clienti (2)) ed è possibile osservare un trend crescente (linea tratteggiata rossa) mentre non è possibile riscontrare alcuna stagionalità.
Tutti i valori seguono un trend crescente eccetto il valore outlier di gennaio 2015 che si discosta dal trend ma questo è dovuto al fatto che per il mese di gennaio sono disponibili solo due recensioni.
Il punto di massimo è rappresentato da marzo 2020 (θ_{medio}= 4,58%) mentre il punto di minimo si ha in maggio 2015 (θ_{medio}= 0,53%).

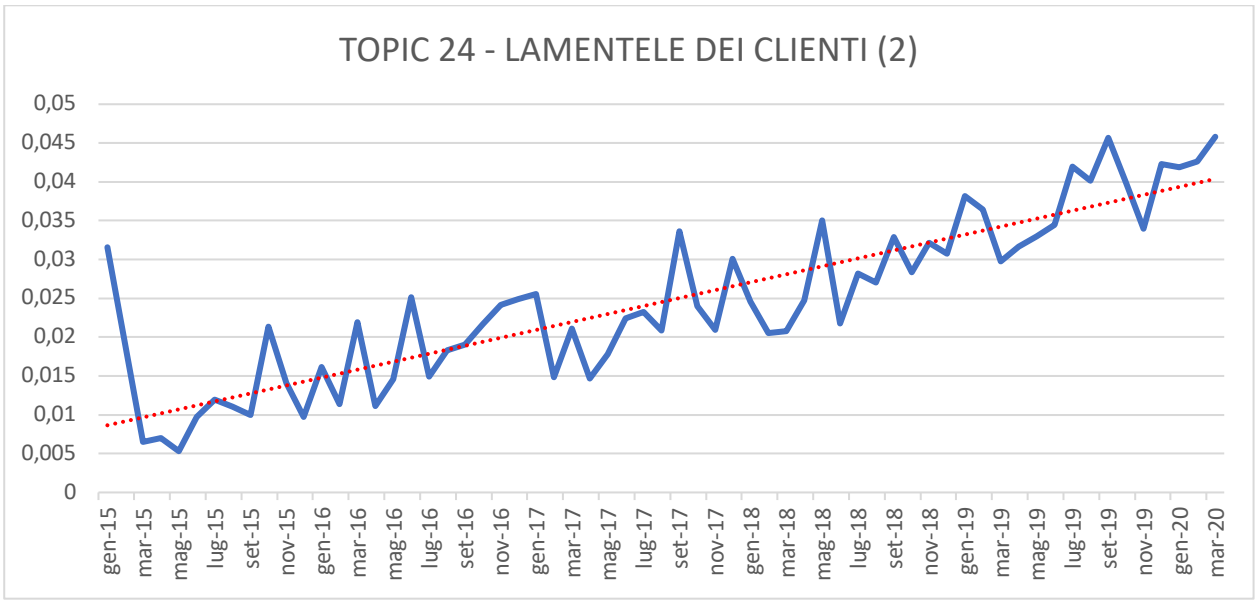


Figura 55. *Andamento temporale topic 24.*

La figura 56 mostra l'andamento del topic 25, ossia l'argomento meno discusso all'interno del dataset e ha dei valori di prevalenza e prevalenza media molto bassi in quasi la totalità delle recensioni.
L'andamento del topic 25 (Problemi con la compagnia aerea) è costante (linea rossa tratteggiata) e tutti i valori si trovano a ridosso del valore 0,015 eccetto il valore di gennaio 2015, il punto di minimo si trova in marzo 2015 (θ_{medio}= 1,15%) e il punto di massimo in luglio 2016 (θ_{medio}= 2,10%).
Non è presente alcuna stagionalità.

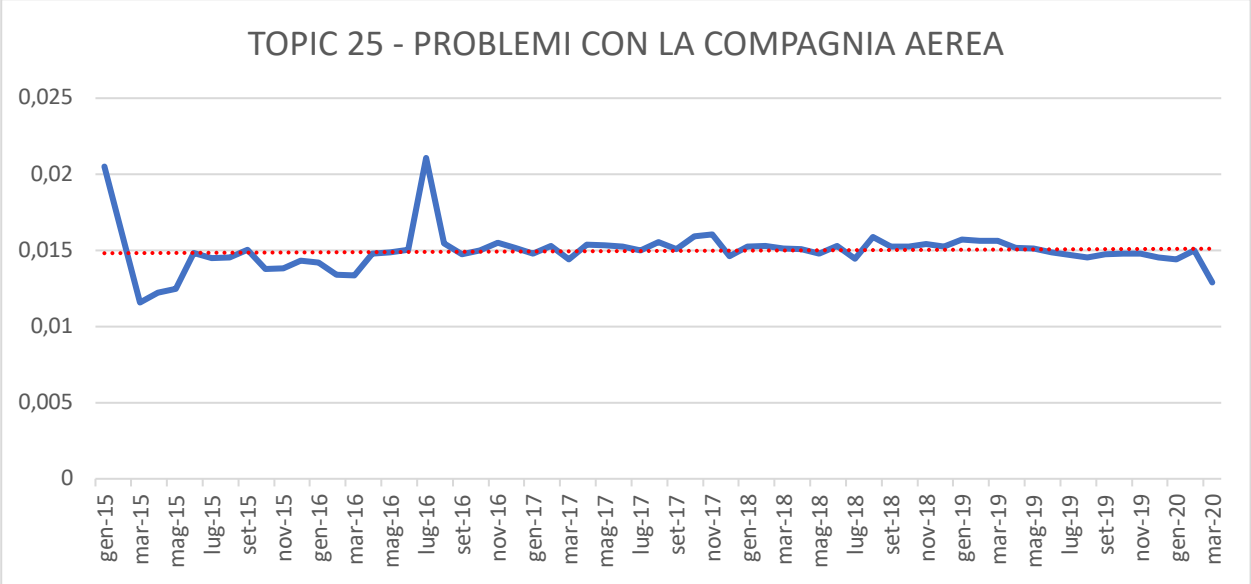


Figura 56. *Andamento temporale topic 25.*

CAPITOLO 5

5.1. CONCLUSIONI

Nel lavoro di tesi esposto nei capitoli precedenti è stata applicata la Text Mining Analysis per analizzare le recensioni estratte dal web tramite web scraping.
Per applicare la Text Mining Analysis sono state applicate in R delle funzioni delle librerie del pacchetto STM, ovvero sono state applicate delle funzioni dell'algoritmo STM.

Dopo aver applicato le funzioni dell'algoritmo STM all'interno dell'ambiente del software statistico R, si è scelto il numero ottimo di topic da analizzare, i topic sono stati etichettati attraverso la fase di etichettatura e infine si è proceduto alla validazione dell'algoritmo per poter constatare se la bontà dei valori ottenuti tramite l'algoritmo STM corrispondeva alla realtà.
Per effettuare la validazione dell'algoritmo sono stati utilizzati 4 indicatori che sono stati applicati ai 4 campioni estratti e ad un campione di dimensione maggiore costituito dalle recensioni dei 4 campioni precedenti e si è potuto verificare che l'indicatore RECALL ha dei valori leggermente superiori alla letteratura considerata mentre l'indicatore PRECISION assume dei valori leggermente inferiori, l'indicatore F – MESAURE rientra nei range della letteratura considerata mentre l'indicatore ACCURACY in tutti i campioni eccetto il terzo e il quarto è superiore alla soglia del 55 % che attesta un buon risultato di validazione dell'algoritmo STM.

In seguito le 10 compagnie aeree del campione sono state divise in base all'area geografica di appartenenza (Medio Oriente, Europa, Europa Low Cost) e al segmento di mercato a cui appartengono e poi si è proseguito con un'analisi sulla percentuale di prevalenza di ciascun topic all'interno di ciascuna recensione.

Inizialmente è stata effettuata un'analisi sul totale delle recensioni del dataset preprocessato che ha mostrato che i principali argomenti discussi sul totale delle recensioni sono le competenze del personale, l'intrattenimento a bordo e l'anticipo / ritardo di un volo, mentre gli argomenti di minor interesse dei clienti che volano con queste 10 compagnie aeree sono stati commenti riguardanti la presenza di famiglie a bordo di un volo, la possibilità di consumare snack e bevande a bordo e problemi riscontrati con la compagnia aerea.
Successivamente sono stati analizzati i risultati per le compagnie aeree delle varie aree geografiche e segmenti di mercato.
Per le compagnie aeree del Medio oriente si è osservato che gli argomenti più discussi all'interno delle recensioni e a cui i clienti danno maggior peso per la scelta della compagnia aerea con cui volare sono innanzitutto l'intrattenimento a bordo e la qualità dei servizi offerti durante il volo e le competenze del personale (hostess, steward, pilota, assistenti dell'ufficio informazioni e reclami, ecc.), mentre gli argomenti di minor interesse e che meno incidono su questo gruppo di recensioni e che influenzano meno la clientela nella scelta del volo riguardano il bagaglio a mano e i bagagli nella stiva e eventuali problemi riscontrati con la compagnia aerea dalla prenotazione alla conclusione dell'esperienza di volo.
In Europa gli argomenti più discussi dai clienti delle 5 compagnie aeree assegnate a questa categoria sono le competenze del personale di bordo, l'intrattenimento durante il volo (film, musica, riviste, ecc.) e la comodità del viaggio mentre gli argomenti meno discussi e sui quali i clienti rilasciano meno contenuti in rete sono riferiti alla presenza di famiglie a bordo e alle necessità che esse richiedono durante un volo (esempio la presenza di bambini piccoli) e la possibilità di consumare snack e bevande durante il volo al di fuori dei pasti forniti dalla compagnia aerea.

Per quanto riguarda l'Europa Low Cost gli argomenti maggiormente discussi nelle recensioni riguardano i bagagli, le competenze del personale e la puntualità e l'orario dei voli.
Gli argomenti meno discussi dai clienti delle compagnie low cost sono la possibilità di avere pasti a bordo durante il volo (soprattutto commenti di clienti che manifestano la possibilità di poter consumare cibi e bevande a bordo ma ad un prezzo elevato e che ogni servizio aggiuntivo a bordo ha un prezzo elevato), coincidenza del volo e l'argomento meno discusso è l'intrattenimento.
Questa analisi mostra che oltre ad appartenere a segmenti di mercato totalmente diversi le compagnie low cost hanno un business model totalmente diverso dalle altre compagnie europee e soprattutto totalmente opposto alle compagnie del Medio Oriente, infatti mentre i "superconnectors" basano il proprio business model sull'intrattenimento e la qualità dei servizi, le compagnie low cost non considerano proprio questo aspetto ma basano la propria strategia di mercato su altri punti di forza come il rapporto qualità / prezzo.

In seguito è stato effettuato un focus sui due principali players europei del settore low cost: Ryanair e Easyjet.
Nonostante appartengano allo stesso segmento di mercato Ryanair e Easyjet mostrano una strategia di mercato differente: Ryanair aggiunge continuamente nuove rotte mentre Easyjet punta a rafforzare le rotte esistenti, Easyjet incrementa il numero dei voli soprattutto sulle rotte più importanti mentre Ryanair punta a servire più aeroporti più che a incrementare il numero dei voli e infine il prezzo, Ryanair mantiene un surplus di guadagno più basso sul singolo cliente per accrescere il numero di clienti mentre Easyjet mantiene dei margini di profitto più elevati sul singolo cliente per massimizzare il guadagno per passeggero trasportato.
Questi aspetti in parte vengono mostrati anche dall'analisi mostrata infatti i principali argomenti discussi dai clienti Ryanair nelle recensioni rilasciate sono il Rapporto qualità / prezzo, il bagaglio e l'orario del volo mentre quelli meno discussi sono la pulizia e manutenzione dell'aeromobile, la qualità dei servizi e l'intrattenimento a bordo.
Gli argomenti più discussi nelle recensioni riferite a Easyjet riguardano il bagaglio, le competenze del personale e l'orario del volo mentre quelli meno discussi sono riferiti alla possibilità di consumare pasti a bordo, alla qualità dei servizi e all'intrattenimento.

Successivamente è stata effettuata un'analisi sulla correlazione dei vari topic ed è stato possibile identificare 4 famiglie di topic mentre i topic che non appartengono a nessuna famiglia sono stati identificati come spaiati.

In conclusione, l'ultima analisi che è stata effettuata sui 25 topic discussi all'interno delle recensioni è stata un'analisi dell'andamento temporale di ciascun topic all'interno del periodo compreso tra gennaio 2015 e marzo 2020 evidenziando per ciascuno di essi se vi è la presenza di punti di massimo locale e assoluto, punti di minimo locale e assoluto, trend e stagionalità.
E' stato possibile osservare dai grafici ottenuti che gli argomenti che sono stati sempre più discussi con il passare del tempo sono stati: la prenotazione e la cancellazione del volo, il decollo / atterraggio, il rapporto qualità / prezzo, la fase di check-in, il bagaglio, anticipo / ritardo del volo e lamentele dei clienti.
Gli argomenti che sono stati sempre meno discussi nelle recensioni tra gennaio 2015 e marzo 2020 riguardano il tipo di classe, la coincidenza con un altro volo da prendere, la comodità del viaggio, qualità dei servizi, manutenzione e pulizia dell'aeromobile e l'intrattenimento a bordo.

Le aziende utilizzano sempre di più la Text Mining Analysis per impostare le proprie strategie di marketing per supportare i propri servizi. La Text Mining Analysis permette di studiare il comportamento, le opinioni e ciò che vuole e richiede un cliente da un servizio e tramite l'analisi temporale è possibile osservare quali sono gli argomenti e ciò che si aspetta il cliente da un servizio nel tempo.

APPENDICE A

Fattori determinanti della qualità dei servizi (Modello PZB, 1985).

Fattori determinanti della qualità nei servizi	Descrizione
1. Accesso	Riguarda la possibilità di accesso e la facilità di contatto.
2. Competenza	Significa avere le know – how e le conoscenze necessarie ad eseguire il servizio.
3. Comunicazione	L'azienda deve adattare il linguaggio alla portata di ogni tipo di cliente, esprimendosi in maniera più evoluta con i clienti più colti e in termini semplici con i clienti meno esperti.
4. Cortesia	Significa avere gentilezza, rispetto, considerazione e amabilità da parte del personale di contatto.
5. Credibilità	Indica fiducia e onestà. Essa comporta l'avere a cuore gli interessi del cliente.
6. Affidabilità	Indica la corrispondenza tra prestazioni e fiducia. Affidabilità significa che l'azienda esegue il servizio nel modo giusto la prima volta e mantiene le promesse.
7. Capacità di risposta	Indica la volontà e prontezza degli addetti nel fornire il servizio.
8. Sicurezza	E' la libertà dal pericolo, dal rischio e dal dubbio.
9. Attività tangibili	Riguardano gli oggetti tangibili del servizio.
10. Capire / conoscere il cliente	Significa adoperarsi per capire i bisogni e gli eventuali problemi che il cliente ha con il servizio.

APPENDICE B

Indicatori di qualità relativi alle attività di gestione aeroportuale - settore passeggeri

Fattore di qualità	*N.*	*Indicatore*	*Unità di misura*
Sicurezza del viaggio	*1*	Livello di soddisfazione del servizio controllo bagagli nell'ottica della sicurezza	% pax soddisfatti
Sicurezza personale e patrimoniale	*2*	Numero eventi (furti e danni) alle auto nei parcheggi a pagamento segnalati al gestore	N° eventi/MPA
	3	Percezione sul livello di sicurezza personale e patrimoniale in aeroporto	% pax soddisfatti
Regolarità del servizio (e puntualità dei mezzi)	*4*	Ritardi nei voli dovuti al Gestore aeroportuale	N° ritardi/Tot. voli pax in partenza
	5	Ritardi complessivi	N° ritardi complessivi/Tot. voli pax in partenza
	6	Recupero sui tempi di transito dei voli arrivati in ritardo	% recuperi sul tempo di transito schedulato
	7	Bagagli disguidati complessivi	N° bagagli disguidati /1.000 pax in partenza
	8	Tempi di riconsegna bagagli	Tempo riconsegna del 1° e dell'ultimo bagaglio nel 90 % dei casi
	9	Tempo di attesa a bordo per lo sbarco del primo passeggero	Tempo di attesa dal Block-On nel 90% dei casi
	10	Percezione complessiva sulla regolarità dei servizi ricevuti in aeroporto	% pax soddisfatti
Pulizia e condizioni igieniche	*11*	Disponibilità toilettes	TPHP/N° toilettes
	12	Percezione sul livello di pulizia e funzionalità delle toilettes	% pax soddisfatti
	13	Percezione sul livello di pulizia in aerostazione	% pax soddisfatti
Comfort nella permanenza in aeroporto	*14*	Disponibilità di spazio per i passeggeri	mq/TPHP
	15	Disponibilità di posti a sedere	TPHP/N° sedute
	16	Disponibilità carrelli portabagagli	TPHP/N° carrelli
	17	Percezione sulla disponibilità di carrelli portabagagli	% pax soddisfatti
	18	Efficienza sistemi di trasferimento pax (ascensori, tapis-roulants, scale mobili)	% tempo funzionamento nell'orario di apertura dello scalo
	19	Percezione sull'efficienza dei sistemi di trasferimento pax	% pax soddisfatti
	20	Percezione sull'efficienza degli impianti di climatizzazione	% pax soddisfatti
	21	Percezione sulla luminosità dell'aerostazione	% pax soddisfatti
	22	Percezione sulla rumorosità in aerostazione	% pax soddisfatti
	23	Percezione complessiva sul livello di comfort	% pax soddisfatti
Servizi aggiuntivi	*24*	Disponibilità telefoni pubblici	TPHP/N° telefoni
	25	Compatibilità orario apertura bar con orario effettivo voli	% voli passeggeri in arrivo/partenza compatibili con l'orario apertura bar nelle rispettive aree
		Percezione su disponibilità/qualità/prezzi:	
	26	Negozi/edicole	% pax soddisfatti
	27	Bar	% pax soddisfatti
	28	Ristoranti	% pax soddisfatti
Servizi per passeggeri a ridotta mobilità	*29*	Disponibilità di percorsi facilitati	Si/No (specificare)
	30	Accessibilità a tutti i servizi aeroportuali	Si/No (specificare)
	31	Disponibilità di personale dedicato su richiesta	Si/No (specificare)
	32	Disponibilità di spazi dedicati	Si/No (specificare)
	33	Disponibilità di sistema di chiamata nel parcheggio	Si/No (specificare)
	34	Disponibilità di sistema di chiamata nel terminal	Si/No (specificare)
	35	Disponibilità di adeguate informazioni e comunicazioni	Si/No (specificare)
Servizi di informazione al pubblico	*36*	Disponibilità punti informazione operativi	TPHP/N° punti informazione
	37	Percezione sull'efficacia dei punti d'informazione operativi	% pax soddisfatti
	38	Presenza di segnaletica interna chiara, comprensibile ed efficace	% pax soddisfatti
	39	Percezione sulla comprensibilità degli annunci	% pax soddisfatti
	40	Percezione complessiva sull'efficacia delle informazioni(1)	% pax soddisfatti
	41	Presenza di Numero Verde / Sito Internet	Si/No (specificare)
	42	Disponibilità di punti informativi per Tour Operators	Si/No (indicazione ubicazione)
Aspetti relazionali e comportamentali	*43*	Percezione sulla cortesia del personale	% pax soddisfatti
	44	Percezione sulla professionalità del personale	% pax soddisfatti
Servizi sportello/varco	*45*	Attesa in coda alle biglietterie	Tempo nel 90% dei casi
	46	Percezione coda alla biglietteria	% pax soddisfatti
	47	Attesa in coda al check in	Tempo nel 90% dei casi
	48	Percezione coda al check in	% pax soddisfatti
	49	Tempo di attesa al controllo radiogeno dei bagagli	Tempo nel 90 % dei casi
	50	Attesa in coda controllo passaporti arrivi/partenze (2)	Tempo massimo nel 90 % dei casi
	51	Percezione coda al controllo passaporti	% pax soddisfatti
Integrazione modale (efficacia collegamenti città-aeroporto)	*52*	Disponibilità, frequenza, puntualità e prezzo collegamenti bus/treno (3) /taxi	% pax soddisfatti
	53	Collegamenti stradali città/aeroporto (4)	% pax soddisfatti
	54	Presenza di segnaletica esterna chiara, comprensibile ed efficace	% pax soddisfatti

(1) Eventualmente dedotta dalle informazioni già raccolte. (2) Onde permettere tale rilevazione, il 13.3.2001 è stata diramata dal Min.dell'Interno specifica informativa a tutte le sedi della Polaria. (3) Ove esistente. (4) Servizi non compresi fra le attività di gestione aeroportuale, ma di opportuna rilevazione da parte del gestore.

Idicatore da misurarsi mediante sondaggio

APPENDICE C

Funzioni del pacchetto STM utilizzate in R.

```
> library("stm")
library("topicmodels")
library("slam")
library ("SnowballCC")
library ("tm")


dataset <- read.csv2("TUTTI I DATI.csv",na.strings==NULL)


# Preprocessing #

set.seed (23456)
processed <- textProcessor(documents = dataset$documents, metadata = dataset)
out <- prepDocuments(documents = processed$documents,
                     vocab = processed$vocab,
                     meta = processed$meta)
docs <- out$documents
vocab <- out$vocab
meta <- out$meta

plotRemoved(processed$documents, lower.thresh = seq(1, 200, by = 100))

# Fitting the model #

out <- prepDocuments(documents = processed$documents,
                     vocab = processed$vocab,
                     meta = processed$meta, lower.thresh = 15)
shortdoc <- substr(out$meta$documents, 1, 200)

# VALUTAZIONE NUMERO DI TOPIC IDEALE
# 4 DIAGNOSTIC DIMENSIONS (held-out likelihood, Residuals, Semantic Coherence, Lower Bound) #
c=(5:50) #c vettore con numeri da 5 a 50
K<-c
storage <- searchK(out$documents, out$vocab, K, data = meta)
plot(storage)

datasetPrevFit <- stm(documents = out$documents, vocab = out$vocab,
                      K = 25, prevalence =~ Company + out$meta$mark ,
                      max.em.its = 75,
                      data = out$meta, init.type = "Spectral")
```

```
# Model selection #

datasetSelect <- selectModel(out$documents, out$vocab, K = 25,
prevalence =~ Company + out$meta$mark , max.em.its = 75, data = out$meta, runs = 20, seed = 8458159)

plotModels(datasetSelect)


# Describing the datasetPrevFit model #

labelTopics(datasetPrevFit, c(1:25))

thoughts1 <- findThoughts(datasetPrevFit, texts = shortdoc,
                          n = 2, topics = 1)$docs[[2]]

thoughts2 <- findThoughts(datasetPrevFit, texts = shortdoc,
                          n = 2, topics = 2)$docs[[2]]

thoughts3 <- findThoughts(datasetPrevFit, texts = shortdoc,
                          n = 2, topics = 3)$docs[[2]]

thoughts4 <- findThoughts(datasetPrevFit, texts = shortdoc,
                          n = 2, topics = 4)$docs[[2]]

thoughts5 <- findThoughts(datasetPrevFit, texts = shortdoc,
                          n = 2, topics = 5)$docs[[2]]

thoughts6 <- findThoughts(datasetPrevFit, texts = shortdoc,
                          n = 2, topics = 6)$docs[[2]]

thoughts7 <- findThoughts(datasetPrevFit, texts = shortdoc,
                          n = 2, topics = 7)$docs[[2]]

thoughts8 <- findThoughts(datasetPrevFit, texts = shortdoc,
                          n = 2, topics = 8)$docs[[2]]

thoughts9 <- findThoughts(datasetPrevFit, texts = shortdoc,
                          n = 2, topics = 9)$docs[[2]]

thoughts10 <- findThoughts(datasetPrevFit, texts = shortdoc,
                          n = 2, topics = 10)$docs[[2]]

thoughts11 <- findThoughts(datasetPrevFit, texts = shortdoc,
                          n = 2, topics = 11)$docs[[2]]

thoughts12 <- findThoughts(datasetPrevFit, texts = shortdoc,
                          n = 2, topics = 12)$docs[[2]]

thoughts13 <- findThoughts(datasetPrevFit, texts = shortdoc,
                          n = 2, topics = 13)$docs[[2]]

thoughts14 <- findThoughts(datasetPrevFit, texts = shortdoc,
                          n = 2, topics = 14)$docs[[2]]
```

```
thoughts15 <- findThoughts(datasetPrevFit, texts = shortdoc,
                           n = 2, topics = 15)$docs[[2]]

thoughts16 <- findThoughts(datasetPrevFit, texts = shortdoc,
                           n = 2, topics = 16)$docs[[2]]

thoughts17 <- findThoughts(datasetPrevFit, texts = shortdoc,
                           n = 2, topics = 17)$docs[[2]]

thoughts18 <- findThoughts(datasetPrevFit, texts = shortdoc,
                           n = 2, topics = 18)$docs[[2]]

thoughts19 <- findThoughts(datasetPrevFit, texts = shortdoc,
                           n = 2, topics = 19)$docs[[2]]

thoughts20 <- findThoughts(datasetPrevFit, texts = shortdoc,
                           n = 2, topics = 20)$docs[[2]]

thoughts21 <- findThoughts(datasetPrevFit, texts = shortdoc,
                           n = 2, topics = 21)$docs[[2]]

thoughts22 <- findThoughts(datasetPrevFit, texts = shortdoc,
                           n = 2, topics = 22)$docs[[2]]


thoughts23 <- findThoughts(datasetPrevFit, texts = shortdoc,
                           n = 2, topics = 23)$docs[[2]]

thoughts24 <- findThoughts(datasetPrevFit, texts = shortdoc,
                           n = 2, topics = 24)$docs[[2]]

thoughts25 <- findThoughts(datasetPrevFit, texts = shortdoc,
                           n = 2, topics = 25)$docs[[2]]
```

```
par(mfrow = c(2, 1), mar = c(.5, .5, 1, .5))
plotQuote(thoughts1, width = 40, main = "Topic 1")
plotQuote(thoughts23, width = 40, main = "Topic 23")

meta$rating <- as.factor(meta$rating)
prep <- estimateEffect(1:25 ~ Company + out$meta$mark , datasetPrevFit,
                       meta = out$meta, uncertainty = "Global")
summary(prep, topics = 1)

plot(datasetPrevFit, type = "summary", xlim = c(0, .3))

# Correlations #

#GRAFO RELAZIONI TRA TOPICS (CORRELAZIONE) metodo "simple"#
#Tutti i topic sono spaiati#
mod.out.corr <- topicCorr(datasetPrevFit,method="simple", cutoff=0.35)
mod.out.corr
plot(mod.out.corr)

#Clustering tramite la modifica del cutoff#

mod.out.corr <- topicCorr(datasetPrevFit,method="simple", cutoff=0.15)
mod.out.corr
plot(mod.out.corr)
```

APPENDICE D

Algoritmo in Python (estratto da GitHub) utilizzato per applicare la tecnica web scraping.

```
# link to Airlines list https://www.tripadvisor.it/Airlines

import time
from selenium import webdriver
import Airline
import argparse
# Some functions
def scapeReviewsFromReviewBox(reviewBoxes):
    reviews = []
    for i in range(5):
        author = reviewBoxes[i].find_element_by_css_selector(
            'a.social-member-event-MemberEventOnObjectBlock__member--35-jC').text

        flightDate = reviewBoxes[i].find_elements_by_css_selector(
            'span.location-review-review-list-parts-EventDate__event_date--1epHa')

        # optional review attribute replace null if no exist
        if len(flightDate) != 0:
            flightDate = flightDate[0].text.replace('Data del viaggio: ', '', 2)
        else:
            flightDate = 'NULL'

        # grab flight labels like trip, type of the flight, type of the flight
        labels = reviewBoxes[i].find_elements_by_css_selector(
            'div.location-review-review-list-parts-RatingLine__labelBtn--e58BL')

        flightFromCity = labels[0].text.split(' - ', 2)[0]
        flightToCity = labels[0].text.split(' - ', 2)[1]
        flightType = labels[1].text
        flightClass = labels[2].text

        # retrieve title of the review
        title = reviewBoxes[i].find_element_by_css_selector(
            'a.location-review-review-list-parts-ReviewTitle__reviewTitleText--2tFRT').text

        # expand text all boxes (always present)
        if i == 0:
            reviewBoxes[i].find_element_by_css_selector(
                'span.location-review-review-list-parts-ExpandableReview__cta--2mR2g').click()
```

```
        # retrieve text of the review
        text = reviewBoxes[i].find_element_by_css_selector(
            'q.location-review-review-list-parts-ExpandableReview__reviewText--gOmRC').text
        # remove new line characters and other junk
        text = text.replace('<br>', ' ')
        text = text.replace('\n', ' ')
        # get the class of the number of bubble MARK
        mark = str(reviewBoxes[i].find_element_by_css_selector('span.ui_bubble_rating').get_attribute('class'))
        # parse the class text
        mark = float(mark.split('_', 20)[-1]) / 10

        review = Airline.Review(author=author,
                                flightDate=flightDate,
                                flightFromCity=flightFromCity,
                                flightToCity=flightToCity,
                                flightType=flightType,
                                flightClass=flightClass,
                                mark=mark,
                                title=title,
                                text=text)

        reviews.append(review)

    return reviews


def main(args):

    driver = webdriver.Chrome('./chromedriver')  # Optional argument, if not specified will search path.

    driver.get(args.company_link);
    #
    time.sleep(15)  # Let the user actually see something!

    companyName = driver.find_element_by_class_name(
        'flights-airline-review-page-airline-review-header-AirlineDetailHeader__airlineName--2JeT1').text

    nReviews = driver.find_element_by_class_name('ui_poi_review_rating  ')
    nReviews = int(nReviews.text.split(' ', 2)[0].replace('.', '', 6))

    avgMark = driver.find_element_by_class_name(
        'flights-airline-review-page-overview-module-OverviewModule__overall_rating--30Bld').text
    avgMark = float(avgMark.split('\n', 2)[0].replace(',', '.', 5))


    print(companyName)
    print(nReviews)
    print(avgMark)
```

```
    airline = Airline.Airline(companyName=companyName,
                              nReviews=nReviews,
                              avgMark=avgMark)

    #iterate over the pages
    while len(airline.reviews)<nReviews and len(airline.reviews) < args.max_rev:

        # there will be 16 boxes bcause this class references to a box that is common for reviews, photos and suggestions tabs. we care only about the first 5 indices of the following list
        reviewBoxes = driver.find_elements_by_xpath(
            '//div[@class="location-review-card-Card__ui_card--2Mri0 location-review-card-Card__card--o3LVm location-review-card-Card__section--NiAcw"]')

        airline.reviews.extend(scapeReviewsFromReviewBox(reviewBoxes))

        #driver.find_element_by_xpath(
        #    '//a[@class="ui_button nav next primary"]').click()
        driver.find_element_by_link_text('Avanti').click()
        #
        time.sleep(7)

    print(airline)

    for r in airline.reviews:
        print(r)

    if args.csv:
        import csv
        data_list = [['author',
                      'flightDate',
                      'flightFromCity',
                      'flightToCity',
                      'flightType',
                      'flightClass',
                      'mark',
                      'title',
                      'text'
                      ]]
        with open(airline.companyName+'_'+str(len(airline.reviews))+'_of_'+str(airline.nReviews)+'_'+str(airline.avgMark)+'.csv', 'w', newline='') as file:
            writer = csv.writer(file, delimiter='|')
            writer.writerows(data_list)
            writer.writerows(airline.getListReviews())

    # time.sleep(5) # Let the user actually see something!
    driver.quit()



if __name__ == '__main__':
    parser = argparse.ArgumentParser()
    parser.add_argument("--company-link", type=str, help="paste the trip advisor link of the company you want to know about",
                        nargs='?', default='https://www.tripadvisor.it/Airline_Review-d8729018-Reviews-Alitalia', const=0)
    parser.add_argument("--max-rev", type=int, help="max reviews to scrape, set to 50 default",
                        nargs='?', default='50', const=0)
    parser.add_argument("--csv", dest='csv', action='store_true', default=False, help="to produce a csv file")

    args = parser.parse_args()

    main(args)
```

APPENDICE E

Lista Custom Stop Words eliminate tramite la funzione “textProcessor”.

“volo”,”voli”,”compagnia”,”compagnie”,”compagn”,”compagni”,”abbstanza”,”abbia”,”abbiamo”,”agli”,”alcune”,”alcuni”,”alla”,”alle”,”allo”,”altra”,”altro”,”altre”,”altri”,”anche”,”ancora”,”andare”,”andata”,”andate”,”andato”,”andavate”,”andavamo”,”andiamo”,”appena”,”detto”,”deve”,”devi”,”devo”,”dice”,”dicendo”,”dico”,”dicono”,”dir”,”dire”,”dopo”,”dove”,”dovevo”,”dovrebbe”,”dovrebbero”,”dovuta”,”dovuto”,”due”,”durante”,”ecc”,”entrare”,”era”,”erano”,”eravate”,”ero”,”miei”,”milano”,”mio”,”modo”,”molta”,”molti”,”molto”,”neanche”,”negli”,”nei”,”nel”,”nella”,”nelle”,”nemmeno”,”nessun”,”nessuna”,”nessuno”,”niente”,”noi”,”nome”,”non”,”nonostante”,”nostro”,”nostra”,”nulla”,”oggi”,”ogni”,”oltre”,”ora”,”ore”,”ormai”,”resto”,”riesce”,”riguarda”,”sara‘“,”sarebbe”,”seconda”,”sei”,”sembra”,”sempre”,”senza”,”sia”,”siamo”,”siano”,”siete”,”solo”,”sono”,”soprattutto”,”sotto”,”spesso”,”sta”,”stata”,”stati”,”stato”,”appunto”,”arrivata”,”assolutamente”,”aver”,”avere”,”aveva”,”avevano”,”avevo”,”avuto”,”ben”,”bene”,”bisogna”,”buon”,”buona”,”buono”,”caso”,”c’era”,”c’ero”,”c’erano”,”certo”,”che”,”chi”,”circa”,”come”,”comunque”,”con”,”cosa”,”cose”,”cosi”,”essere”,”faccio”,”fanno”,”far”,”farci”,”fare”,”fate”,”fatta”,”fatto”,”fine”,”forse”,”forte”,”fosse”,”fuori”,”gia’“,”giorno”,”giorni”,”giu’“,”gli”,”gran”,”hai”,”hanno”,”ieri”,”parte”,”per”,”perche”,”perche‘“,”pero”,”pero‘“,”piu”,”piu‘“,”poca”,”poche”,”pochi”,”poi”,portato”,”posso”,”possono”,”poter”,”potete”,”potrebbe”,”prendere”,”presso”,”prima”,”probabilmente”,”proprio”,”puo”,”puo‘“,”pur”,”stessa”,”stesso”,”sto”,”sua”,”sul”,”sulla”,”sullo”,”suo”,”tanta”,”tante”,”tanto”,”tipo”,”tra”,”tre”,”troppa”,”troppo”,”trova”,”trovare”,”trovata”,”trovato”,”trovo”,”tuo”,”tutta”,”tutte”,”tutti”,”tutto”,”una”,”uno”,”un’ora”,”cosi‘“,”credo”,”cui”,”da”,”dai”,”dal”,”dalla”,”dalle”,”dare”,”dato”,”davvero”,”degli”,”dei”,”del”,”della”,”delle”,”dello”,”inoltre”,”invece”,”italia”,”loro”,”lui”,”mai”,”mandato”,”meglio”,”meno”,”mezzo”,”mia”,”mie”,”purtroppo”,”qua”,”qualche”,”quale”,”quando”,”quanto”,”quasi”,”quattro”,”quel”,”quella”,”quello”,”questa”,”queste”,”questi”,”questo”,”qui”,”quindi”,”raggiungere”,”recata”,”vado”,”vari”,”varie”,”vengo”,”vengono”,”venivo”,”veramente”,”vero”,”verso”,”via”,”viene”,”visto”,”voi”,”volta”,”voto”,”lufthansa”,”emir”,”emirates”,”alitalia”,”british”,”airways”,”viaggi”,”ritorno”,”busi”,”business”,”roma”,”doha”,”dubai”,”bagag”,”bagagli”,”ne’“,”ne”,”personal”,”persona”,”new”york”,”londra”,”pagar”,”pagare”,”euro”,”anno”,”anni”,”volt”,”volte”,”unora”,”air”,”franc”,”france”,”nessuno”,”francoforte”,”francofort”,”madrid”,”barcellona”,”ryanair”,”iberia”,”etihad”,”monaco”,”easyjet”,”easy”,”jet”,”america”,”bangkok”,”bogota”,”bogotã”,”grazi”,”grazie”,”prego”,”¬”,”molt”,”inglese”,”italiano”,”francese”,”decisamente”,”decisament”,”personali”,”chiesto”,”â”,”poco”,”moltitudine”,”molte”,”abu”,”dhabi”,”dabi”,”arrivati”,”parigi”,”secondo”,”gia”,”iu”,”bologna”,”dacqua”,”d’acqua”,”neo”,”partito”,”qatar”,”business”,”volato”,”diverso”,”divers”,”diversi”,”diversa”,”diverse”,”portar”,”portare,”,”particolarmente”,”allandata”,”all’andata”,”malpensa”,”linate”,”alghero”,”cuore”,”london”,”bologna”,”roma”,”torino”,”milano”.

APPENDICE F

Lista dei 25 topic con le relative 7 parole Highest Prob, FREX, LIFT e SCORE.

ETICHETTA	HIGHEST PROB.	FREX	LIFT	SCORE	ETICHETTA	HIGHEST PROB.	FREX	LIFT	SCORE
1. Prenotazione volo (1)	pagamento	costa	nazionalità	nazionalità	2. Tipo di classe	economy	premium	assicuro	assicuro
	biglietto	soldi	spesi	soldi		class	soddisfatto	economy	economy
	posto	vuoi	vuoi	pagamento		viaggiato	rimasto	world	class
	costa	comprare	truffa	biglietto		volta	deluso	peggioramento	premium
	pagato	costa	compri	pagato		rispetto	raggio	deluso	rimasto
	doppio	supplemento	costa	carta		aereo	intercontinentali	premium	raggio
	soldi	doppio	cercano	doppio		raggio	recensioni	usano	deluso
3. Famiglia	posto	bambini	consigliarla	consigliarla	4. Pasti a bordo	cibo	vecchio	imparare	imparare
	bambini	famiglia	passeggini	bambini		colazione	colazione	croissant	colazione
	piccoli	piccoli	bimbi	famiglia		cena	datato	immangiabile	cena
	famiglia	figlio	adulti	piccoli		servito	freddo	tavolini	vecchio
	vicini	parecchi	vicini	vicini		poco	immangiabile	formaggio	servito
	marito	bimbi	parecchi	marito		pasto	bagno	scomoda	freddo
	figlia	vicini	posto	figlio		vecchio	scadente	pane	cibo
5. Coincidenza volo	scalo	san	affermare	affermare	6. Comodità viaggio	viaggio	viaggio	swiss	viaggio
	destinazione	destinazione	pietroburgo	scalo		consiglio	piacevole	rilassante	swiss
	coincidenza	scalo	intermedio	destinazione		hostess	consiglio	agio	piacevole
	tratta	heathrow	messico	coincidenza		piacevole	gentili	piacevole	consiglio
	diretto	transito	las	città		viaggiato	cordiali	cordiali	gentili
	tempo	coincidenza	francisco	san		sicuramente	rilassante	sorridenti	comodo
	città	francisco	san	francisco		comodo	consigliare	viaggio	lungo

7. Cancellazione volo	giorno	cancellato	Pasqua	Pasqua	**8. Qualità servizi (1)**	servizi	servizi	minimo	servizi
	aeroporto	giorno	sostenuto	giorno		livello	complesso	igienici	minimo
	rimborso	spese	alloggio	cancellato		media	ottimo	offerti	ottimo
	perso	hotel	cancellato	rimborso		ottimo	offerti	servizi	buono
	cancellato	successivo	riprotezione	perso		buono	positivo	essenziali	livello
	successivo	albergo	albergo	hotel		bordo	buono	complesso	complesso
	notte	sera	hotel	successivo		complesso	precisi	precisi	media

9. Snack & bevande	mangiare	acqua	spiacente	spiacente	**10.Decollo/ atterraggio**	atterraggio	pilota	gola	gola
	bere	bere	santo	acqua		aereo	atterraggio	bravo	atterraggio
	acqua	panini	domingo	bere		decollo	decollo	vento	decollo
	piedi	fondo	panini	piedi		passeggeri	condizioni	pilota	pilota
	bicchiere	solita	povero	bicchiere		pilota	comandante	discesa	comandante
	file	avanti	ritrovati	dovere		pista	pista	atterraggio	discesa
	passeggeri	carrello	solita	file		condizioni	vento	pista	pista

11. Manutenzione e pulizia	aerei	aerei	venuta	aerei	**12. Qualità servizi (2)**	servizio	eccellente	town	ottimo
	pulito	nuovi	aerei	venuta		ottimo	impeccabile	cape	servizio
	nuovi	puliti	nuovi	puliti		bordo	lounge	letti	town
	migliori	aeromobili	moderni	nuovi		top	top	impeccabile	eccellente
	tratta	tenuti	puliti	aeromobili		eccellente	eccezionale	ineccepibile	lounge
	aeromobili	vecchi	generazione	confortevoli		migliore	curato	curato	impeccabile
	confortevoli	lunghi	lunghi	vecchi		class	catering	eccellente	curato

13. Rapporto qualità/prezzo (1)	prezzo	rapporto	opportunità	prezzo	**14. Competenze personale (1)**	puntualità	gentilezza	signori	signori
	costo	prezzo	rapporto	opportunità		cortesia	professionalità	eleganza	puntualità
	qualità	economico	competitivo	rapporto		gentilezza	cortesia	precisione	cortesia
	rapporto	costo	qualità/ prezzo	costo		pulizia	puntualità	gentilezza	gentilezza
	biglietto	low-cost	imbattibile	qualità		professionalità	disponibilità	professionalità	professionalità
	rispetto	qualità/ prezzo	prezzo	economico		disponibilità	precisione	rapidità	pulizia
	low-cost	basso	economico	qualità/prezzo		comodità	efficienza	estrema	disponibilità

15. Chech-in					16. Lingua				
	gate	gate	costare	costare		italiano	inglese	funzionare	funzionare
	fila	capire	documento	gate		nota	negativa	glutine	inglese
	persone	signori	addetto	fila		inglese	francese	odore	italiano
	passeggeri	corsa	ammassati	banco		hostess	speci	miglia	nota
	controlli	banco	passaporti	arriviamo		negativa	funzionare	francese	negativa
	chech-in	documento	signorina	capire		paio	nota	parlare	continuo
	hostess	controlli	capire	documento		lingua	parlare	negativa	lingua

17. Posto a bordo					18. Bagaglio				
	spazio	gambe	eccessivamente	spazio		bagaglio	mano	irrigidita	bagaglio
	sedili	spazio	angeles	gambe		mano	trolley	zainetti	mano
	gambe	sufficiente	los	eccessivamente		stiva	stiva	tracolla	stiva
	posto	benissimo	stretto	sedili		trolley	bagaglio	borsetta	trolley
	spazio	angeles	uscita	stretto		valigia	borsa	borsa	irrigidita
	stretto	los	gambe	benissimo		imbarcare	misura	mano	borsa
	poco	stretto	allungare	sufficiente		cabina	borsetta	zaino	valigia

19. Rapporto qualità/prezzo (2)					20. Prenotazione volo (2)				
	low	cost	proporzione	low		chech-in	agosto	foglio	foglio
	cost	low	stracciati	cost		prenotazione	the	aiutato	prenotazione
	prezzo	prezzo	competitivi	prezzi		online	online	the	online
	costi	vinci	vinci	proporzione		problema	line	agosto	check-in
	volta	gratta	lotteria	gratta		line	effettuare	time	the
	vuole	vendita	gratta	vinci		sito	errore	book	agosto
	pagare	profumi	cost	costi		momento	risolvere	aiutarmi	mail

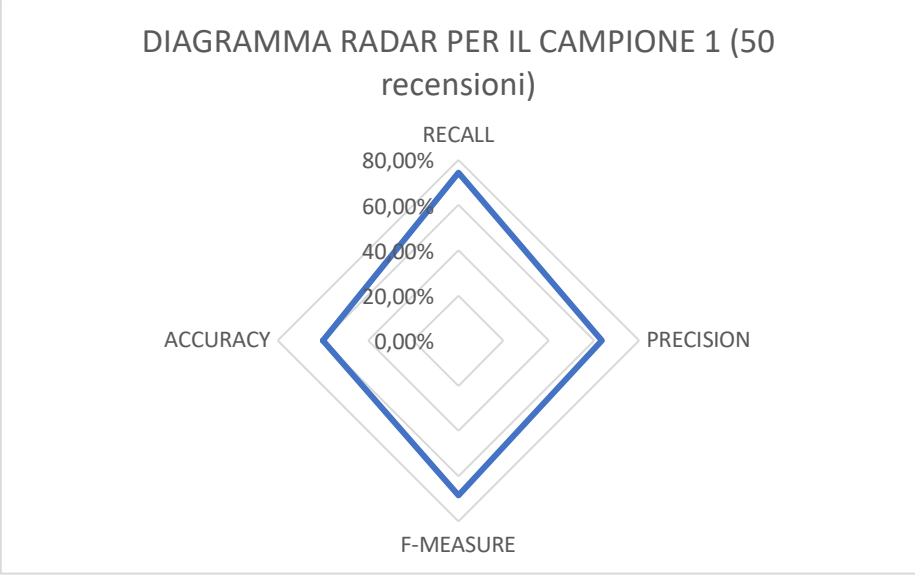
21. Competenze personale (2)					22. Anticipo/ ritardo volo				
	personale	gentile	fuerteventura	personale		ritardo	arrivi	Puntual-mente	puntualmente
	bordo	disponibilità	cordiale	gentile		partenza	partenza	recuperato	ritardo
	gentile	cortese	gentile	fuerteventura		minuti	partiti	leggero	minuti
	disponibile	pulito	pulito	disponibile		orario	ritardo	ritardo	partenza
	puntuale	cordiale	professionale	puntuale		anticipo	recuperato	orario	orario
	aereo	professionale	qualificato	cortese		arrivi	minuti	lieve	arrivo
	cortese	puntuale	preparato	pulito		partito	partito	arrivi	anticipo

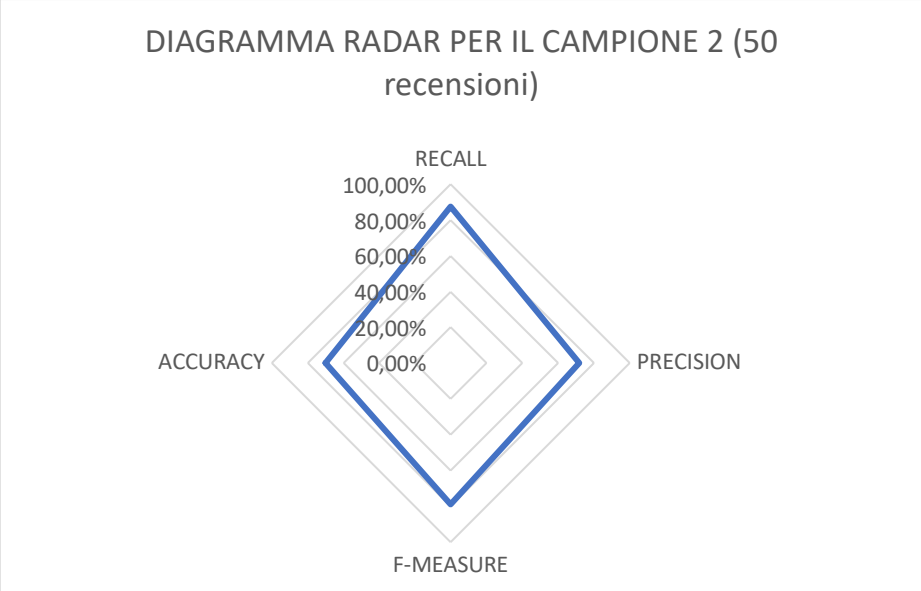
23.Intrattenimento	cibo	film	speziato	film	**24. Lamentele dei clienti**	pessimo	scortese	prendere	prendere
	film	giochi	videogiochi	Intratteni-mento		ritardi	pessimo	aria	ritardi
	scelta	Intratteni-mento	documentari	speziato		poco	condizio-nata	aereo	pessima
	intrattenimento	musica	dentifricio	cibo		aria	aria	condizio-nata	scortese
	pasti	comodi	calze	musica		personale	personale	pessima	condizio-nata
	bevande	ampia	vasta	giochi		aereo	poco	scortese	peggior
	comodi	vasta	tappi	pasti		terra	terra	imbarazzante	vergogna

25. Problemi riscontrati con la compagnia aerea	servizio	peccato	compromesso	compromesso
	bordo	solito	alimenti	servizio
	problemi	problemi	certezza	bordo
	peccato	giusto	idem	problemi
	tratta	tratta	birra	peccato
	solito	penso	peccato	tratta
	volta	certezza	giusto	solito

APPENDICE G

Diagrammi radar per i 4 indicatori della validazione dell'algoritmo.





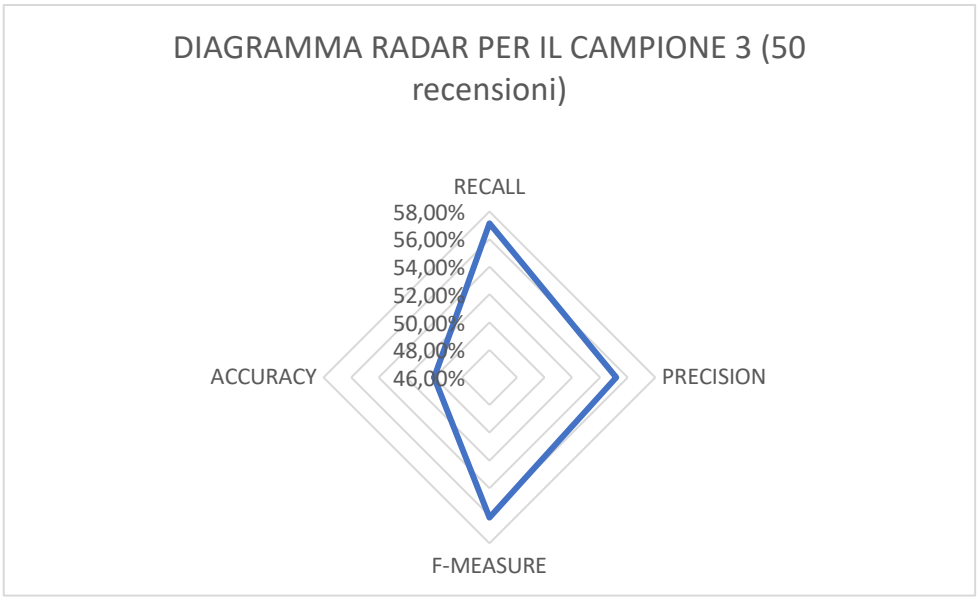


DIAGRAMMA RADAR PER IL CAMPIONE 3 (50 recensioni)
RECALL
58,00%
56,00%
54,00%
52,00%
50,00%
48,00%
46,00%
PRECISION
F-MEASURE
ACCURACY

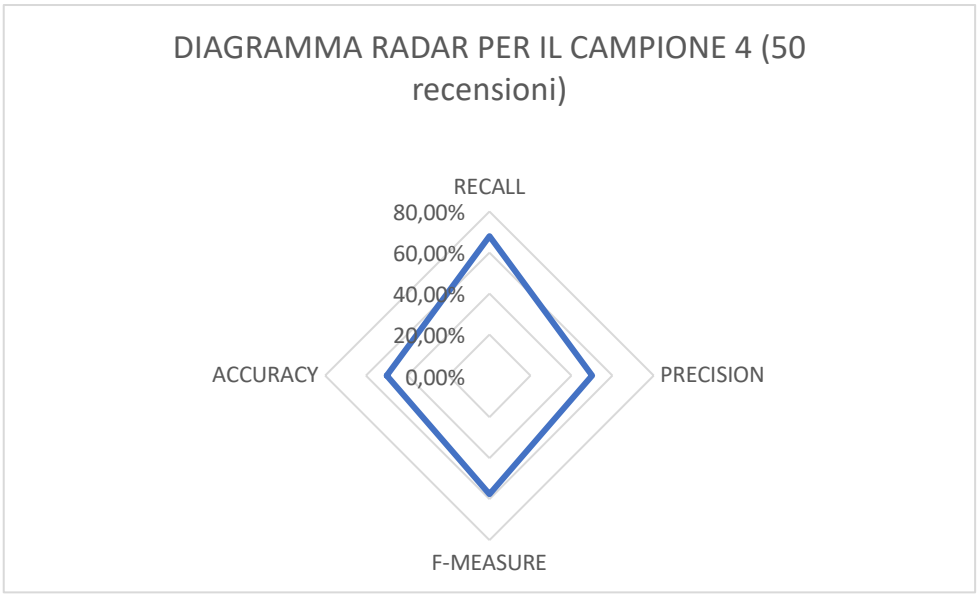


DIAGRAMMA RADAR PER IL CAMPIONE 4 (50 recensioni)
RECALL
80,00%
60,00%
40,00%
20,00%
0,00%
PRECISION
F-MEASURE
ACCURACY

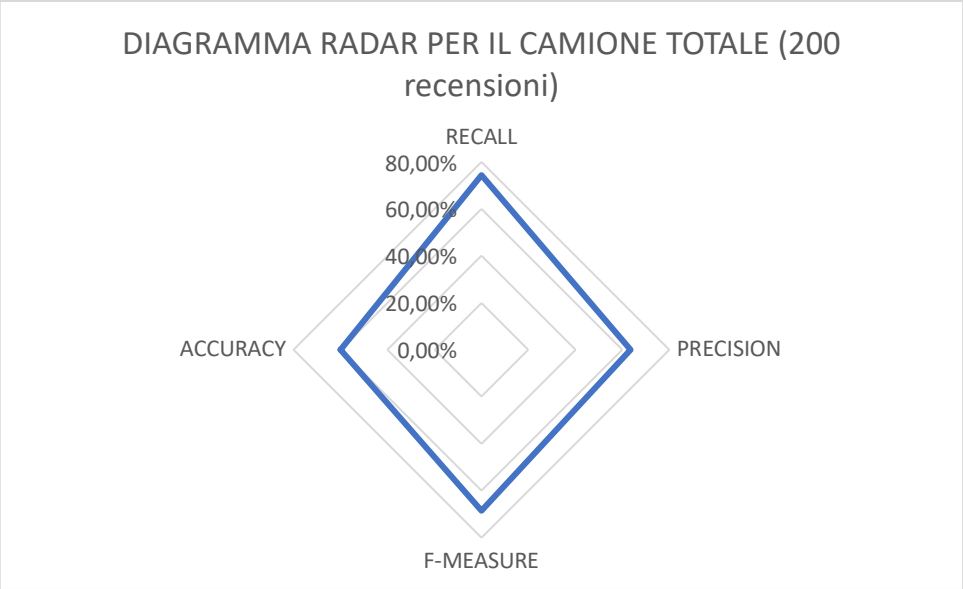


DIAGRAMMA RADAR PER IL CAMIONE TOTALE (200 recensioni)
RECALL
80,00%
60,00%
40,00%
20,00%
0,00%
PRECISION
F-MEASURE
ACCURACY

BIBLIOGRAFIA E SITOGRAFIA

"Algoritmo in Python per fare web scraping", GitHub, link:https://github.com/gigpir/DynamicWebScraper

https://www.tesionline.it/tesi/brano/confronto-ryanair-e-easyjet/11961

"Feedback e Opinioni", Il Mio Volo Cancellato, link: https://www.ilmiovolocancellato.it/sito/feedback/

"Il caso Trip Advisor, recensioni false e l'importanza del monitoraggio del proprio brand sul web", link: https://www.sitiwebshop.it/blog-webshop/123-il-caso-tripadvisor-le-recensioni-false-e-l-importanza-del-monitoraggio-del-proprio-brand-sul-web.html

"L'inarrestabile successo delle compagnie aeree del Golfo", Economist, 7 maggio 2015, link: https://www.ilpost.it/2015/05/07/compagnie-aeree-emirates-etihad-qatar-turkish/

"Market share of major airlines in India in financial year 2019 based on international traffic", fonte: Statista, link: https://www.statista.com/statistics/643889/market-share-of-leading-airlines-india/

"Modello di business low cost nel trasporto aereo – 116onfront Ryanair e Easyjet", link: https://www.tesionline.it/tesi/brano/confronto-ryanair-e-easyjet/11961

"User generated content", Wikipedia, link: https://en.wikipedia.org/wiki/User-generated_content

"Voli e recensioni per le Compagnie aeree", TripAdvisor, link: https://www.tripadvisor.it/Airline_Review-d10533118-Reviews-La-Compagnie

A.Bene, V. Cantarelli, F. Carlucci, U. Colombo, P. Piccari, A. Ruscitti, U. Turello *"Manuale di controllo di qualità e affidabilità",* ISEDI Istituto Editoriale Internazionale, Milano, settembre 1974.

A.Calabrese, Nathan Levialdi Ghiron, *"Qualità dei servizi nel "XXI secolo",* Enciclopedia Treccani, 2010, link: http://www.treccani.it/enciclopedia/qualita-dei-servizi_%28XXI-Secolo%29/

A.Di Bartolomeo, *"Ryanair ed EasyJet: ultime regole bagaglio a mano e costi per chi non le rispetta",* Investire Oggi – Quotidiano economico finanziario, 21 agosto 2019.

A.Gelbukh, *"Computational Linguistics and Intelligent Text Processing",* Springer International Publishing, Switzerland, 2015.

A.K. Nassirtoussi, S. Aghabozorgi, I Ying Wah, David Chek Ling Ngo, *"Text mining for market prediction: A systematic review",* University of Malaya, Kuala Lumpur, Malaysia, 9 giugno 2014, link : https://romisatriawahono.net/lecture/rm/survey/information%20retrieval/Nassirtoussi%20-%20Text%20Mining%20for%20Market%20Prediction%20-%202014.pdf

A.M. Mineo "Guida all'utilizzo dell'Ambiente Statistico R", Dipartimento di Scienze Statistiche e Matematiche "S. Vianelli", Università degli Studi di Palermo

A.Minini, *"Web scraping"*, link: http://www.andreaminini.com/seo/web-scraping

A.Parasuraman, VA. Zeithaml, LL. Berry, *"Refinement and reassessment of the SERVQUAL scale",* Journal of retailing, 1991.

A.Parasuraman, VA. Zeithaml, LL. Berry, *"Refinement and reassessment of the SERVQUAL scale",* Journal of retailing, 2002.

A.Parasuraman, VA. Zeithaml, LL. Berry, *"SERVQUAL: a multiple – item scale for measuring consumer perceptions of service quality",* Journal of retailing, 1988.

A.Visentin, *"Identificazione delle dimensioni di qualità latenti tramite l'analisi di "user generated contents" (UGC) con algoritmi di text mining: il caso delle strutture ospedaliere italiane",* Politecnico di Torino, Marzo 2020.

A.Zenarolla, *"Costruire qualità sociale: indicazioni teoriche e operative per lo sviluppo della qualità nei servizi",* Franco Angeli s.r.l. , Milano, Italy , 2007.

B. Liu, *"Sentiment Analysis and Opinion Mining",* Morgan & Claypool Publishers, Maggio 2012.

B.Pang, L. Lee, S. Vaithyanathan, *"Thumbs up? Sentiment classification using machine learning techniques",* 2002, Conference on Empirical methods in Natural Language Processing (EMNLP), link: https://arxiv.org/pdf/cs/0205070.pdf

B.Simonetta, *“eCommerce: perché le recensioni negative possono essere un’opportunità”*, il Sole24ore, 7 gennaio 2019.

C. Dal Monte, *“Perché è importante gestire le recensioni online”,* agosto 2019, link: https://www.soiel.it/news/dettaglio/perche-importante-gestire-le-recensioni-online/

DM. Blei, JD. Lafferty, B. Gao, I Bose, *“A correlated topic model of science”*, Princeton University and Carnegie Mellon University, 2007.

E. Marro, *“Ryanair, Easyjet e le altre: le 6 “tasse occulte” delle compagnie aeree low cost”,* il Sole24ore, 17 novembre 2018.

ENAC, *“La qualità dei servizi nel trasporto aereo – Le carte dei servizi standard – Linee Guida”,* 31 ottobre 2014 link: https://www.enac.gov.it/sites/default/files/allegati/2018-Lug/All.1_Linee_guida.pdf

F. Domanico, *“Il trasporto aereo passeggeri in Europa: Contestable Theory, Core Theory o semplicemente Low Cost Carriers?”,* Rivista Italiana di Politiche Pubbliche,2006.

F. Formica, *“Trasporto aereo”,* La Repubblica, 29 aprile 2020, link: https://www.repubblica.it/argomenti/trasporto_aereo

F. Franceschini, *“Dai prodotti ai servizi – Le nuove frontiere per la misura della qualità*”, UTET Libreria Srl, Torino, maggio 2001.

F. Franceschini, M. Galetto, D.A. Maisano, L. Mastrogiacomo *“Ingegneria della Qualità Applicazioni ed Esercizi”,* CLUT, Torino, ottobre 2016.

F. Gilardi, Charles R. Shipan, B. Wuest, *“The diffusion of policy perceptions: evidence from a structural topic model”,* University of Zurich, 2015.

G. Miner, J. Elder IV, A. Fast, T. Hill, R. Nisbet, D. Delen, *“Practical Text Mining and Statistical Analysis for Non – structured Text Data Applications”,* Academic Press is an imprint of Elsevier
225 Wyman Street, Waltham, MA 02451, USA The Boulevard, Langford lane, Kidlington, Oxford, OX5 1GB, UK , 2012.

G.Amati, S.Angelini, A,Caterina Carli, G.Gambosi, D.Pasquini, G.Rossi, P.Vocca, *"Analisi temporale degli eventi su Twitter",* link: http://www.isticom.it/documenti/rivista/rivista_2017_2018/03_analisi_twitter.pdf

Hongning Wang, Duo Zhang, ChengXiang Zhai, *"Structural Topic Model for Latent Topical Structure Analysis",* Department of Computer Science University of Illinois at Urbana-Champaign Urbana IL, 61801 USA.

INSIDE MARKETING, *"User Generated Content"*, link: https://www.insidemarketing.it/glossario/definizione/user-generated-content/

J. Joseph Cronin, Jr. & Steven A. Taylor, *"SERVPERF versus SERVQUAL: Reconciling Performance – Based and Perceptions – Minus – Expectations Measurement of Service Quality",* Journal of Marketing, 1994, link: https://journals.sagepub.com/doi/pdf/10.1177/002224299405800110
JF O'Connell, "The rise of the Arabian Gulf carriers: An insight into the business model of Emirates Airline", Journal of Air Transport Management, 2011.

L. Cillis, "*Aerei, Italia preda delle low cost: ora sono il 51% del mercato",* Economia&Finanza, 5 giugno 2018.

L. Mastrogiacomo, F. Barravecchia, F. Franceschini, F. Marimon *"Mining quality determinants of Product-Service Systems from unstructured User-Generated Contents: The case of car-sharing"* , Politecnico di Torino (DIGEP), Corso Duca degli Abruzzi 24, 10129, Torino (Italy).

L. Pinto, *"Structural Topic Model per le scienze sociali e politiche",* aprile 2019.

Lewis, S. C., Zamith, R., & Hermida, A. (2013), *"Content Analysis in an Era of Big Data: A Hybrid Approach to Computational and Manual Methods.",* Journal of Broadcasting & Electronic Media

M. E. Roberts, B.M. Stewart, D. Tingley, E.M. Airoldi *"The Structural Topic Model and Applied Social Science",* 2013.

M. E. Roberts, B.M. Stewart, D.Tingley *"STM: An R Package for Structural Topic Models",* Journal of Statistical Software.

M. Roberts, *“Structural Topic Models”,* UC San Diego, 25 maggio 2017.

N Hu, T Zhang, B Gao, I Bose, *“What do hotel customers about? Text analysis using structural topic model”.*

PC. Cacciabue, I. Oddone, I. Rizzolo, *“Sicurezza del trasporto aereo”,* Springer, 2010.

Pengtao Xie, Eric P. Xing, *“Integrating Document Clustering and Topic Modeling”,* 2013, link: https://arxiv.org/pdf/1309.6874.pdf

R. Lawson, *“Web Scraping with Python”,* Published by Packt Publihing Ltd. , ottobre 2015.

R. Mitchell, *“Web Scraping with Python: Collecting More Data from the Modern Web”,* pubblicato da O’Reilly Media Inc., 1005 Gravenstein Highway North, Sebastopol, California, USA, 2018.

Regolamenti ENAC (Ente Nazionale per l’Aviazione Civile).

S. Colapaoli, *“Strategie e performance delle compagnie aeree low cost. I casi Ryanair, Easyjet e Virgin Express”,* 2003, link: https://www.tesionline.it/tesi/strategie-e-performance-delle-compagnie-aeree-low-cost-i-casi-ryanair-easyjet-e-virgin-express/10428

S. Franco, *“Studio delle determinanti della qualità nei Product Service Systems tramite tecniche di text mining”,* Politecnico di Torino, Anno Accademico 2019-2020.

S. Velupillai, H. Dalianis, M. Hassel, *“Developing a Standard for De-Identifying Electronic Patient Records Written in Swedish: Precision, Recall and F-measure in a Manual and Computerized Annotation Trial”,* 2009, link: https://pubmed.ncbi.nlm.nih.gov/19482543/

S. Salustri *“ 7 Strumenti gratuiti per fare Scraping”,* 25 aprile 2017 link: https://www.stefanosalustri.com/blog/7-strumenti-gratuiti-per-fare-scraping/

S. Sensini *“Come fare web scraping”,* 11 Novembre 2019 link: https://www.apogeonline.com/articoli/come-fare-web-scraping-con-python-serena-sensini/

T. Daugherty, MS Eastin, L. Bright, *"Exploring Consumer Motivations for Creating User Generated Content",* Journal of interactive advertising, 2008, link: https://d1wqtxts1xzle7.cloudfront.net/6591554/Exploring_Consumer_Motivations_for_Creating_User-Generated_Content.pdf?response-content-disposition=inline%3B+filename%3Dexploring_consumer_motivations_for_creat.pdf&Expires=1593010459&Signature=f39xSSq0PsraoSL0F0mo~s6CCcrCKfvd9J0GoffYR9IieIobpQkHdu8w-7wkftoljnckNHO4uUqOL2CuvWnPEHJ6ejH3TLCng5dcSgbpuMHQSIRpJZS0GkDb6PuOo9gx6bbbFqm705sKbypWtH6hUisbO0Ux~sUyV82VUKGgANpE0YKAL6ikbYvxzS-Szw9nip6nGzUFQM4yzzlkFRnOYPcY2mGl4eMBJyz0pqpPglcoJ60tPidU88jDGtXzbEkji1SogmglXRJSVP5VnFE3-e9XlGkdGBOvAwv3ujU2~FShbTULgFB7nX3qLPSresBHA3~WVhnhHk-aRRASQ56KZA__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA

T. Hospedales, Shaogang Gong, Tao Xiang, *"A Markov Clustering Topic Model for Mining Behaviour in Video",* School of Electronic Engineering and Computer Science Queen Mary University of London, London E1 4NS, UK , 2009.
Trasporto aereo – Autorità di regolazione dei trasporti, link: https://www.autorita-trasporti.it/trasporto-aereo/

U. Arrigo, A. Giuricin, *"Gli effetti della liberalizzazione del trasporto aereo e il ruolo delle compagnie low cost un confronto USA – Europa",* Università di Pavia, settembre 2006.

W. Kasper, M. Vela, "Sentiment Analysis for Hotel Reviews", Germany, 2011 link: https://d1wqtxts1xzle7.cloudfront.net/41728359/finalVersion_KasperVela_2012.pdf?1454075333=&response-content-disposition=inline%3B+filename%3Dsentiment_Analysis_for_Hotel_Reviews.pdf&Expires=1592831149&Signature=VwRL6kb~lPECQjiZhFIbqlf35yM7zSU1usG7eaHzvNblDFHylxczJkyXjM1o5ev8DrvIWcaaDajfiNie3f~slbdlu37045Ugy6T8kstyqEtA3qczwizK48-u6CAQxCqXnVfYtaap37tsBcMLWaxTBDdeSH-OYQrlQHpvkA9ntDqq55meltGQch89kyMqTrfoZbFIOsDtUQV5tpd3BXE88Xk6rar5dDgp3eilEjfj0e6Alrz-8b4dk7wM6j42G~lKpsfUnuhqLKy3mku7WyxAcFfzEOPPRfpbXxPaG9pJXXtN1IhqsKW6CstFQfEaz7aGTV6j1dGLrBGt-i5hwZvr6g__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA

W. Sun, C. Chou, A.W. Stacy, H. Ma, J.Unger, P. Gallaher, *"SAS and SPSS macros to calculate standardized Cronbach's alpha using the upper bound of the phi coefficient for dichotomous items",* University of Southern California, Los Angeles, California, 2007.

Xingwei Zhu, Zhao-Yan Ming, Xiaoyan Zhu, Tat-Seng Chua, *"Topic Hierarchy Construction for the Organization of Multi-Source User Generated Contents",* State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology, Department of Computer Sci. and Tech., Tsinghua

University 2Department of Computer Science, School of Computing, National University of Singapore, Singapore , Agosto 2013.

Yuanpeng J. Huang, Robert Powers, Gaetano T. Montelione, *"Protein NMR Recall, Precision, and F-measure Scores (RPF Scores): Structure Quality Assessment Measures Based on Information Retrieval Statistics",* 17 maggio 2004.

Yuen – Hsien Tseng, Chi – Jen Lin, Yu – l Lin, *"Text mining techniques for patent analysis",* Information processing and Management, 2007.

Z. Chen, B. Liu, *"Topic Modeling using Topics from Many Domains, Lifelong Learning and Big Data",* International conference on machine learning, 2014.

RINGRAZIAMENTI

Innanzitutto vorrei ringraziare il mio relatore di tesi, il Prof. Luca Mastrogiacomo per avermi seguito e aiutato nella stesura del mio lavoro di tesi di Laurea Magistrale, inoltre vorrei ringraziare il Prof. Fiorenzo Franceschini e il Prof. Federico Barravecchia che insieme al Prof. Luca Mastrogiacomo hanno scaturito in me un grande interesse per il mondo dei servizi e in particolare per la qualità nei servizi.

Desidero fare il ringraziamento più grande ai miei genitori, che mi hanno sempre sostenuto durante tutto il mio percorso universitario e non solo. Mi sono sempre stati vicini sia nei momenti di difficoltà quando non riuscivo a passare un esame o ero triste o abbattuto e sia nei momenti di felicità condividendo con me i miei successi ed essendo orgogliosi per quello che stavo facendo.
Li ringrazio dal profondo del cuore per avermi dato la passibilità di studiare e per essermi stati vicino durante tutto questo tempo, devo a loro quello che sono diventato e che sto diventando, grazie ai loro insegnamenti e ai sani principi con i quali mi hanno cresciuto.

Un ringraziamento speciale va ai miei nonni materni e a mia zia Mariagrazia perché oltre a volermi un mondo di bene, hanno sempre pregato per me e sono sempre stati orgogliosi e felici per i miei successi universitari.
Li ringrazio davvero dal profondo del cuore perché mi sono sempre stati vicini e perché nonostante la lontananza, poiché studiando a Torino ci vedevamo meno, mi hanno sempre fatto sentire la loro vicinanza e so che su di loro potrò sempre contare.

Un grande ringraziamento va a te, Giulia, insieme abbiamo passato tutto il periodo universitario e ci siamo sempre stati l'uno per l'altro e siamo cresciuti insieme e con me hai condiviso tutti i momenti, sia quelli tristi che quelli felici che ci sono stati durante il mio percorso universitario. Ti ringrazio perchè mi sei sempre stata vicina in ogni momento e ti ringrazio per il bene che mi hai sempre voluto.

Infine vorrei ringraziare i fratelli del collegio Universitario Villa San Giuseppe, il luogo meraviglioso che mi ha ospitato per tutti gli anni in cui sono stato a Torino per frequentare il Politecnico.
Ringrazio dal profondo del cuore Fratel Igino , Fratel Alessandro, Fratel Arcangelo, Fratel Antonio, Fratel Gianluigi che sono stati dei veri educatori e mi hanno aiutato e guidato durante tutto il mio percorso di studi.
Sono stato molto felice di aver trascorso il mio periodo di studi in Villa San Giuseppe perché ho avuto modo di conoscere molte persone e ho stretto delle belle amicizie che dureranno nel tempo.