UNIVERSITÁ DEGLI STUDI DI MILANO-BICOCCA

Scuola di Economia e Statistica

Corso di laurea in Scienze Statistiche ed Economiche

Analisi di testi dal web e tecniche di Text Mining: recensioni di ristoranti italiani a Londra

Relatore: Prof. Roberto Boselli

Tesi di Laurea di:

Aina Belloni

Matricola 829768

Anno Accademico 2019/2020

Indice

Introduzione

Il presente elaborato tratta un tema attualmente molto rilevante che è l'analisi dei dati testuali del web. L'avvento e la rapida crescita delle risorse tecnologiche e del Web per la condivisione di opinioni come i social media e i siti di rating, come TripAdvisor, hanno aperto nuovi orizzonti per analizzare e comprendere il pensiero e i gusti dei consumatori. Identificare le informazioni che possono facilitare il processo decisionale tra migliaia di pagine web e social media feed rappresenta un processo complesso e dispendioso. In quest'ottica, le tecniche di text mining ed opinion mining aiutano a estrarre i dati di testo in modo più efficiente offrendo la capacità di trasformare le informazioni testuali in conoscenze utili.

L'obiettivo di questa tesi è quindi quello di estrarre informazioni a partire dall'opinione delle persone contenuta, nel mio caso, nelle recensioni di ristoranti, ossia trasformare dati non strutturati in dati accessibili e comprensibili a tutti.

Il lavoro si è suddiviso in diverse fasi, la prima è stata il web scraping delle recensioni su TripAdvisor, utilizzando un programma in Python, si è passati poi ad altri due software, Orange e Rapidminer, che hanno consentito la pulizia dei dati e tutto il pre-processing.

Il focus infine volge su tecniche di Text Mining, Sentiment Analysis e Aspect Based Sentiment Analysis dirette a individuare e quantificare il contenuto sentimentale del testo e a scoprire cluster naturali di recensioni.

1. I siti di rating

Oggi viviamo in un contesto nel quale l'opinione della gente è un aspetto fondamentale, le persone sono sempre più critiche e anche grazie alla recente proliferazione di siti e applicazioni del Web (come le chat, blog, conferenze, e-commerce, social media, …) l'impulso di esprimere ognuno le proprie idee è in continuo aumento.

Le piattaforme che consentono la condivisione delle proprie opinioni sono sempre di più e stanno guadagnando sempre più importanza, esse facilitano la ricerca di informazioni ed influenzano il processo di decisione da parte del cliente. Tra le più utilizzate, se non la più utilizzata, troviamo TripAdvisor.

1.1. TripAdvisor

TripAdvisor fu fondato negli Stati Uniti a Febbraio del 2000 da Stephen Kaufer, è un sito di viaggi leader a livello mondiale, permette agli utenti di facilitare la pianificazione dei propri viaggi e di fornire recensioni su hotel, ristoranti o altre attrazioni locali. Comprende più di 200 mila strutture, 30 mila destinazioni ed attualmente sono presenti oltre 830 milioni di recensioni. Si basa sull'idea che i viaggiatori o utenti si affidino al sito e all'opinione degli altri utenti per prendere decisioni in merito.

Ciò che lo rende così importante è la grande quantità di *contenuto generato dagli utenti* (dall'inglese User-Generated Content, UGC) presente, ossia un contenuto creato dagli utenti e pubblicato in Internet tramite piattaforme di rete sociale.

Offre così un vantaggio sia per i clienti che in questo modo hanno la possibilità di comparare strutture, raccogliere informazioni riguardanti la qualità e il prezzo in modo da poter scegliere quella più adatta a loro; sia per i manager o proprietari delle strutture stesse che sono spinti a migliorare certi aspetti per mantenere un'immagine positiva, la cosiddetta web reputation e in modo da incrementare la customer loyalty.

L'informazione relativa all'UGC ha subito un rapido incremento negli ultimi anni e ha portato molte organizzazioni a voler analizzare l'opinione dei clienti. Diventa così fondamentale saper applicare combinazioni di tecniche di analisi dei testi e dati non strutturati per estrarre informazioni utili dalle semplici opinioni espresse in linguaggio naturale tramite questi canali.

Attraverso il metodo di analisi dei contenuti, O'Connor (2010) nel suo studio ha confermato che i dati presentati su TripAdvisor sono significativi e appropriati da utilizzare durante la pianificazione dei viaggi, dal momento che un qualsiasi utente può consultare tutto ciò che gli altri pubblicano, valutazioni quantitative, qualitative, oltre a foto e immagini. Considerando il numero di visitatori del sito è evidente che il contenuto viene consultato.

Nella seguente figura (Figura 1) possiamo osservare, grazie a uno studio condotto nel 2010, che il fattore che più influisce nella scelta di una struttura è l'esperienza degli altri viaggiatori.

Figura 1

2. Text mining e dati non strutturati

E' proprio dall'esigenza di approfondire la notevole utilità dell'User-Generated Content che nasce il Text Mining, ovvero un'analisi che estrae informazioni dai testi per fini predittivi, di indagine o statistici. Permette in particolare di trasformare un testo (dati non strutturati) in dati strutturati tramite l'impiego di algoritmi di Natural Language Processing.

Quando parliamo di Text Mining dobbiamo ricordarci che esso è parte di una realtà più ampia chiamata Analisi automatica dei testi (AAT) che comprende:

- Analisi lessicale (linguaggio)
- Analisi testuale (discorso)
- Text mining (strutturazione ed estrazione di informazione dai testi)
- Text Analytics (applicazione di algoritmi di analisi a testi gia strutturati)

Essa si serve di statistica, information retrieval e linguistica computazionale per ottenere dati strutturati.

Focalizzando l'attenzione sul Text Mining, possiamo dire che è un campo multidisciplinare, che impiega diverse tecniche tra cui :

- Clustering
- Sentiment Analysis
- Aspect-based Sentiment Analysis
- Topic Extraction
- Categorizzazione del testo

2.1. Recensioni

Una recensione è un resoconto scritto di un'esperienza di viaggio, condiviso sul portale con altri viaggiatori. Si compone di:

- Valutazione da 1 (pessimo) a 5 (eccellente)
- Titolo
- Testo
- Eventuali immagini

Oltre alle valutazioni, il sito incoraggia l'elaborazione attraverso i commenti in modo che gli altri utenti ottengano il maggior numero di informazioni possibili. Il testo di una recensione è ciò che la caratterizza maggiormente ed essendo un testo scritto in linguaggio naturale lo si classifica come dato non strutturato. I dati non strutturati si differenziano da quelli strutturati o semi-strutturati in quanto hanno una forma interna che non si può ricondurre ad un database o a campi pre-progettati. I dati strutturati invece sono quei dati che possono essere memorizzati o organizzati all'interno di un database.

Scrivere una recensione, oltre ad essere un modo di esprimere un opinione positiva piuttosto che negativa, ha il vantaggio di poter considerare diversi criteri come la soddisfazione generale, il rapporto qualità-prezzo, la pulizia, il servizio, e molti altri.
Per questo risulta interessante non solo analizzare la recensione nel suo complesso ma anche divisa per criteri o argomenti.

2.2. Aspect-based sentiment analysis

La Sentiment Analysis come abbiamo anticipato è una tecnica di Text Mining che permette di determinare la polarità positiva o negativa di un testo libero non strutturato. Pang, Lee e Vaithyanathan furono i primi ad applicare metodi di Machine Learning su recensioni in modo da suddividerle in 3 categorie : positive, negative e neutre. Spesso però questa semplice suddivisione non è sufficiente, le recensioni ad esempio contengono al loro interno diversi aspetti (come il servizio, la pulizia, il prezzo, la qualità, …), alcuni dei quali possono essere positivi ed altri negativi, per analizzare quindi diversi aspetti si utilizza la Aspect-Based Sentiment Analysis. Nel settore turistico in particolare può essere più importante valutare ogni singolo aspetto del servizio, piuttosto che la soddisfazione generale di ogni cliente.

Xiang ed altri ricercatori analizzarono la relazione tra i diversi aspetti e la valutazione generale. Hanno analizzato i dati applicando il metodo LDA (Latent Dirichlet Allocation), un metodo di modellazione degli argomenti per identificare i 5 temi principali.

Da questi studi si è rilevato che gli aspetti più rilevanti sono:

- Basic Service
- Value
- Landmarks&Attractions
- Dining&Experience
- Core Product

3. Analisi di recensioni di TripAdvisor

Fino ad ora abbiamo osservato quanto sia fondamentale la condivisione sul Web delle opinioni da parte dei consumatori, illustrando le principali tecniche di estrazione delle informazioni utili. In questo capitolo mettiamo in pratica alcune delle tecniche principali di text mining per ricavare dati utili da recensioni di TripAdvisor, focalizzando l'attenzione sui ristoranti.

3.1. Fase di scraping e descrizione dataset

La prima fase del lavoro è stata raccogliere i dati necessari. La tecnica utilizzata prende il nome di Web Scraping e consiste nell'estrarre informazioni e dati dal Web. L'obiettivo nel mio caso era quello di scaricare le valutazioni e i testi delle recensioni di TripAdvisor, in particolare di ristoranti italiani a Londra e grazie ad un codice Python sul notebook Jupyter creato appositamente per dati provenienti da TripAdvisor ho raccolto 15533 recensioni da 41 diversi ristoranti.

Per la scelta dei 41 ristoranti si sono utilizzati i seguenti filtri :

- Luogo : London city
- Cucina : Pizza, Northern-Italian, Central-Italian, Southern-Italian, Italian, Tuscan, Romana, Lazio, Sicilian, Neapolitan, Campania, Sardinian, Emilian
- Lingua recensioni: English

Lo script ha permesso di raccogliere tutte le recensioni per ogni link di ristorante e trascrivere i dati in un file Excel costituito dai seguenti attributi per ogni recensione:

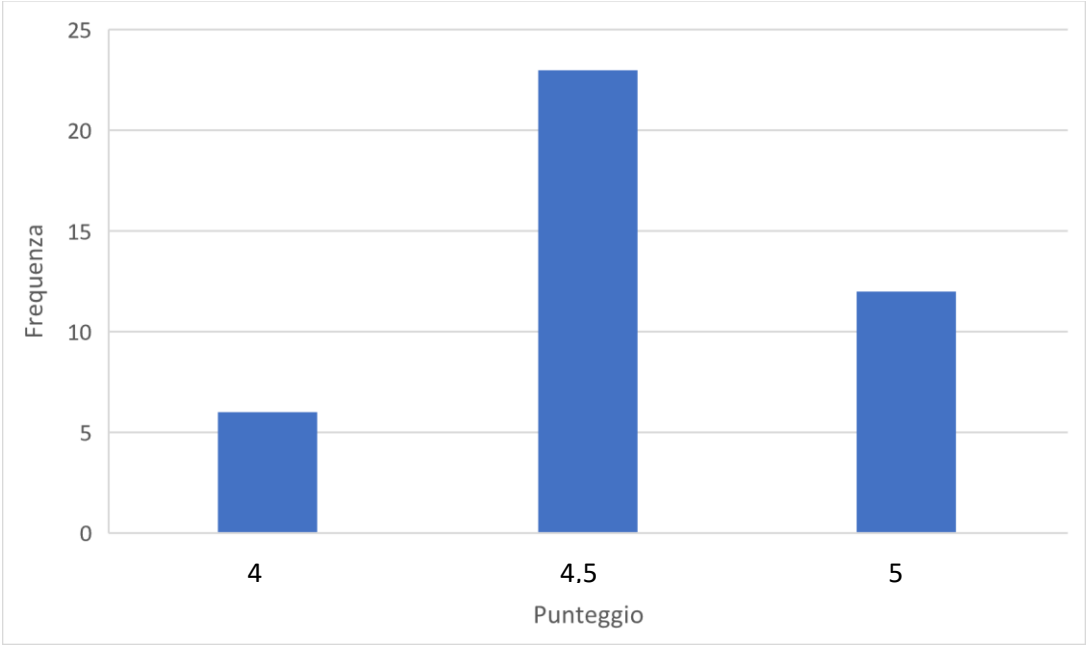
- ID ristorante
- Valutazione (da 1 a 5)
- Testo

In un altro file sono stati raccolti altri dati riguardanti i 41 ristoranti:

- ID
- Nome
- Range prezzo
- Cucina
- Valutazione (tra 1 e 5)

Come possiamo notare dal Grafico 1, i ristoranti considerati hanno valutazioni complessive comprese tra 4 e 5, ovvero buoni se non ottimi, in particolare osserviamo che la maggioranza ha punteggio pari a 4.5 .

Grafico 1 - Numero di ristoranti per valutazione



3.2. Text Mining

Dopo aver raccolto i dati necessari, procediamo con l'analisi utilizzando le principali tecniche di Text Mining. I testi in linguaggio naturale sono dati non strutturati, per questo abbiamo bisogno di trasformarli così da poterli analizzare sia qualitativamente che quantitativamente. La prima fase di questo processo è il cosiddetto Pre-Processing che consiste in fasi di pulizia e strutturazione dei dati.

Per questa fase ho utilizzato due piattaforme di data science e visualizzazione dei dati : Rapidminer e Orange. In particolare in Rapidminer ho utilizzato le estensioni di Rosette e MeaningCloud.

La pulizia dei dati effettuata consisteva in:

- Tokenization, suddivisione del testo in tokens (parole singole) rimuovendo gli spazi e la punteggiatura
- Lower case, trasformare tutte le lettere in minuscolo
- Stopwords, eliminare parole inutili e ridondanti ai fini dell'analisi (come articoli, preposizioni, ecc)
- Filter tokens by lenght, considera le parole di lunghezza compresa tra 2 e 25
- Generare bigrammi e trigrammi

Ho provato anche a inserire lo *stemming* nel pre-processing, ossia la riduzione delle parole alla propria radice, ma non essendo un passaggio fondamentale e non avendo grosse differenze nei risultati finali ho deciso di lasciare le parole nella loro forma originale.

Una volta pulito il testo, sono stati creati vettori di parole tramite la tecnica di Information Retrieval TF-IDF che ha permesso di assegnare ad ogni parola il suo peso per ogni documento (in questo caso per ogni testo di recensione). L'obiettivo è quello di dare più importanza ai termini che compaiono nella recensione, ma che in generale sono poco frequenti. Otteniamo così una matrice di termini-documenti composta da T colonne (una colonna per token), D righe (recensioni) e in ogni cella il peso TF-IDF che esso ha in ogni recensione. Nel mio caso si è creata una matrice 15533x20257.

Il peso TF-IDF viene assegnato nel seguente modo :

$$w_{d,t} = tf\ (d,t) \times idf(d,t)$$

Dove :

- $tf(d,t)$ è la term frequency, ovvero sono le occorrenze del token t nella recensione d
- $idf(d,t) = \log\left(\frac{N}{df(t)}\right)$ è la inverse document frequency, infatti n è il numero di documenti mentre $df(t)$ è il numero di documenti contenenti il token t

Ho proseguito visualizzando graficamente tramite Wordcloud (nuvola di parole), come si osserva nel Grafico 2, i termini più importanti e frequenti.
Mentre nella Figura 3 è possibile visualizzare le 20 parole singole e i 20 bigrammi che sono più ricorrenti in tutte le recensioni prese in considerazione.

Grafico 2 - Wordcloud

Figura 3 - 20 parole e bigrammi più frequenti

Word	Word Count
food	11083
good	7746
pizza	7025
great	6981
service	6895
restaurant	6318
italian	5788
place	5244
staff	5105
really	3820
friendly	3792
london	3694
back	3389
one	3300
time	3199
delicious	3131
nice	2830
excellent	2813
table	2770
best	2706

Word	Word Count
italian food	1259
food good	1049
staff friendly	978
italian restaurant	881
food service	867
great food	821
food great	810
pizza pizza	753
good food	752
food delicious	714
good service	674
definitely back	660
friendly staff	649
great service	648
really good	631
service food	617
restaurant food	565
service friendly	560
pizza good	528
food excellent	526

3.3. Sentiment analysis e topic extraction

La Sentiment Analysis è la parte del Text Mining che cerca di estrarre ed identificare l'opinione da un testo. Lo scopo di questa analisi è misurare il sentiment (positività e negatività di una recensione) e valutare le emozioni espresse in linguaggio naturale. Si basa su un processo computazionale che richiede un lessico specifico per determinare il contesto generale, assegnando ad ogni recensione una polarità.

Per procedere con la sentiment ho utilizzato due programmi:

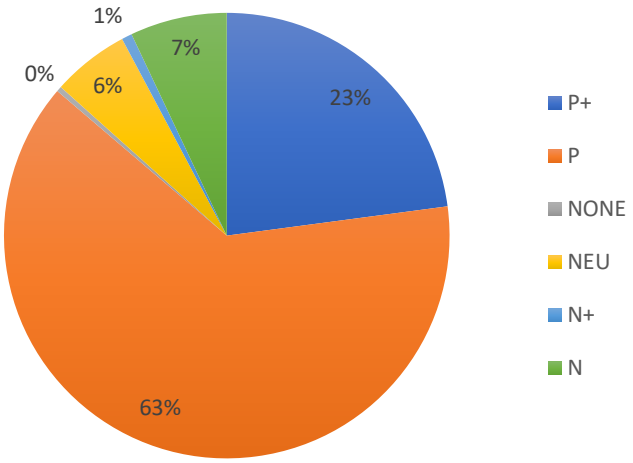
- Orange, in particolare il metodo VADER che calcola la polarità di un testo assegnando valori tra -1 (estremo negativo) e 1 (estremo positivo)

$$compound = \frac{score}{\sqrt{score^2 + \alpha}}$$

- Rapidminer, con estensione MeaningCloud Text che restituisce le seguenti sigle P (positivo), P+ (molto positivo), N (negativo), N+ (molto negativo), NEU (neutro), NONE (non classificato) , per ogni testo.

Nel seguente grafico a torta, Grafico 4, viene rappresentata la distribuzione della classificazione in sentiment delle recensioni. Osservo che la componente positiva, in particolare P, è nettamente maggiore di quella negativa o neutra, e fortunatamente le recensioni non classificate sono in numero estremamente ridotto.

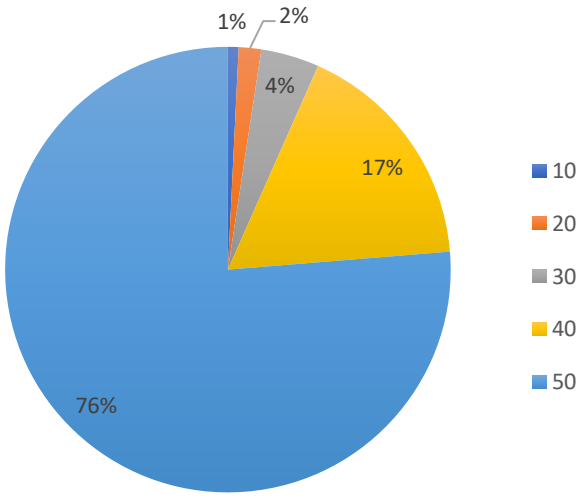
Grafico 4 – Rappresentazione a torta della variabile Sentiment



Possiamo associare ad ogni sigla quello che potrebbe essere il valore corrispondente in termini di valutazione su TripAdvisor :

- P+ → 50
- P → 40
- NEU → 30
- N → 20
- N+ → 10

Grafico 5 – Rappresentazione a torta della variabile Rate



Vediamo nel Grafico 5 una rappresentazione della variabile Rate, ovvero della valutazione che ogni utente assegna al ristorante, e che dovrebbe rispecchiare ciò che ha scritto nel testo della recensione.
Confrontando infatti i due grafici e le percentuali annesse, notiamo che effettivamente le percentuali si assomigliano nonostante nella sentiment prevalga P e nelle valutazioni prevalga il 50.

Una possibile spiegazione potrebbe essere il fatto che gli utenti soddisfatti anche se non in maniera estrema tendano ad assegnare punteggi molto alti (50), mentre la sentiment tende a sottovalutare la positività e assegna solo a casi estremi la sigla P+.

Come mostrato nel Grafico 6, è interessante osservare anche quali sono i termini che più hanno influito nell'assegnazione del valore di sentiment, sia negativamente che positivamente così ho creato un grafico per la visualizzazione dei 20 termini più negativi e 20 termini più positivi.

Grafico 6 – Parole che più influiscono sulla sentiment positiva e negativa

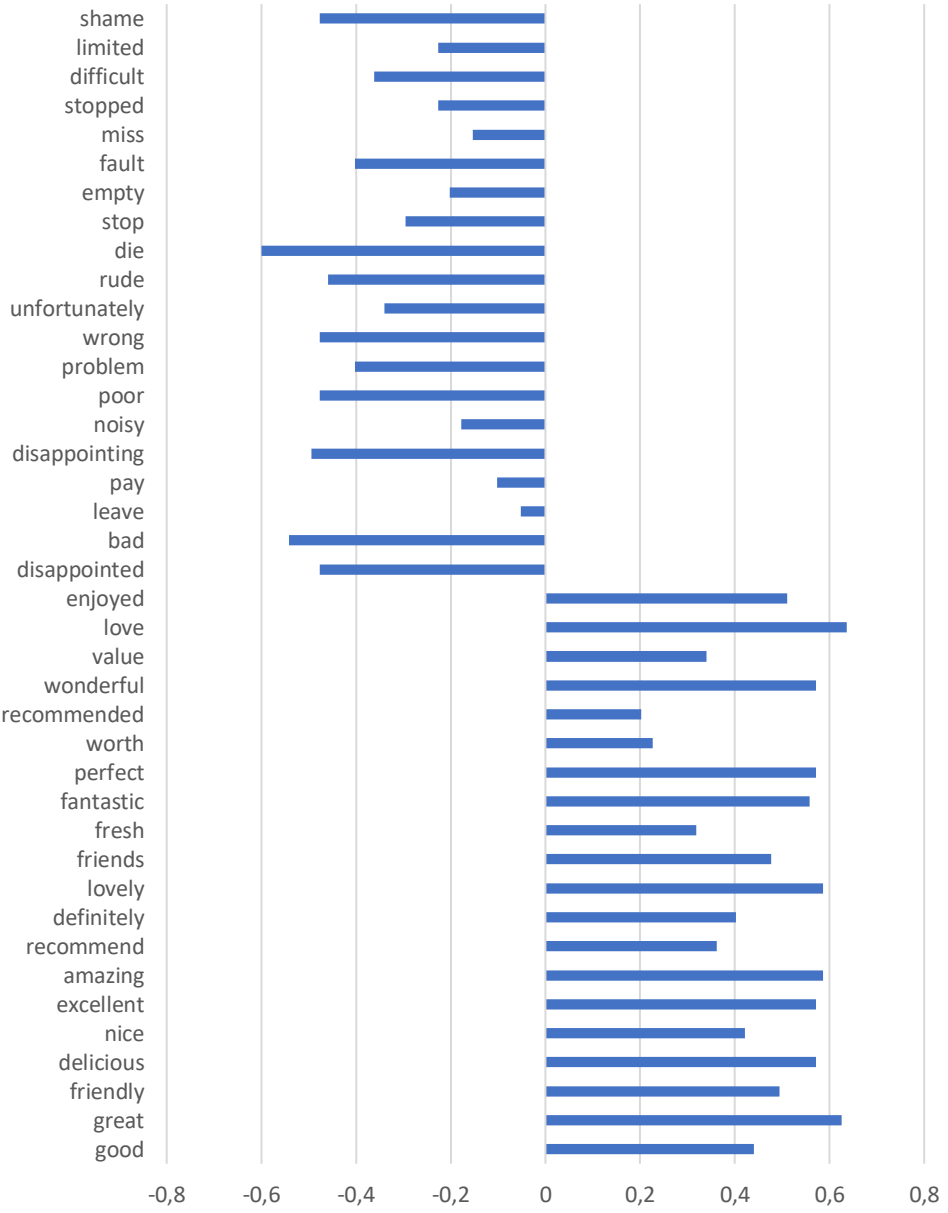
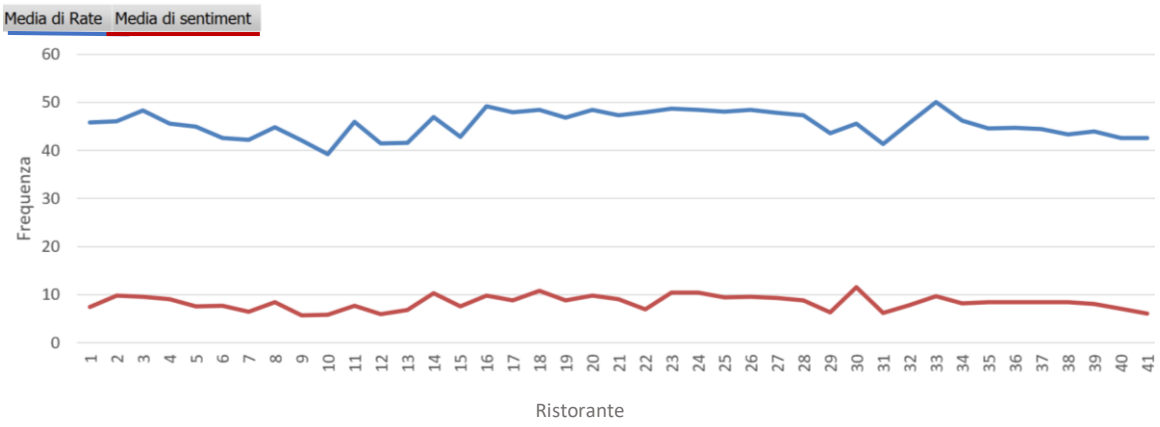


Grafico 7 – Media Rate e Media Sentiment per ristorante



Il Grafico 7 mostra l'andamento per ristorante della media della variabile Rate e della media della variabile Sentiment. Confrontando i due andamenti notiamo che effettivamente c'è un riscontro tra il sentiment del testo e il punteggio assegnato. La correlazione delle medie infatti è di 0,8879. Considerando invece i dati non aggregati di tutti i ristoranti la correlazione risulta essere di 0,5514

Successivamente è stato applicato un algoritmo di Information Extraction per la Topic Extraction, sempre utilizzando l'estensione Meaning Cloud di Rapidminer, essa si occupa di estrarre elementi con alto potere informativo da dati non strutturati, come :

- Concetti
- Entità
- Date
- Citazioni

L'estrazione degli argomenti consente agli utenti di rivedere rapidamente un elenco di frasi chiave e concetti per ottenere il succo di un articolo o di un documento. Si utilizza per capire quali sono le idee più comuni tra i documenti, in questo caso recensioni. La conoscenza delle frasi chiave e dei concetti in ogni documento consente agli utenti di etichettare, ordinare e organizzare automaticamente i propri dati, rendendoli più utili per analizzare e gestire il database.

3.4. Aspect-Based Sentiment Analysis

Sempre grazie a Rapidminer e l'estensione di Meaning Cloud è stato possibile fare un'analisi di sentiment più approfondita. Con la topic extraction si sono trovati i temi e concetti principali e che ricorrevano spesso all'interno delle recensioni.

Così ho pensato a quattro macro argomenti che potessero essere interessanti da valutare separatamente e per ognuno ho assegnato dei concetti chiave che li caratterizzassero.

Gli argomenti presi in considerazione sono :

- Cibo e bevande
- Servizio
- Atmosfera
- Prezzo

Le parole chiave assegnate ad ognuno di questi sono stati :

- Per cibo e bevande: food, pizza, pasta, dinner, beverage, wine, meal
- Per servizio : service, staff
- Per atmosfera: resturant, italian, atmosphere, place
- Per prezzo: price, bill

L'ABSA per ognuna delle parole chiave assegna un valore di sentiment che può essere : positivo (P, P+) , negativo (N,N+) , neutro (NEU), non classificato (NONE). Sommando i punteggi ottenuti ottengo il seguente grafico.

Grafico 8 – Frequenza assoluta dei quattro aspetti

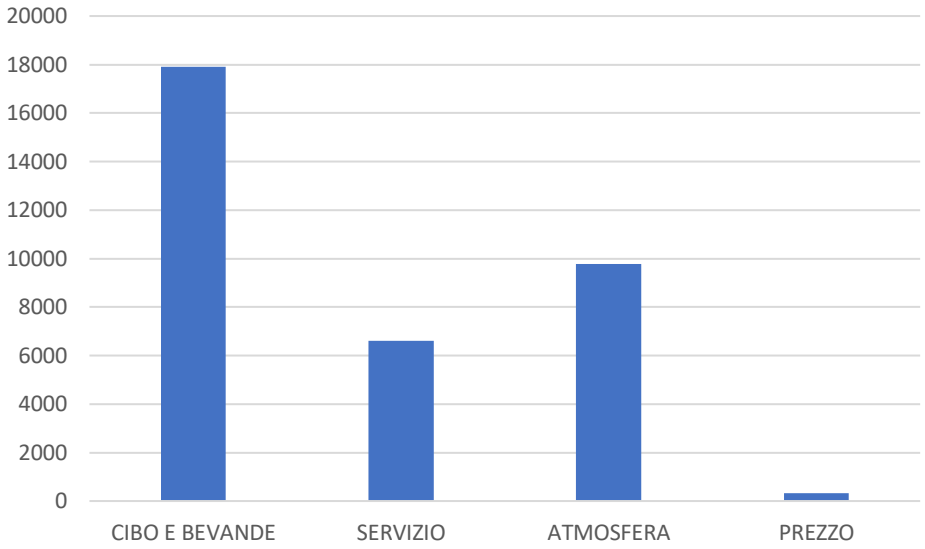
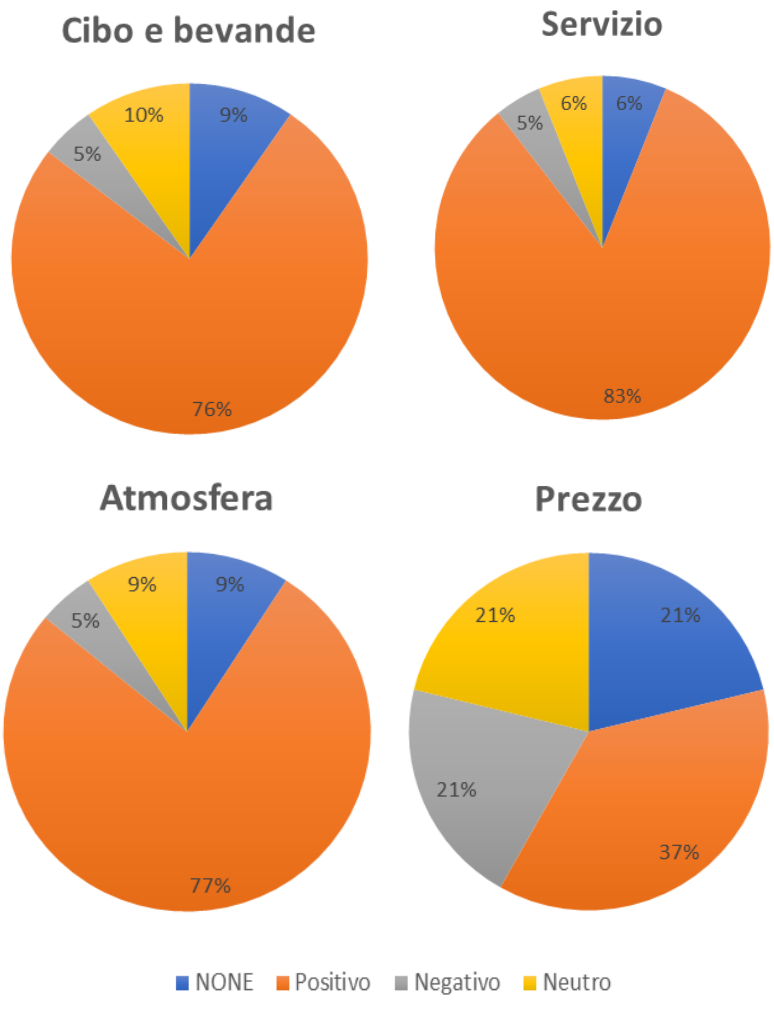


Grafico 9 - Percentuali di sentiment per ogni aspetto



Dai grafici 8 e 9 soprastanti posso dire che l'aspetto principale di cui parlano le recensioni di ristoranti è il cibo, in particolare la qualità e la varietà del menu. Tenendo in considerazione che la maggior parte delle recensioni analizzate sono recensioni positive, di conseguenza anche l'Aspect Based Sentiment Analysis ha in prevalenza risultati positivi, infatti per i primi 3 aspetti le percentuali sono circa quelle che ci si aspettava di avere.

La situazione cambia per quanto riguarda il prezzo, esso ha solo il 37% dei casi positivi, molto meno rispetto agli altri. E' da considerare anche che ci sono molte meno recensioni che trattano l'argomento prezzo rispetto a quelle che trattano gli altri argomenti, e moltissimi non classificati. Questo è un limite non da poco ma comunque ci fa pensare che un consumatore tende a parlare di prezzo raramente e spesso quando la recensione è una critica.

3.5. Similarità e clustering

Grazie al modello spazio-vettoriale, avendo creato la Document-Term Matrix è possibile individuare le recensioni simili tra loro. Per fare ciò è necessario definire una misura di *similarità* (o al contrario *distanza*) tra documenti.

Essa può essere calcolata in diversi modi :

- Distanza euclidea

$$d_e(D_i, D_j) = \sqrt{\sum_{t=1}^{m} (tf_{it} - tf_{jt})^2}$$

- Distanza cosine

$$d_{cos}(D_i, D_j) = \frac{\sum_{t=1}^{m} tf_{it} tf_{jt}}{\sqrt{\sum_{t=1}^{m} tf_{it}{}^2} \sqrt{\sum_{t=1}^{m} tf_{jt}{}^2}}$$

- Coefficiente di correlazione di Pearson

$$d_P(D_i, D_j) = \frac{1 - \rho(D_i, D_j)}{2}$$

Ho scelto di utilizzare la distanza cosine, poiché misura l'angolo tra i vettori rappresentanti i documenti, quindi i documenti nella stessa direzione sono considerati molto vicini. Assume valori tra 0 (similarità minima) e 1 (similarità massima).

Viene creata così una Distance Matrix, essendo però una matrice molto grande è difficile visualizzarla tutta, così ho preso un campione di solo 500 recensioni e ho osservato che le due recensioni che hanno similarità maggiore, pari a 0.35, sono :

Recensione 1
"*If you love pizza, definitely go here. lovely selection of pizza, just that bit different from everything else. Great service, would definitely come back*"

Recensione 2
"*I really love this place. So busy because food is delicious. I've been there many times and always everything is ok. It's not restaurant. It's pizza place . Amazing pizza place. Definitely I will come back."*

Effettivamente notiamo che il testo delle due recensioni è simile, parla della pizza utilizzando aggettivi simili, e in entrambe l'autore dice che sicuramente tornerà a mangiare in quel ristorante.

Sempre utilizzando come misura di similarità quella del coseno, e scelto come linkage l'average linkage procediamo con il clustering delle recensioni.

Il clustering è un metodo non supervisionato nonchè una tecnica di apprendimento che utilizza dati di addestramento in cui non è nota la variabile target. Consiste nell'individuare le osservazioni che condividono caratteristiche simili sulla base di precisi criteri, così da poterle raggruppare.

Può essere gerarchico (crea una decomposizione gerarchica degli oggetti) o partizionale (costruisce una serie di partizioni degli oggetti e ne valuta la qualità). Il clustering consente di accedere più velocemente all'argomento di interesse e di individuarne i legami con altri argomenti.

L'approccio che ho adottato nella mia ricerca è stato quello partizionale, in particolare il metodo k-means, un algoritmo iterativo che suddivide un insieme di oggetti in K gruppi sulla base dei loro attributi, calcolando per ogni iterazione i K centroidi.

L'algoritmo prevede l'assegnazione casuale di un valore da 1 a K ad ogni osservazione, calcola il valore del centroide, ovvero il vettore che contine le medie delle variabili per le osservazioni nel cluster, e assegna ogni osservazione al cluster per il quale il centroide risulta più vicino (in termini di similarità). Si ripetere questo procedimento finchè non si raggiunge la convergenza e ogni osservazione è assegnata al cluster corretto.

Ho scelto K=4 e ho applicato il metodo K-Means tramite il software Rapidminer ottenendo così le recensioni divise in 4 cluster.

```
Cluster 0: 1100 items
Cluster 1: 4918 items
Cluster 2: 6373 items
Cluster 3: 3142 items
Total number of items: 15533
```

Per ogni cluster è interessante visualizzare le tabelle centroidi, in particolare i 14 attributi più centrali per ogni cluster, come si mostra nella Figura 10.

Figura 10 – Tabelle centroidi

Attribute	cluster_0 ↓
amazing	0.082
bunga	0.040
night	0.035
great	0.033
fun	0.029
entertainment	0.029
birthday	0.026
food	0.026
recommend	0.025
definitely	0.024
thank	0.023
staff	0.022
fantastic	0.022
time	0.021

Attribute	cluster_1 ↓
great	0.032
italian	0.032
friendly	0.029
food	0.028
place	0.027
staff	0.026
restaurant	0.026
good	0.026
service	0.024
excellent	0.024
atmosphere	0.024
london	0.024
lovely	0.022
recommend	0.021

Attribute	cluster_2 ↓
dishes	0.032
good	0.024
table	0.024
restaurant	0.023
pasta	0.022
wine	0.020
tapas	0.020
menu	0.019
italian	0.019
service	0.018
food	0.017
ordered	0.017
order	0.016
excellent	0.015

Attribute	cluster_3 ↓
pizza	0.100
pizzas	0.035
good	0.029
great	0.023
place	0.023
nice	0.020
london	0.020
friendly	0.018
service	0.018
staff	0.017
toppings	0.016
delicious	0.015
go	0.015
dough	0.015

Notiamo che i 4 cluster contengono termini nelle tabelle centroidi che si possono ricondurre a 4 macro argomenti :

- Cluster 0 : Esperienza del consumatore
- Cluster 1 : Posto, atmosfera, servizio
- Cluster 2: Cibo, menu
- Cluster 3: Pizza

La maggior parte delle recensioni tratta diversi aspetti, quindi risulta difficile raggruppare le recensioni secondo questo criterio, comunque possiamo dire le recensioni appartenenti a ciascuno dei quattro cluster parlerà in prevalenza dei concetti citati sopra.

Infine per valutare la performance del Clustering ho calcolato l'*average within centroid distance*, cioè la media della distanza tra il centroide e tutti gli esempi di un cluster, che è risultata essere – 0,974.

Conclusioni

Nel presente elaborato si è visto come il web influisca nei viaggi e di conseguenza nel settore turistico. Se i social network continueranno ad espandersi gli argomenti dell'user-generated content relativi al turismo saranno sempre più influenti nella pianificazione di un viaggio. Come ho mostrato questo è uno dei fattori che già adesso più influenza i viaggiatori e la reputation delle strutture.

Al fine di provare a valutare l'opinione dei consumatori per questa ricerca sono state raccolte molte recensioni da TripAdvisor. I metodi di Text Mining, Sentiment Analysis e Clustering applicati hanno permesso di sintetizzare e analizzare la raccolta di recensioni, comprendendo così quali aspetti sono di maggiore interesse per i consumatori e sono maggiormente correlati alla soddisfazione generale.

Dopo aver identificato e quantificato il sentimento dell'unità di testo (overall polarity), la conclusione raggiunta è che i recensori si mostrano nei confronti del servizio decisamente positivi. La maggior parte delle applicazioni di sentiment analysis volge a classificare l'intero documento in positivo o negativo presupponendo che all'interno del testo il soggetto sia uno solo e il sentimento si riferisca solo ad esso. In realtà, analizzando le relazioni tra espressioni di sentimento e soggetti, è possibile fare analisi più approfondite e precise.

La valutazione che poteva semplicemente essere fatta dalla componente umana leggendo e interpretando ciascuna recensione è stata invece svolta in maniera automatizzata dal computer, riconoscendo i sentimenti e i temi principali

trattati. Per grandi quantità di recensioni risulta quindi assai utile, se non indispensabile, usufruire di queste tecniche per poter estrarre informazioni valide.

Concludo affermando che il progetto presentato è un ulteriore conferma dell'importanza che l'opinione della gente ha nell'influenzare le decisioni. Le strutture stesse dovrebbero riconoscere ciò e intensificare l'utilizzo di tecnologie per processi come il Text Mining al fine di migliorare le proprie strategie di business e la propria reputation.

Bibliografia

Afzaal M., Fong A., Usman M., (2019). *Predictive aspect-based sentiment classification of online tourist reviews*, Journal of Information Science 2019, Vol. 45(3) 341–363

Anderson CK., (2012). *The impact of social media on lodging performance*, Cornell Hospitality Report.

Anjos F., Anjos S., Limberger P., Meira J., (2014). *Satisfaction in hospitality on TripAdvisor.com: An analysis of the correlation between evaluation criteria and overall satisfaction,* Tourism & Management Studies.

Baggio R., Costa C., Miguéns J., (2008). *Social media and Tourism Destinations: TripAdvisor Case Study.*

Chen Z., Fei G., Liu B., Mukherjee A., (2016). *Discovering Correspondence of Sentiment Words and Aspects,* Computational Linguistics and Intelligent Text Processing.

Gan C. H.K., Pezenka I., C. Weismayer, (2018). *Aspect-Based Sentiment Detection: Comparing Human Versus Automated Classifications of TripAdvisor Reviews*, Information and Communication Technologies in Tourism.

Geng X., Liao J., Yang L., (2016). *A web sentiment analysis method on fuzzy clustering for mobile social media users*, Journal on Wireless Communications and Networking.

Khoo C. S.G., Na J., Thet T.T., (2010). *Aspect-based sentiment analysis of movie reviews on discussion boards*, Journal of Information Science.

Nizamuddin SK. MD., (2015). *Marketing utility of TripAdvisor for Hotels: An importance-performance analysis*, Journal of Tourism.

Ravi K., Ravi V., (2015). *A survey on opinion mining and sentiment analysis: Tasks, approaches and applications*, Knowledge-Based Systems 89, 14–46

Sitografia

https://www.tripadvisor.com

https://jupyter.org/

https://www.wikipedia.org/

Ringraziamenti

Infine, vorrei dedicare qualche riga a tutti coloro che mi sono stati vicini in questo percorso di crescita personale e professionale.

Un sentito grazie al mio relatore Roberto Boselli per la sua disponibilità e tempestività ad ogni mia richiesta. Grazie per avermi fornito ogni materiale utile alla stesura dell'elaborato.

Senza il supporto morale della mia famiglia, soprattutto quello dei miei genitori, non sarei mai potuta arrivare fin qui. Grazie per esserci sempre stati soprattutto nei momenti di sconforto.

Ringrazio il mio fidanzato Giovanni per avermi sempre incoraggiata. Grazie per tutto il tempo che mi hai dedicato. Grazie perché ci sei sempre stato.

Ringrazio i miei amici e i miei compagni di università per essermi stati accanto in questo periodo intenso e per gioire, insieme a me, dei traguardi raggiunti.

Grazie a tutti, senza di voi non ce l'avrei mai fatta.