

Matteo Gurrieri

STUDENTE PRESSO LUMSA

```
8 .box{  
9   position: absolute;  
10  top: 50%;  
11  left: 50%;  
12  transform: translate(-50%, -50%);  
13  width: 400px;  
14  padding: 40px;  
15  background: #fff;  
16  box-sizing: border-box;  
17  box-shadow: 0 15px 25px #ccc;  
18  border-radius: 10px;  
19 }  
20 .box h2{  
21   margin: 0 0 30px;  
22   padding: 0;}
```

Report

Introduzione

Il progetto di Text Mining è stato incentrato sull'analisi delle recensioni in inglese dello Scholars Lounge Irish Pub di Roma presenti su TripAdvisor. Situato nel cuore di Roma, accanto all'imponente Piazza Venezia, il pub gode di una reputazione prestigiosa non solo nella capitale italiana, ma in tutta l'Italia stessa. Fondato nel settembre 2005 da Declan Crean e Celestino Cucchiarelli, l'attività è diventata un punto di riferimento per la vita notturna della Città Eterna.

Dataset

Per la realizzazione del progetto, è stato utilizzato il software Data Miner. Tuttavia, durante il processo di estrazione, è emerso un limite tecnico. Il software non è stato in grado di estrarre l'intera recensione, ma ha ottenuto solo il contenuto fino alla parola "More". Per visualizzare l'intera recensione, sarebbe stato necessario cliccare sul termine "More" nell'apposita sezione.

Il dataset è composto ufficialmente da 1725 righe e 7 colonne:

- **Username:** il nome dell'utente che ha scritto la recensione.
- **Review Number:** il numero della recensione scritta dall'utente.
- **Title:** il titolo della recensione scritta dall'utente.
- **Text:** il testo della recensione scritta dall'utente.
- **Date:** la data in cui è stata scritta la recensione.
- **Rating:** la valutazione numerica assegnata dall'utente.
- **Id:** l'identificatore univoco assegnato a ciascuna recensione nel dataset.

Operazioni sul Dataset

Per ottenere il dataset descritto in precedenza, sono state apportate modifiche al dataset originale sia in Python che in R. Di seguito vengono riportati i dettagli delle modifiche effettuate:

1. **Creazione della colonna 'combined_rating':** È stata creata una nuova colonna che combina i valori delle colonne 'Rating', 'Rating.1', 'Rating.2', 'Rating.3', 'Rating.4'. Successivamente, queste ultime sono state rimosse dal dataset.
2. **Rinominazione delle colonne:** Le colonne 'i..Username' e 'combined_rating' sono state rinominate rispettivamente in 'Username' e 'Rating'.
3. **Pulizia della colonna 'Date':** È stata rimossa la parte di testo "Date of visit: " dalla colonna 'Date' nel DataFrame. Questa operazione ha consentito di ottenere solo la data relativa alla visita.
4. **Pulizia della colonna 'Review Number':** Sono state rimosse le stringhe " reviews" e " review" dalla colonna 'Review Number'. Questa operazione ha consentito di ottenere solo il numero di recensioni senza alcun testo aggiuntivo.
5. **Creazione del nuovo DataFrame 'df':** È stato creato un nuovo DataFrame che contiene gli stessi dati del DataFrame originale 'ScholarsEn'.
6. **Elaborazione della colonna 'Date':** Sono stati estratti gli ultimi 4 caratteri dalla colonna 'Date' nel DataFrame 'df'. Questa operazione ha consentito di ottenere l'anno delle visite al pub.
7. **Aggiunta della colonna 'Id':** È stata aggiunta una colonna che contiene un numero incrementale da 1 alla lunghezza del DataFrame 'df'.
8. **Conversione delle colonne 'Text' e 'Title':** Le colonne 'Text' e 'Title' nel DataFrame 'df' sono state convertite in tipo stringa per garantire una corretta gestione del testo.

Premessa

La differenza principale tra le analisi effettuate in R e Python risiede nell'approccio adottato per l'elaborazione dei dati. In Python, le analisi sono state condotte sui titoli e sul testo delle recensioni combinati, mentre in R sono state svolte separatamente. Questa scelta è stata determinata dall'obiettivo di cercare nuove informazioni e non solo replicare gli stessi passaggi in entrambi i linguaggi di programmazione.

Analisi in Python

Tokenizzazione, Stopwords e POS Tagging

Prima di procedere con l'analisi effettiva, è stata eseguita la tokenizzazione e la rimozione delle stopwords. Inizialmente, sulla colonna 'review' del DataFrame 'df', sono stati rimossi tutti i caratteri non alfabetici utilizzando la libreria regex. Successivamente, tutti i valori sono stati convertiti in minuscolo per uniformare il testo.

Poi è stato creato un nuovo DataFrame ridotto chiamato 'IRDR' che include solo le colonne 'Id', 'Rating', 'Date' e 'Review'. Su questo DataFrame è stata effettuata la tokenizzazione utilizzando la funzione 'word_tokenize' e la rimozione delle stopwords in inglese. Le stopwords erano state precedentemente scaricate dalla libreria nltk. Questa operazione ha consentito di ottenere una lista di parole significative, escludendo le parole comuni che non apportano informazioni rilevanti.

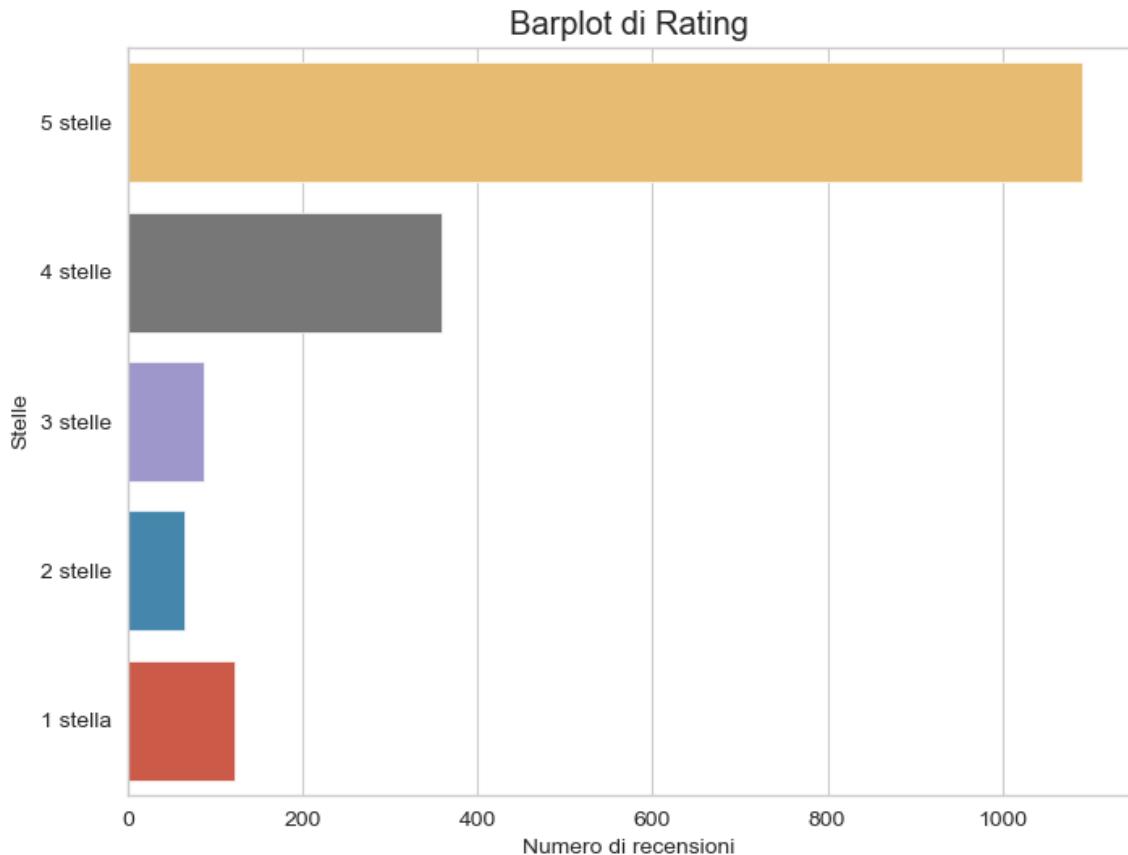
All'interno del notebook, è stata effettuata nuovamente la tokenizzazione su una recensione casuale dal dataset. Inoltre, è stato applicato il POS tagger ai token utilizzando la funzione 'pos_tag' della libreria nltk. Questo approccio ha fornito una visione più chiara del funzionamento di queste funzioni e ha permesso di identificare la categoria grammaticale di ciascuna parola nel testo.

Per effettuare la lemmatizzazione è stato utilizzato anche il modulo 'nlp' della libreria spacy all'interno della funzione 'lematized_review'. Questo metodo ha consentito una diversa strategia di rimozione delle stopwords basata su modelli di linguaggio più avanzati.

Infine, il risultato di tutte queste operazioni è stato assegnato alla variabile 'clean_df_review', che contiene le recensioni pulite e pronte per l'applicazione delle tecniche di Text Mining.

Analisi Colonna Rating

Per analizzare la colonna 'Rating', abbiamo iniziato trovando la frequenza dei diversi punteggi stellari in base al numero delle recensioni presenti nel dataset. Abbiamo rappresentato graficamente questi dati utilizzando un barplot.

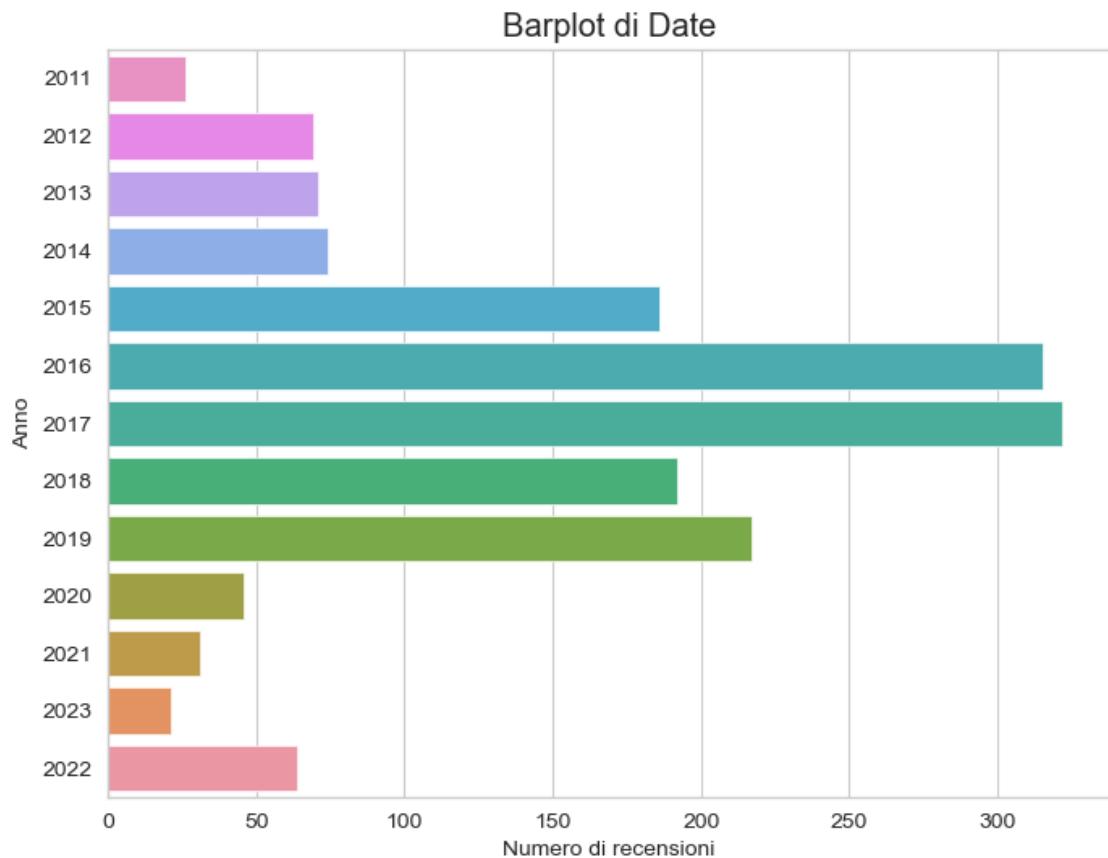


È emerso immediatamente che su un totale di 1725 recensioni presenti nel dataset, la maggior parte di esse ha assegnato 5 stelle al locale. Questo risultato non è sorprendente, considerando che lo Scholars Lounge Irish Pub, come descritto nel paragrafo introduttivo, è uno dei locali più rinomati di Roma.

Questa prima analisi ci ha fornito un'indicazione preliminare: esplorando più approfonditamente le recensioni, ci aspettiamo di trovare un maggior numero di parole positive rispetto a quelle negative nei confronti del locale.

Analisi Colonna Date

Per quanto riguarda la colonna 'Date', abbiamo effettuato un'analisi simile calcolando la frequenza degli anni in base al numero delle recensioni presenti nel dataset. Questo ci ha permesso di visualizzare facilmente la distribuzione delle recensioni nel corso degli anni.



Dal grafico, è evidente che la maggior parte delle recensioni è stata effettuata durante gli anni 2016-2017. Inoltre, si nota un drastico calo delle recensioni a partire dal 2020. Questo calo potrebbe essere attribuito all'inizio della pandemia di COVID-19.

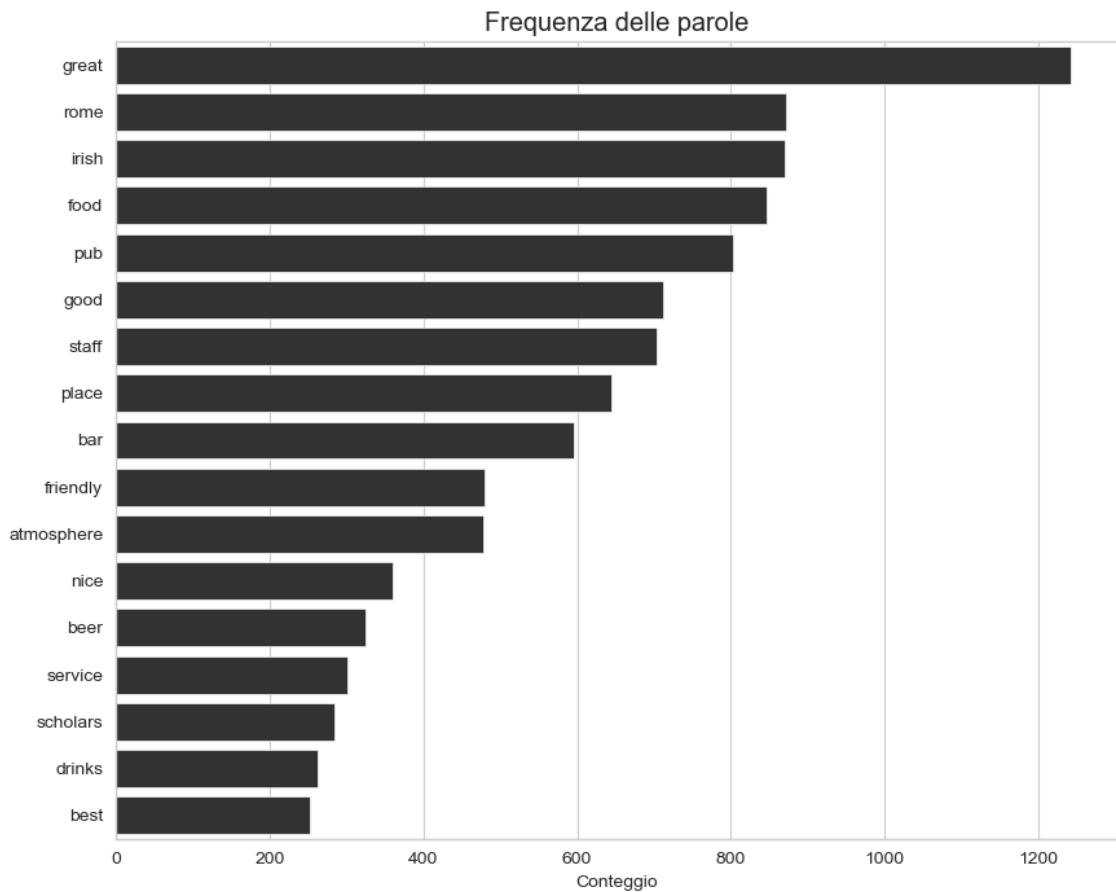
Quest'analisi ci fornisce una panoramica temporale delle recensioni e ci aiuta a comprendere meglio l'andamento nel tempo dell'attività di recensione degli utenti per

il locale.

Frequenza delle Parole

Successivamente, abbiamo creato un grafico a barre per analizzare le parole più frequenti nelle recensioni. In particolare, ci siamo focalizzati sulle parole che hanno una frequenza superiore a 250.

Questo ci permette di identificare facilmente le parole più comuni nelle recensioni, fornendoci una panoramica delle parole che vengono menzionate più frequentemente dagli utenti.



Tra le parole più utilizzate nelle recensioni, possiamo notare che alcune si distinguono per la loro frequenza. In cima alla lista troviamo parole come "great", "rome", "irish", "food" e "pub". Queste parole chiave indicano elementi che i recensori ritengono rilevanti nell'esperienza allo Scholars Lounge Irish Pub di Roma.

La presenza frequente della parola "great" suggerisce che gli utenti hanno una percezione positiva del locale, forse indicando un servizio eccellente, un'atmosfera accogliente o una vasta selezione di bevande e cibi. "Rome" è un termine che probabilmente indica l'ubicazione del pub, evidenziando l'attrattiva turistica della

città e l'opportunità per i visitatori di godersi una serata in un ambiente internazionale.

Le parole "irish" e "pub" riflettono la natura stessa dell'esercizio, sottolineando l'identità culturale del locale come autentico pub irlandese. È probabile che gli utenti abbiano apprezzato l'atmosfera unica, la musica tradizionale e la possibilità di gustare bevande tipiche dell'Irlanda.

Infine, la menzione della parola "food" suggerisce che la componente culinaria del pub è stata apprezzata dagli utenti. Questo potrebbe indicare la presenza di piatti tradizionali irlandesi o di un menu variegato che soddisfa le preferenze dei clienti.

Continuando a scorrere tra le parole più utilizzate, iniziamo a notare che un gran numero sono aggettivi positivi come: "good", "friendly", "nice" e "best". Come detto in precedenza, essendo per lo più recensioni con 5 stelle, è normale che tra le parole con una frequenza alta siano presenti parole positive riguardo al locale.

Successivamente sono state effettuate ulteriori ricerche concentrandosi sulla relazione tra le parole e l'anno di pubblicazione della recensione, così come la relazione tra le parole e il punteggio assegnato.

	Rating	word	count		Date	word	count
0	5	great	968	0	2023	food	19
1	5	rome	639	1	2023	staff	15
2	5	irish	571	2	2023	bar	14
3	5	food	558	3	2023	place	12
4	5	pub	540	4	2023	rome	12
...
8481	1	year	1	12771	2011	win	1
8482	1	years	1	12772	2011	wings	1
8483	1	yes	1	12773	2011	wingsmore	1
8484	1	yet	1	12774	2011	withmore	1
8485	1	youll	1	12775	2011	womens	1

Nel primo caso, si è cercato di valutare se ci fossero differenze nelle parole utilizzate a seconda del punteggio assegnato alle recensioni. Nel secondo caso, l'obiettivo era individuare eventuali tendenze o variazioni nelle parole utilizzate nelle recensioni nel corso degli anni. Tuttavia, anche in questo caso, non sono state riscontrate evidenze rilevanti o significative. Le parole chiave individuate erano simili a quelle precedentemente identificate, senza variazioni sostanziali in base agli anni presi in considerazione e al punteggio assegnato.

Questi risultati suggeriscono che le parole più frequenti e significative nelle recensioni rimangono coerenti indipendentemente dall'anno di pubblicazione o dal punteggio

assegnato. Le parole chiave individuate inizialmente continuano a essere rilevanti in tutto il dataset, sottolineando ancora una volta la presenza di elementi positivi e apprezzati nelle esperienze dei clienti presso lo Scholars Lounge Irish Pub di Roma.

Word Cloud

Successivamente, abbiamo scelto di visualizzare le 100 parole più frequenti all'interno delle recensioni utilizzando un WordCloud a forma di birra, grazie all'utilizzo della libreria wordcloud. In questo tipo di visualizzazione, la dimensione delle parole è direttamente proporzionale alla loro frequenza all'interno del testo.



Esaminando il WordCloud, possiamo notare che le parole più grandi corrispondono alle stesse parole menzionate in precedenza. Ancora una volta, queste parole chiave sono prevalentemente positive. Tra queste, emergono combinazioni come "great atmosphere", "great pub" e "great food", che evidenziano ulteriormente la presenza di un'atmosfera eccezionale, un pub straordinario e un cibo di qualità.

Il WordCloud ci offre un'ulteriore conferma che le recensioni riguardanti lo Scholars Lounge Irish Pub di Roma sono principalmente positive. Le parole più frequenti sono strettamente legate a esperienze piacevoli, servizio di qualità, atmosfera accogliente e cibo eccellente.

Questa rappresentazione visiva ci permette di cogliere immediatamente i concetti principali che emergono dalle recensioni, evidenziando l'entusiasmo e

I'apprezzamento degli utenti per l'esperienza nel locale.

L'analisi delle parole più utilizzate nelle recensioni ci fornisce una panoramica iniziale dell'esperienza complessiva dei clienti al Scholars Lounge Irish Pub di Roma. Tuttavia, per ottenere una visione più completa, è stato necessario approfondire ulteriormente l'analisi dei testi, ad esempio mediante l'analisi dei sentimenti, l'identificazione di temi chiave o l'applicazione di algoritmi di classificazione.

Sentiment Analysis

La sentiment analysis è stata effettuata prima utilizzando il modello VADER (Valence Aware Dictionary for Sentiment Reasoning), che tiene conto sia della polarità (positiva/negativa) che dell'intensità delle emozioni. Oltre a fornire informazioni sui punteggi di positività e negatività, VADER fornisce anche informazioni sul sentimento di una frase. Vader è molto intelligente nel riconoscere frasi positive e negative in base alle parole presenti nella frase.

```
# Calcolo del sentiment polarity score per la frase 'I am so happy!'
sia.polarity_scores('I am so happy!')
```

```
{'neg': 0.0, 'neu': 0.318, 'pos': 0.682, 'compound': 0.6468}
```

```
# Calcolo del sentiment polarity score per la frase 'This is the worst thing ever'
sia.polarity_scores('This is the worst thing ever')
```

```
{'neg': 0.451, 'neu': 0.549, 'pos': 0.0, 'compound': -0.6249}
```

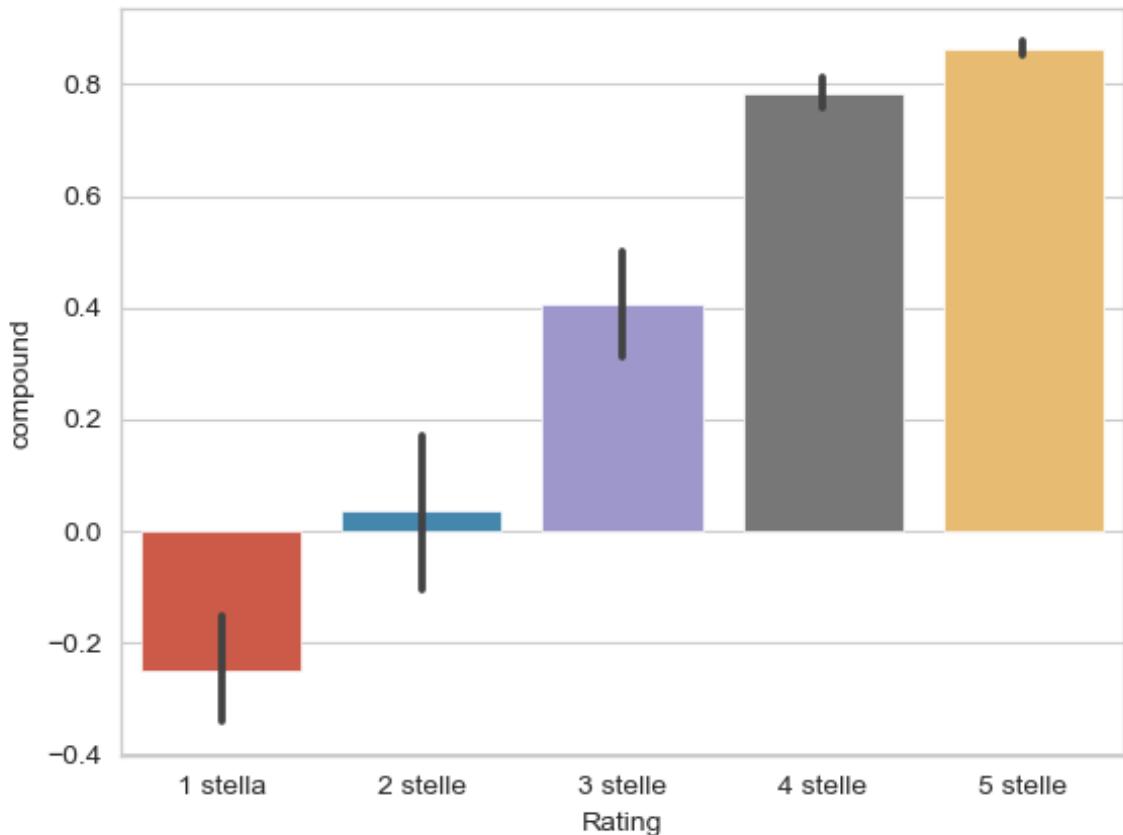
Il punteggio complessivo (Compound score) in Vader è una misura che somma tutti i punteggi lessicali che sono stati scalati tra -1 (estremamente negativo) e +1 (estremamente positivo). La funzione SentimentIntensityAnalyzer() di VADER analizza una stringa e produce un dizionario di punteggi suddivisi in quattro categorie: Negativo, Neutrale, Positivo e Compound, ottenuti normalizzando i restanti tre punteggi.

```
# Utilizzo della funzione .head() per visualizzare le prime 5 righe del nuovo DataFrame vaders
vaders.head()
```

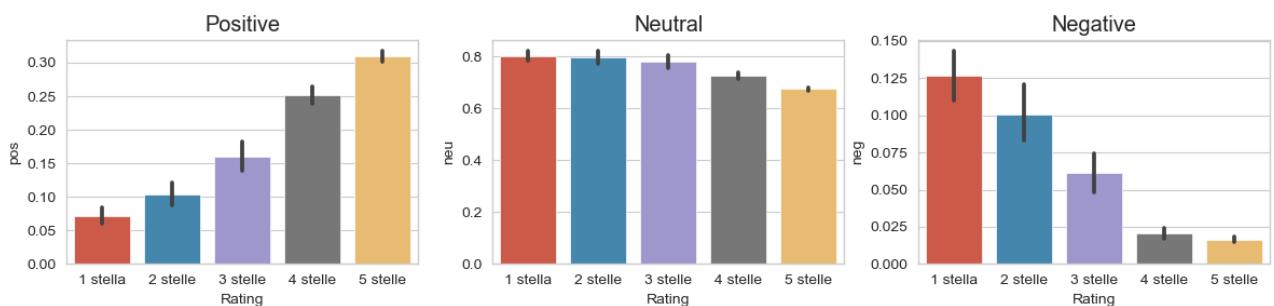
	Id	neg	neu	pos	compound	Username	Review Number	Title	Text	Date	Rating	review	review_stop
0	1	0.000	0.660	0.340	0.9428	Mon-coach i	37	We had so much fun	"We had so much fun. It was my first time goin...	2022	5.0	we had so much fun we had so much fun it was m...	[time, go, club, drink, perfect, music, karao...
1	2	0.094	0.840	0.067	-0.3197	GoldstarDancer	31	Shocking Overcrowding	On our trip to Rome we spent about 1/2 hrs in ...	2023	2.0	shocking overcrowding on our trip to rome we s...	[shock, overcrowding, trip, rome, spend, schol...
2	3	0.000	0.589	0.411	0.9810	theokietravelbird	47	Best Pub!	We went there for St. Patrick's Day and had a ...	2023	5.0	best pub we went there for st patricks day and...	[good, go, patrick, wonderful, time, music, am...
3	4	0.138	0.782	0.080	-0.8317	AndreaE2327	1	Very rude security staff	They let enter first just the girls they like...	2023	1.0	very rude security staff they let enter first ...	[rude, security, staff, enter, girl, like, fri...
4	5	0.139	0.763	0.098	-0.2505	R7169YHdarrenm	1	Overpriced and terrible food	For the price they charge the food is terrible...	2023	1.0	overpriced and terrible food for the price the...	[overpriced, terrible, food, price, charge, fo...

Abbiamo calcolato i punteggi di polarità per tutte le recensioni presenti nel nostro dataframe, ottenendo valori per la negatività, la neutralità, la positività e il punteggio complessivo di ciascuna recensione. Successivamente, abbiamo combinato questi dati con il dataframe originale, creando un nuovo dataframe chiamato "Vader" che contiene i punteggi di polarità per le recensioni.

Compound su Rating delle recensioni



Il grafico mostra il punteggio complessivo ottenuto da Vader in base alle stelle assegnate alle recensioni. Come ci si poteva aspettare, i punteggi più alti sono associati alle recensioni con 5 stelle, mentre le recensioni con 2 stelle risultano abbastanza neutre e quelle con 1 stella presentano un punteggio negativo. Questo risultato conferma che Vader è in grado di identificare correttamente il sentimento delle recensioni. È naturale che le recensioni con punteggio più elevato abbiano un punteggio migliore, poiché sono caratterizzate da un'esperienza positiva, mentre le recensioni con un punteggio più basso hanno un punteggio peggiore, indicando un'esperienza negativa.



Il secondo grafico mostra i punteggi positivi, neutri e negativi in base alle stelle assegnate alle recensioni. Nel primo subplot è evidente come i punteggi positivi siano concentrati principalmente nelle recensioni da 5 stelle, mentre diminuiscono man mano che il punteggio diminuisce fino ad arrivare alle recensioni da 1 stella. Al

contrario, nel terzo subplot, i punteggi negativi sono più elevati nelle recensioni da 1 stella e diminuiscono nelle recensioni da 5 stelle. Per quanto riguarda i punteggi neutrali, i valori sembrano essere abbastanza costanti per tutte le stelle, anche se le recensioni da 1 stella presentano un valore neutrale più elevato rispetto alle recensioni da 5 stelle.

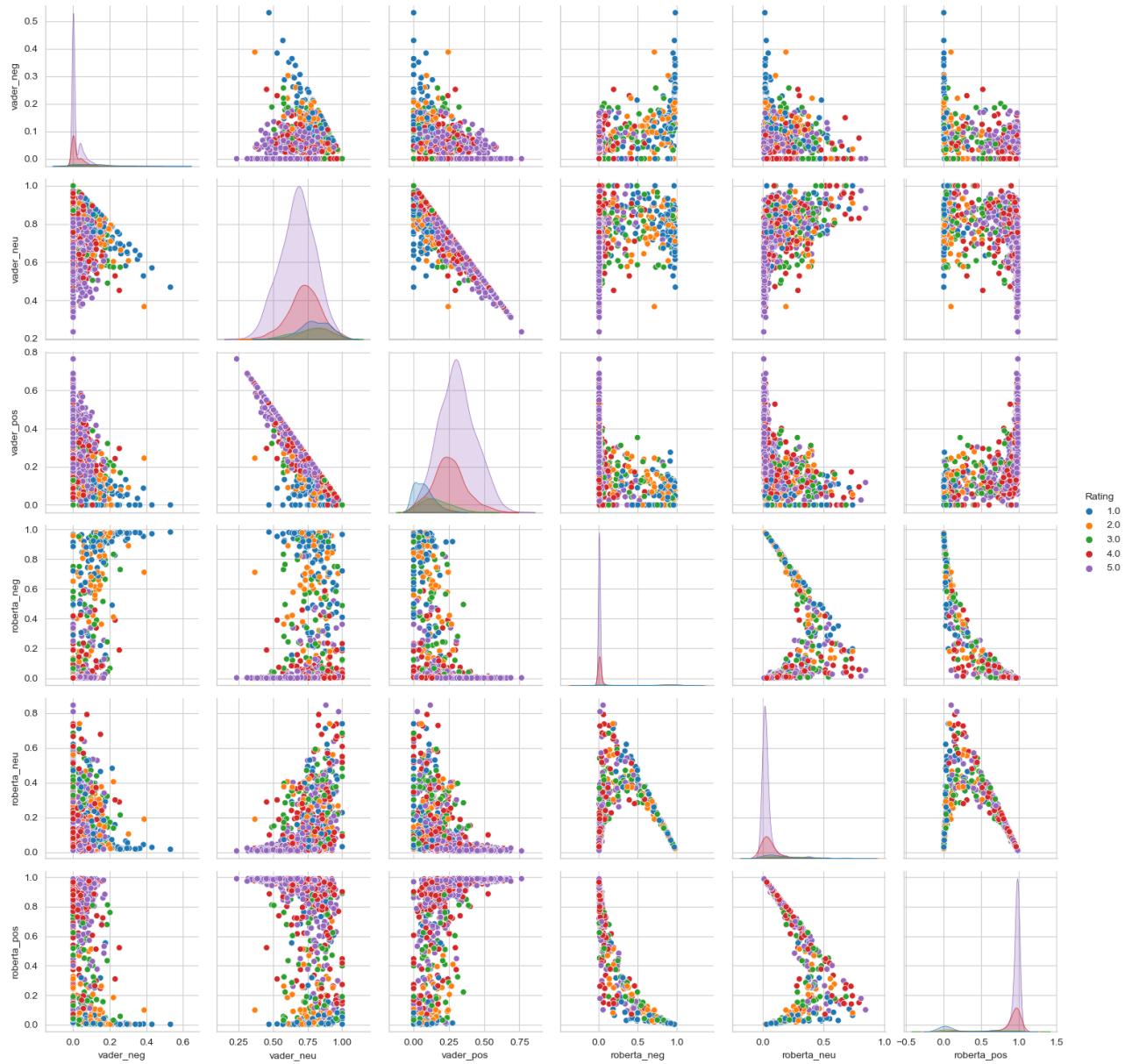
Successivamente, abbiamo utilizzato il modello RoBERTa per condurre l'analisi. RoBERTa è un modello di trasformatori che è stato pre-addestrato su un vasto corpus di testi in lingua inglese utilizzando un metodo di auto-supervisione. RoBERTa analizza una stringa di testo e produce un dizionario di punteggi nelle tre categorie: Negativo, Neutro e Positivo.

```
disappointed the place has a wonderful irish pub feel and can imagine when it is full there would be some atmosphere we popped in for a drink and food at the end of a long day of sightseeing  
we had a pint of lager and a double gin and tonic the pint price is the going rate for rome but the gin was a little overpriced  
gin was what you would expect the lager on the other hand was foul it tasted and smelt like sweaty feet  
we decided not to try the food and left with the lager undrunkmore  
{'neg': 0.02, 'neu': 0.883, 'pos': 0.097, 'compound': 0.8338}  
{'roberta_neg': 0.40062717, 'roberta_neu': 0.3998137, 'roberta_pos': 0.19955912}
```

Per valutare le differenze tra i risultati di VADER e RoBERTa, abbiamo eseguito un test su un esempio. I punteggi ottenuti dai due modelli sono risultati significativamente diversi. Mentre VADER ha classificato l'esempio come sentiment neutro, RoBERTa ha attribuito un punteggio più elevato al sentiment negativo.

Questo esempio mette in evidenza come i diversi modelli di analisi del sentiment possano produrre valutazioni discordanti per lo stesso testo. Mentre VADER si basa su un lessico predefinito e considera l'intensità delle parole nel determinare il sentiment, RoBERTa utilizza un approccio più complesso basato su modelli di trasformatori e l'apprendimento supervisionato. La differenza nei punteggi ottenuti può essere attribuita alle diverse metodologie e alla complessità dei modelli.

A questo punto, abbiamo deciso di confrontare i punteggi di VADER e RoBERTa su tutte le recensioni. Se osserviamo i punteggi di VADER, le recensioni positive sono più concentrate verso destra per le recensioni a 5 stelle. Per il modello RoBERTa, invece, i punteggi si trovano molto più a destra. È evidente che il modello VADER è leggermente meno sicuro nelle sue previsioni rispetto al modello RoBERTa, che separa in modo netto i punteggi di positività, neutralità e negatività per ciascuno di questi valori predetti. Naturalmente, le recensioni a 5 stelle presentano un punteggio di positività elevato, mentre quelle a 1 stella hanno un punteggio peggiore.



È interessante notare come l'utilizzo di diversi modelli possa portare a variazioni nei punteggi e nelle classificazioni delle recensioni. Questo sottolinea l'importanza di considerare e confrontare più approcci nell'analisi del sentiment, al fine di ottenere una visione più completa e accurata delle opinioni espresse nel testo.

Nonostante la maggior parte dei testi sia classificata correttamente, ci saranno comunque alcune frasi ambigue o erroneamente classificate. A volte possono sembrare positive, ma in realtà sono negative. Allo stesso modo, le frasi positive a volte possono sembrare negative. Ora vedremo alcuni testi che i nostri modelli classificano erroneamente.

```
# Selezionare la recensione con la valutazione 'Rating' pari a 1, ordinata per il punteggio 'roberta_pos' in ordine decrescente e selezionare il valore nella results_df.query('Rating == 1').sort_values('roberta_pos', ascending = False)['review'].values[0]
```

'birthday party it was one of my favourite places to come before i did my birthday party there \ni loved the place before because it is the only pub where i feel the international atmosphere and that is why i decided to make my birthday party theremore'

```
# Selezionare la recensione con la valutazione 'Rating' pari a 1, ordinata per il punteggio 'vader_pos' in ordine decrescente e selezionare il valore nella results_df.query('Rating == 1').sort_values('vader_pos', ascending = False)['review'].values[0]
```

'should be called linekers sports bar in italian we have been here two evenings running both evenings started well with the friendly familiarity of a great irish bar oak brass whiskey guinness and typical best friend irish waiter there are numerous tvs throughout the bar showing champions league and europa league football allmore'

```
# Selezionare la recensione con la valutazione 'Rating' pari a 5, ordinata per il punteggio 'roberta_neg' in ordine decrescente e selezionare il valore nella results_df.query('Rating == 5').sort_values('roberta_neg', ascending = False)['review'].values[0]
```

'a door to ireland in the middle of rome i entered this pub with prejudices because theres millions of place like this all over italy but its hard they catch the true irish spirit well i couldnt be more wrong\basically all the personal comes from ireland or at least have relatives you canmore'

```
# Selezionare la recensione con la valutazione 'Rating' pari a 5, ordinata per il punteggio 'vader_neg' in ordine decrescente e selezionare il valore nella results_df.query('Rating == 5').sort_values('vader_neg', ascending = False)['review'].values[0]
```

'an oasis in rome we loved the scholars we were suffering from crowd overload and fairly desperate for a cold pint and a bite to eat then we stumbled upon this lovely atmospheric friendly irish bar and knew wed fallen on our feet it was a real havenmore'

Nel primo esempio, abbiamo preso in considerazione un testo che è stato classificato come positivo da RoBERTa, ma la valutazione del cliente è di 1 stella. Inizialmente, questo potrebbe sembrare contraddittorio, in quanto il testo contiene parole positive come "amore". Tuttavia, analizzandolo più attentamente, diventa evidente che il testo è in realtà negativo, in quanto il cliente si lamenta della sua esperienza presso il locale durante la sua festa di compleanno. Pertanto, RoBERTa ha classificato erroneamente il testo.

Anche il terzo e il quarto esempio sono chiari esempi di errata classificazione sia da parte di RoBERTa che di VADER. Questa misclassificazione è accettabile in questo caso, poiché non è stato possibile considerare l'intera recensione utilizzando il software Data Miner.

Topic Modeling

La modellazione degli argomenti è un metodo utilizzato per classificare i temi presenti in un documento. Nel nostro caso, abbiamo seguito i seguenti passaggi:

- 1. Abbiamo creato un dizionario a partire dalle recensioni pulite utilizzando la libreria gensim e la classe Dictionary. Questo dizionario rappresenta un mapping tra parole uniche e identificatori numerici.**
- 2. Successivamente, abbiamo creato un corpus utilizzando il modello bag-of-words (BOW) per ogni recensione nel dataset pulito. Il corpus è una rappresentazione numerica delle recensioni, in cui ogni parola è rappresentata dal suo identificatore nel dizionario e il suo conteggio nella recensione.**
- 3. Abbiamo applicato il modello LDA (Latent Dirichlet Allocation) utilizzando la funzione LdaMulticore() di gensim. Abbiamo specificato num_topics=6 per creare un modello LDA con 6 argomenti. L'argomento chunksize indica il numero di documenti da utilizzare in ciascun blocco di elaborazione, mentre passes indica il numero di volte che il modello deve elaborare l'intero corpus.**

Questi passaggi ci hanno permesso di creare un modello LDA che è in grado di classificare gli argomenti presenti nelle recensioni.

```

[(0,
[('good', 0.02844367),
 ('food', 0.025269046),
 ('beer', 0.022207228),
 ('staff', 0.02079964),
 ('nice', 0.016814645),
 ('visit', 0.016800229),
 ('irish', 0.01630163),
 ('friendly', 0.015107621),
 ('place', 0.011667512),
 ('rome', 0.011322025)]),
(1,
[('great', 0.026506089),
 ('place', 0.024392087),
 ('food', 0.020235416),
 ('beer', 0.013144059),
 ('irish', 0.0117692305),
 ('nice', 0.011734887),
 ('good', 0.011138825),
 ('service', 0.0088787945),
 ('stop', 0.0088535845),
 ('chicken', 0.00864452)])],
(2,
[('rome', 0.031773787),
 ('good', 0.025878757),
 ('irish', 0.024523439),
 ('night', 0.023405353),
 ('place', 0.013401851),
 ('go', 0.013400362),
 ('food', 0.012124442),
 ('scholar', 0.011624968),
 ('drink', 0.011290523),
 ('time', 0.0072042914)]),
(3,
[('good', 0.051754437),
 ('irish', 0.03440265),
 ('great', 0.030134788),
 ('food', 0.02822849),
 ('place', 0.025298407),
 ('rome', 0.024464762),
 ('beer', 0.019495593),
 ('staff', 0.013468893),
 ('nice', 0.011601005),
 ('service', 0.01074538)]),
(4,
[('great', 0.058378834),
 ('rome', 0.036650382),
 ('staff', 0.032753337),
 ('irish', 0.029755674),
 ('food', 0.026476828),
 ('friendly', 0.021939384),
 ('atmosphere', 0.020730112),
 ('place', 0.019305997),
 ('good', 0.016181719),
 ('scholar', 0.014017879)],
(5,
[('drink', 0.05047168),
 ('great', 0.032104734),
 ('good', 0.015894417),
 ('pint', 0.014042235),
 ('night', 0.013227001),
 ('time', 0.0125924),
 ('come', 0.011722964),
 ('nice', 0.010471933),
 ('staff', 0.010052804),
 ('expensive', 0.009914028)])
]

```

I topic models generati tramite l'analisi del modello LDA ci offrono una panoramica interessante delle principali tematiche emerse dalle recensioni dei clienti. Ognuno di questi argomenti rappresenta un aspetto chiave dell'esperienza vissuta presso il locale.

Il primo argomento riguarda l'apprezzamento per la qualità del cibo e della birra offerti dalla locale. I clienti hanno probabilmente elogiato la bontà dei piatti e la vasta selezione di birre disponibili.

Il secondo argomento sottolinea l'importanza dell'atmosfera e dell'ospitalità offerte dal pub. I clienti hanno probabilmente apprezzato l'ambiente accogliente e il personale amichevole, che ha contribuito a creare un'esperienza positiva per i visitatori.

Il terzo argomento evidenzia il connubio unico tra la cultura romana e l'identità irlandese presenti nel locale. I clienti hanno probabilmente apprezzato la possibilità di immergersi nella tradizione irlandese mentre si trovavano a Roma, creando un'atmosfera unica e coinvolgente.

Il quarto argomento sottolinea nuovamente l'importanza della qualità del cibo e del servizio. I clienti hanno probabilmente riconosciuto la bontà degli ingredienti utilizzati nella preparazione dei piatti, così come l'attenzione e la cortesia del personale.

Il quinto argomento mette in luce la reputazione del locale come luogo eccezionale. I clienti hanno probabilmente elogiato la sua grandezza e raccomandato vivamente il pub ad altri visitatori, contribuendo a diffondere la sua fama.

Infine, il sesto argomento si concentra sull'apprezzamento per le bevande, come birra e cocktail, offerte dal pub. I clienti hanno probabilmente evidenziato la qualità delle bevande e la piacevolezza delle serate trascorse a sorseggiarle. È importante notare che la menzione della parola "expensive" suggerisce che il locale non è economico.

Complessivamente, questi topic models ci permettono di comprendere meglio l'esperienza dei clienti presso lo Scholars. Essi evidenziano l'importanza della qualità del cibo e delle bevande, dell'atmosfera accogliente e dell'ospitalità del personale, nonché della reputazione e della peculiarità del locale.

Durante l'analisi, abbiamo anche calcolato i valori di Perplessità e Coerenza, che ci forniscono ulteriori indicazioni sull'efficacia del modello. La Perplessità, con un valore di -6.756, rappresenta una misura di quanto il modello sia in grado di predire correttamente i documenti nel corpus. In questo caso, un valore più basso indica una maggiore precisione e coerenza nelle previsioni del modello.

È importante notare che la Perplessità è un valore negativo, il che può sembrare controiduitivo. Tuttavia, in questo contesto, il valore negativo non altera il suo significato. L'obiettivo è sempre quello di ottenere un valore di Perplessità più basso possibile, poiché ciò indica una migliore aderenza del modello ai dati di input e una maggiore capacità di generalizzazione.

D'altra parte, la Coerenza, con un valore di 0.2667, rappresenta la misura della coesione dei topic generati dal modello. Un valore di coerenza più alto indica che i termini all'interno di ogni topic sono strettamente correlati e che i topic stessi sono distinti tra loro. Sebbene il valore di coerenza possa sembrare relativamente basso, è importante considerare che la coerenza dipende anche dalla natura del corpus e dalla complessità degli argomenti trattati. In questo contesto, un valore di coerenza di 0.2667 può essere considerato accettabile.

In definitiva, la Perplessità e la Coerenza sono misure complementari che ci aiutano a valutare la qualità del modello LDA. La Perplessità ci fornisce informazioni sulla sua capacità di predizione, mentre la Coerenza evidenzia la coesione e la distinzione dei topic generati.

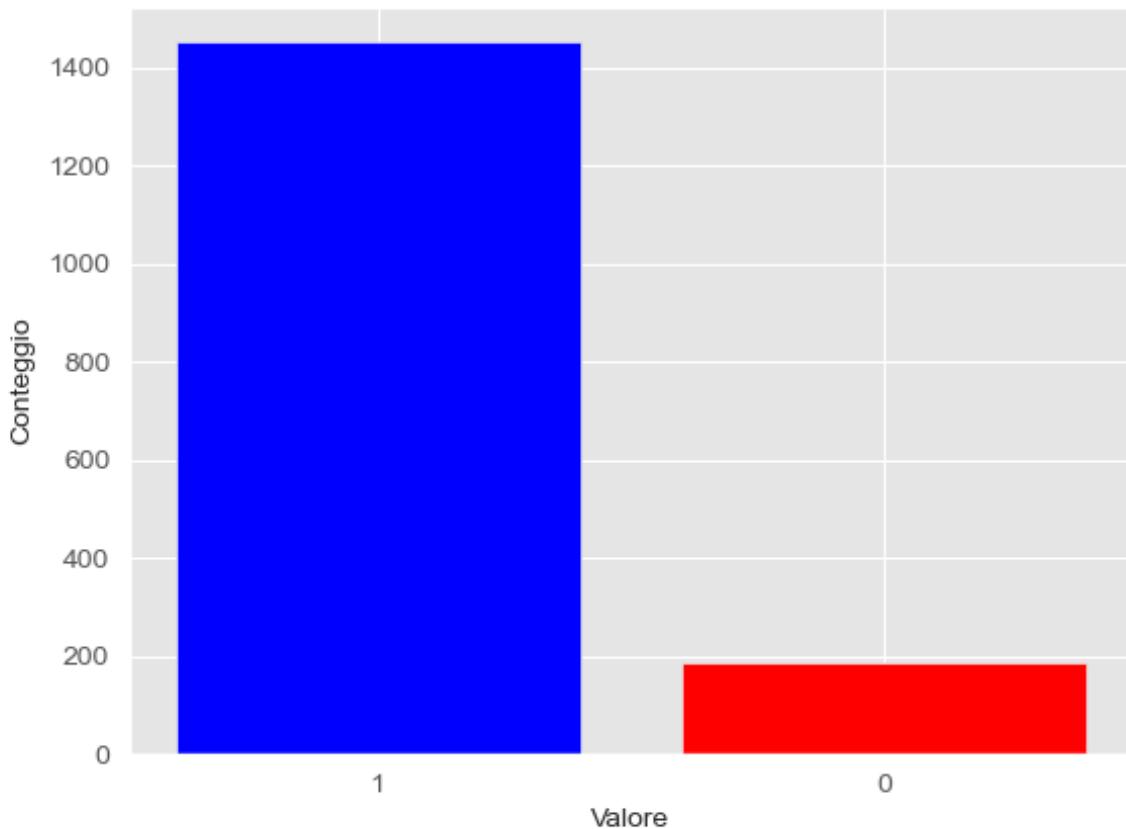
Parole con coefficiente più grande e più piccolo

Nel nostro studio, abbiamo adottato una metodologia di analisi basata sulla classificazione delle recensioni come positive o negative utilizzando un modello di regressione logistica. Questa scelta ci ha permesso di esaminare le caratteristiche linguistiche delle recensioni e di identificare i termini che contribuiscono maggiormente alla classificazione di una recensione come positiva o negativa.

Per prima cosa, abbiamo convertito la colonna 'Rating' nel nostro DataFrame in un tipo di dati intero, al fine di facilitare le operazioni successive. Successivamente, abbiamo

eliminato le righe che avevano un valore di 'Rating' pari a 3, poiché volevamo concentrarci sulle recensioni positive e negative.

Per creare un DataFrame più focalizzato, abbiamo aggiunto una nuova colonna chiamata 'Positively Rated', che indica se una recensione è valutata positivamente (1) o meno (0). Abbiamo assegnato il valore 1 alle recensioni con un punteggio superiore a 3 e il valore 0 alle recensioni con un punteggio inferiore o uguale a 3. Questo ci ha permesso di identificare chiaramente le recensioni positive e negative nel nostro dataset.



Successivamente, abbiamo suddiviso il DataFrame in un set di addestramento e un set di test utilizzando la funzione `train_test_split`. Il set di addestramento è stato utilizzato per addestrare il nostro modello di regressione logistica, mentre il set di test è stato utilizzato per valutarne le prestazioni.

Per trasformare le recensioni testuali in una rappresentazione numerica, abbiamo utilizzato il vettorizzatore `CountVectorizer`, che ha convertito le recensioni in una matrice di conteggio delle parole. Questa rappresentazione ci ha consentito di identificare le parole chiave nelle recensioni che sono associate a una valutazione positiva o negativa.

Abbiamo quindi creato un'istanza del modello di regressione logistica e lo abbiamo addestrato utilizzando il set di addestramento vettorizzato e i valori target corrispondenti alle recensioni positive e negative. La regressione logistica è stata

scelta perché è un algoritmo di classificazione efficace per problemi binari come il nostro.

Successivamente, abbiamo utilizzato il modello addestrato per effettuare previsioni sul set di test utilizzando il vettorizzatore e abbiamo calcolato l'area sotto la curva ROC (AUC) come misura delle prestazioni del nostro modello. L'area sotto la curva ROC fornisce una valutazione della capacità del nostro modello di distinguere tra recensioni positive e negative. Nel nostro caso, l'AUC è risultata essere 0.7615, il che indica una buona capacità predittiva del nostro modello.

Infine, abbiamo estratto le parole con i coefficienti più grandi e più piccoli dal modello di regressione logistica. I coefficienti più grandi indicano le parole che hanno una maggiore influenza nella classificazione delle recensioni come positive, mentre i coefficienti più piccoli indicano le parole che sono associate a una classificazione negativa.

I 10 coefficienti più grandi sono: ['great', 'friendly', 'excellent', 'rome', 'amazing', 'fantastic', 'best', 'mark', 'good', 'football']. Queste parole sono indicative di una valutazione positiva delle recensioni. Al contrario, i 10 coefficienti più piccoli sono: ['terrible', 'poor', 'no', 'horrible', 'disappointed', 'not', 'bad', 'over', 'awful', 'worst']. Queste parole sono associate a una valutazione negativa delle recensioni.

Questo ci fornisce una panoramica delle parole chiave che influenzano la valutazione positiva o negativa delle recensioni e ci aiuta a capire meglio i fattori che possono influire sull'opinione degli utenti.

In conclusione, l'utilizzo della classificazione attraverso la regressione logistica ci ha permesso di analizzare le recensioni e identificare le parole chiave che influenzano la valutazione positiva o negativa. I risultati ottenuti, insieme all'area sotto la curva ROC, indicano che il nostro modello ha una buona capacità predittiva. Ciò ci consente di comprendere meglio le opinioni degli utenti e di individuare le parole chiave che influenzano la percezione generale di un prodotto o servizio.

Analisi in R

Nel contesto di R, abbiamo seguito un processo simile a quello svolto in Python per la pulizia del dataset e l'analisi delle recensioni. Dopo aver eseguito le operazioni di pulizia dei dati, abbiamo creato due dataset distinti: uno per i titoli delle recensioni e uno per il testo delle recensioni.

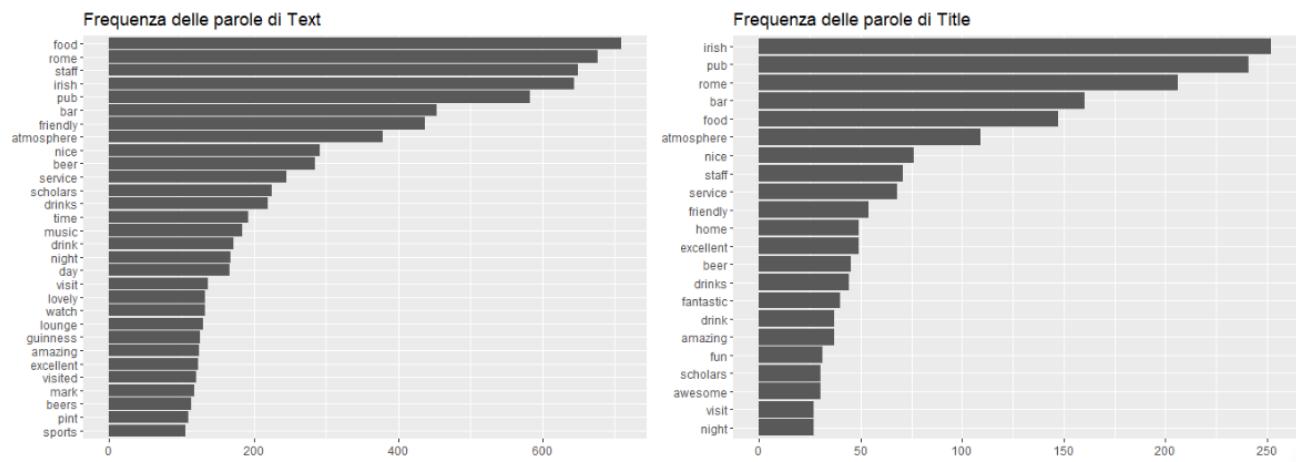
Per effettuare la tokenizzazione del testo delle recensioni, abbiamo utilizzato la funzione `unnest_tokens()` del pacchetto `tidytext`. Questa funzione ci ha consentito di suddividere il testo delle recensioni in singole parole o token. Inoltre, per rimuovere le stopwords, abbiamo utilizzato la funzione `str_detect()` del pacchetto `stringr`.

Successivamente, abbiamo analizzato nuovamente le colonne "Rating" e "Date" nei nuovi dataset creati. Queste colonne hanno prodotto lo stesso risultato già analizzato

in precedenza, confermando la distribuzione delle valutazioni e la temporaneità delle recensioni.

Frequenza delle parole

Come per l'analisi svolta in Python, anche in R abbiamo iniziato analizzando la frequenza delle parole. Tuttavia, in questo caso abbiamo condotto l'analisi separatamente per il testo e il titolo delle recensioni.



L'analisi della frequenza delle parole condotta in R non ha prodotto risultati diversi rispetto all'analisi effettuata in Python. Inoltre, non sono emerse differenze significative nelle frequenze delle parole tra il titolo e il testo delle recensioni. I risultati ottenuti confermano la coerenza tra i due approcci e suggeriscono che le parole più comuni all'interno dei titoli e dei testi delle recensioni sono simili.

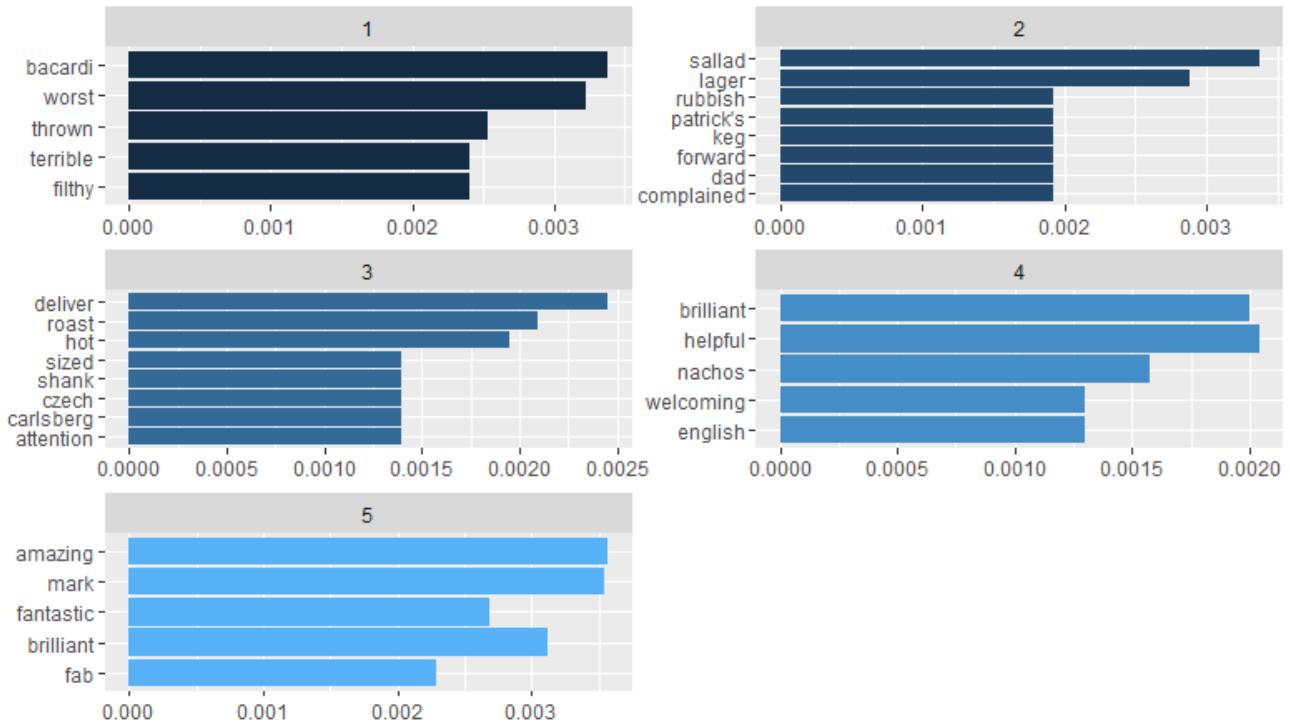
Proseguendo con l'analisi della frequenza delle parole, abbiamo esteso il nostro studio in R includendo l'analisi basata sull'anno e sulle stelle delle recensioni. Sia per il testo che per i titoli delle recensioni, abbiamo esaminato le parole più frequenti in base all'anno di pubblicazione e al punteggio assegnato. Tuttavia, da entrambe le analisi, non sono emersi risultati significativi o degni di nota.

TF-IDF

La misurazione del TF-IDF (Term Frequency-Inverse Document Frequency) è un'importante tecnica utilizzata nell'analisi del testo per valutare l'importanza di una parola in un documento rispetto all'intero corpus. Il TF-IDF tiene conto sia della frequenza di una parola nel documento (TF) che dell'importanza della parola nell'intero corpus (IDF).

Nel codice fornito, è stato calcolato il valore del TF-IDF per le parole raggruppate per valutazione (Rating) nel dataframe "words_by_stars". Utilizzando la funzione "bind_tf_idf", il TF-IDF è stato legato alla parola, al Rating e alla frequenza n. Il dataframe è stato poi ordinato in modo decrescente in base al valore del TF-IDF.

TF-IDF di Text



Osservando le parole con i punteggi più alti di TF-IDF per ciascuna valutazione, emerge un quadro interessante delle caratteristiche più rilevanti associate a diverse esperienze di recensione.

Nel caso delle recensioni con una valutazione di 1 stella, le parole come "worst", "terrible" e "filthy" indicano un'esperienza negativa e deludente. La presenza di parole come "bacardi" potrebbe suggerire un'associazione con un'esperienza di bevande alcoliche poco gradevole.

Per le recensioni con una valutazione di 2 stelle, la presenza di parole come "sallad" e "lager" potrebbe indicare un'insoddisfazione riguardo all'offerta di cibo o bevande.

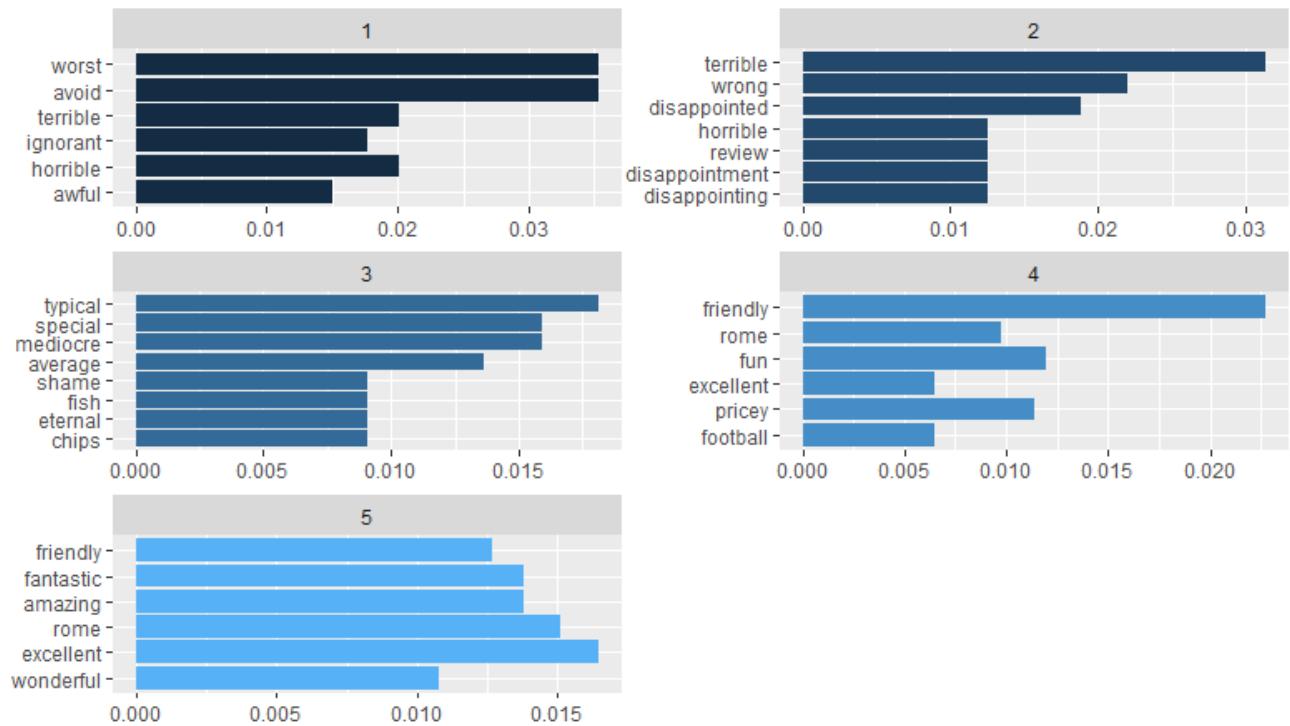
Le recensioni con una valutazione di 3 stelle presentano parole come "deliver", "roast" e "hot", che potrebbero indicare una valutazione neutra o una moderata soddisfazione. La presenza di parole come "sized", "shank" e "czech" potrebbe suggerire l'inclusione di aspetti specifici relativi al cibo o all'ambiente.

Le recensioni con una valutazione di 4 stelle mostrano parole come "brilliant", "helpful" e "welcoming", che suggeriscono un'esperienza positiva con apprezzamenti per il servizio e l'ospitalità ricevuti. La presenza di parole come "nachos" ed "english" potrebbe indicare una valutazione positiva riguardo a specifici piatti o esperienze multilingue offerte.

Infine, le recensioni con una valutazione di 5 stelle presentano parole come "amazing", "fantastic" e "brilliant", che indicano un'esperienza eccezionalmente positiva e

soddisfacente. Questi risultati suggeriscono che le recensioni con una valutazione di 5 stelle sono caratterizzate da forti elogi e una valutazione estremamente positiva dell'esperienza complessiva.

TF-IDF di Title



L'analisi TF-IDF condotta sulla colonna "Title" in base alle frequenze e al Rating ha fornito risultati interessanti riguardo alle parole con punteggio più alto per ciascuna valutazione.

Nel caso delle recensioni con una valutazione di 1 stella, le parole con un alto punteggio di TF-IDF come "worst", "avoid", "terrible", "ignorant", "horrible" e "awful" indicano un'esperienza estremamente negativa e deludente.

Per le recensioni con una valutazione di 2 stelle, le parole con un alto punteggio di TF-IDF come "terrible", "wrong", "disappointed", "horrible", "disappointment" e "disappointing" indicano un'esperienza insoddisfacente e deludente.

Le recensioni con una valutazione di 3 stelle presentano parole con un alto punteggio di TF-IDF come "typical", "special", "mediocre", "average" (medio), "shame", "fish" e "chips". Queste parole suggeriscono un'esperienza neutra o mediamente soddisfacente. È interessante notare la menzione di "fish and chips", che potrebbe indicare che il piatto offerto dal locale non era all'altezza delle aspettative, ma viene comunque considerato come una valutazione media o neutra.

Le recensioni con una valutazione di 4 stelle presentano parole con un alto punteggio di TF-IDF come "friendly", "fun", "excellent", "pricey" (costoso). Queste parole indicano un'esperienza generalmente positiva, con un'enfasi sull'ospitalità, il divertimento e la

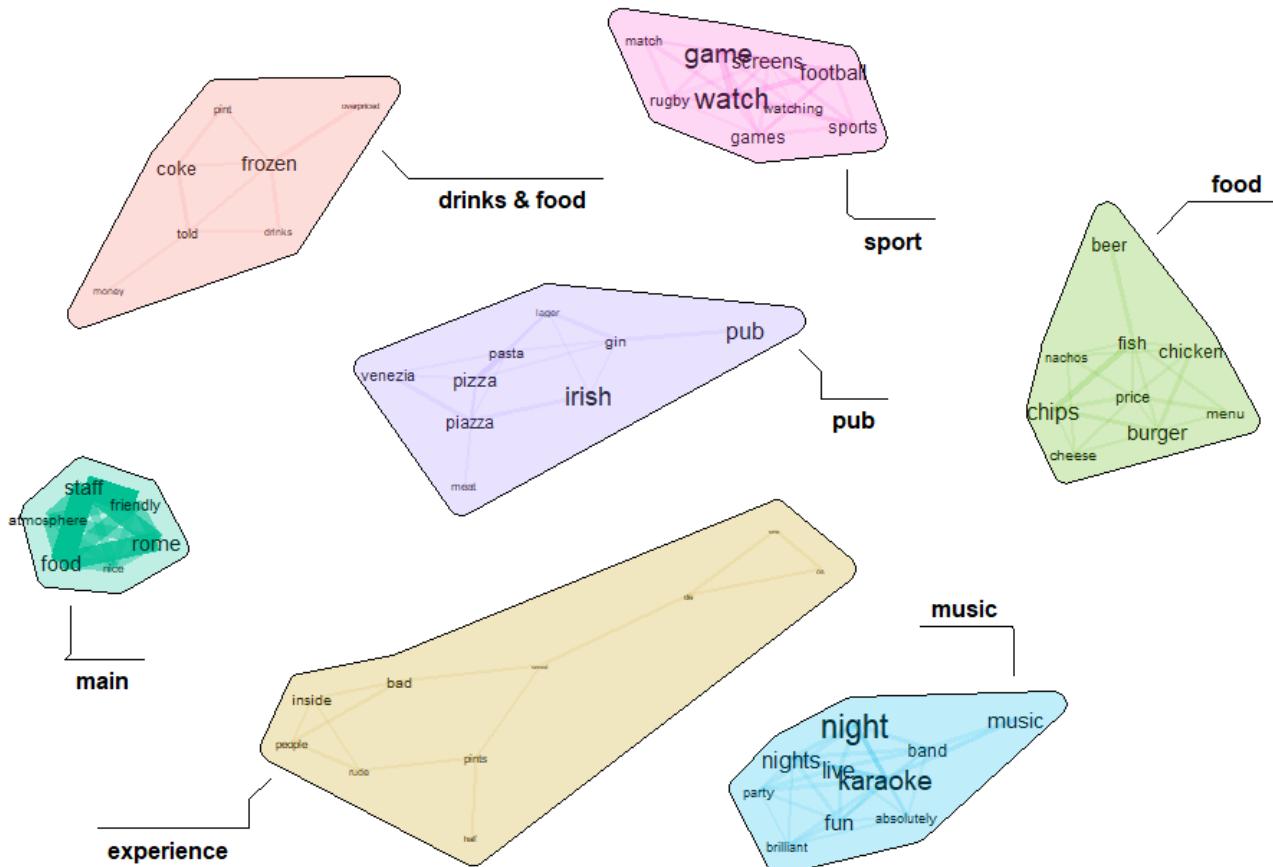
qualità eccellente. Tuttavia, potrebbe essere menzionato anche il fatto che alcuni potrebbero considerare i prezzi un po' alti.

Infine, le recensioni con una valutazione di 5 stelle presentano parole con un alto punteggio di TF-IDF come "friendly", "fantastic", "amazing", "excellent" e "wonderful". Queste parole evidenziano un'esperienza eccezionalmente positiva e soddisfacente, vengono elogiate l'ospitalità, la qualità eccellente e l'esperienza complessiva del locale.

Topic Modeling

Per l'analisi dei topic, abbiamo adottato un approccio di topic modeling utilizzando la libreria BTM (Biterm Topic Model) in R. Abbiamo selezionato un insieme di parole dalla colonna "word" del dataset pulito, che rappresentava i testi delle recensioni.

BTM model



Abbiamo specificato di voler generare 7 topic utilizzando la libreria BTM, e questi sono i risultati ottenuti:

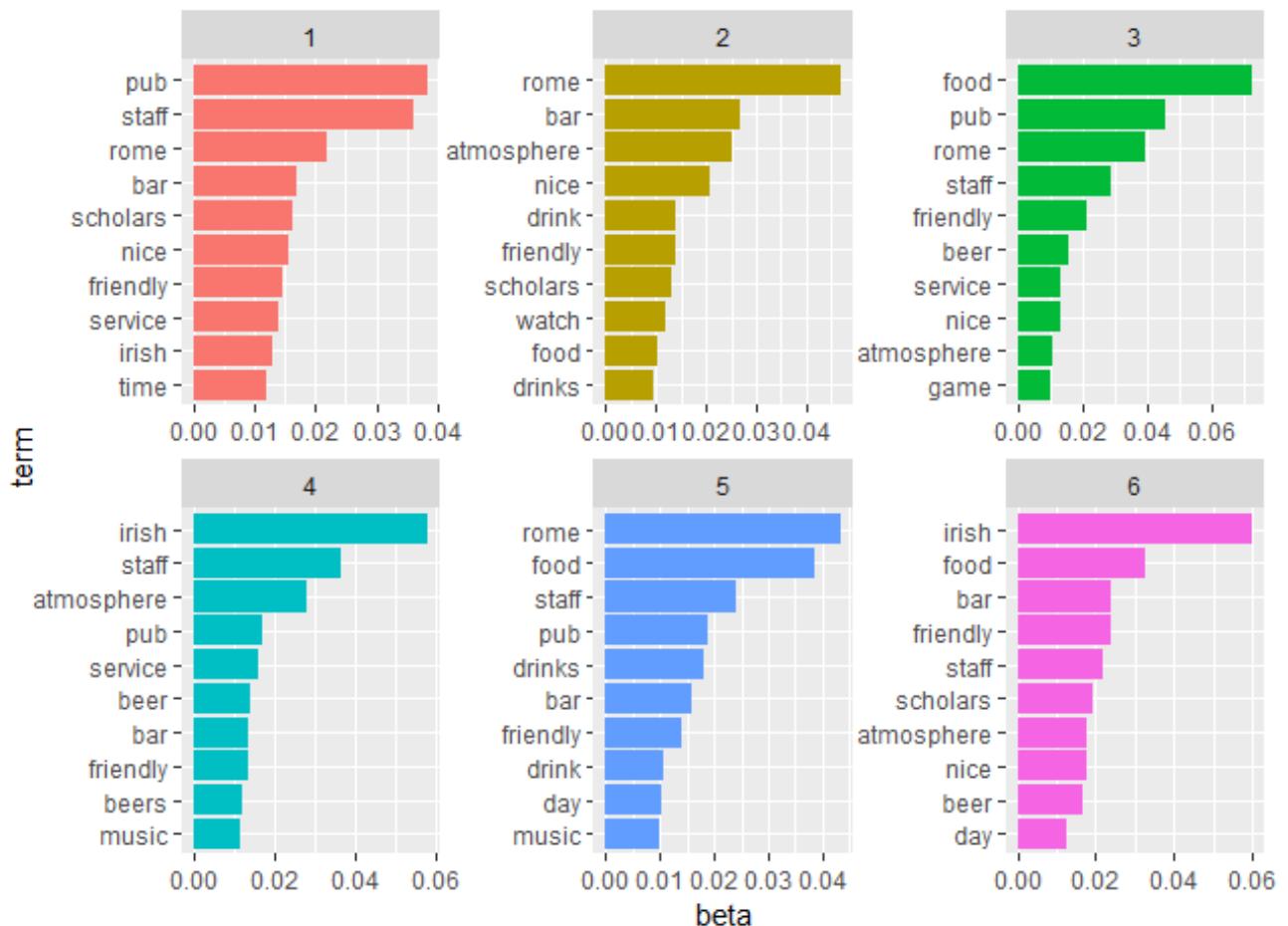
- Il primo topic è caratterizzato dai seguenti termini chiave: "food", "rome", "irish", "staff", "pub", "friendly", "atmosphere". Questi termini rappresentano la descrizione principale del locale, come già riscontrato in precedenza.**
- Il secondo topic è caratterizzato dai seguenti termini chiave: "watch", "game", "football", "screens", "bar". Questo argomento è chiaramente legato allo sport e si riferisce alla visione di partite ed eventi sportivi. I termini suggeriscono la**

- presenza di schermi dove è possibile seguire le partite. È probabile che le recensioni associate a questo topic riguardino l'ambiente sportivo del locale.**
- 3. Il terzo topic è caratterizzato dai seguenti termini chiave: "frozen", "coke", "food", "chicken", "drink". Questo argomento sembra concentrarsi sulle bevande e sul cibo offerto dal locale. I termini indicano la presenza di bevande ghiacciate, Coca-Cola, cibo (in particolare pollo) e altre bevande. Potrebbe rappresentare recensioni focalizzate sulla qualità delle bevande, la varietà di cibo e le esperienze culinarie nel complesso.**
 - 4. Il quarto topic è caratterizzato dai seguenti termini chiave: "bad", "inside", "people", "pints", "rude". Questo argomento sembra affrontare gli aspetti negativi dell'esperienza nel locale. I termini chiave come "bad", "rude" e "pints" potrebbero indicare recensioni negative sul servizio, l'atmosfera o l'atteggiamento del personale. Questo topic potrebbe evidenziare le esperienze meno soddisfacenti dei clienti.**
 - 5. Il quinto topic è caratterizzato dai seguenti termini chiave: "night", "karaoke", "live", "music", "fun". Questo argomento si concentra sulla musica e sull'aspetto divertente dell'esperienza nel locale. I termini suggeriscono la presenza di serate musicali, karaoke dal vivo e un'atmosfera divertente. Le recensioni associate a questo topic potrebbero essere incentrate sulla qualità delle esibizioni musicali e sull'atmosfera generale del locale.**
 - 6. Il sesto topic è caratterizzato dai seguenti termini chiave: "chips", "burger", "beer", "food", "chicken". Questo argomento è chiaramente legato al cibo, con menzione specifica di patatine fritte, hamburger, birra e pollo. Potrebbe rappresentare recensioni o commenti sulla qualità del cibo e delle bevande offerte dal locale, con una focalizzazione su piatti come patatine, hamburger, pollo e birra.**
 - 7. Il settimo topic è caratterizzato dai seguenti termini chiave: "irish", "pub", "pizza", "piazza", "beer". Questo argomento sembra essere centrato sul concetto di pub e sulla sua atmosfera. Potrebbe rappresentare recensioni o commenti sul pub stesso, inclusa l'atmosfera, la varietà di birre e la presenza di piatti come la pizza.**

Dopodiché, abbiamo effettuato la trasformazione del dataset tokenizzato e pulito in una matrice documento-termine (Document-Term Matrix, DTM) utilizzando la funzione `cast_dtm()` della libreria `tidytext`. Successivamente, abbiamo utilizzato la funzione `LDA()` della libreria `topicmodels` con `k=6` per creare un modello di Latent Dirichlet Allocation (LDA) con 6 argomenti. Questa funzione restituisce un oggetto che contiene tutti i dettagli relativi all'adattamento del modello, comprese le parole associate a ciascun argomento.

Per convertire la matrice documento-termine in una tabella di dati ordinata, abbiamo utilizzato la funzione `tidy()`. Il pacchetto `tidytext` fornisce questo metodo per estrarre le probabilità per-argomento-per-parola, chiamate "beta". Quello che abbiamo fatto è stato trasformare il modello in un formato in cui c'è una riga per ogni termine e un argomento corrispondente. Il modello calcola la probabilità che quel termine sia generato da quell'argomento.

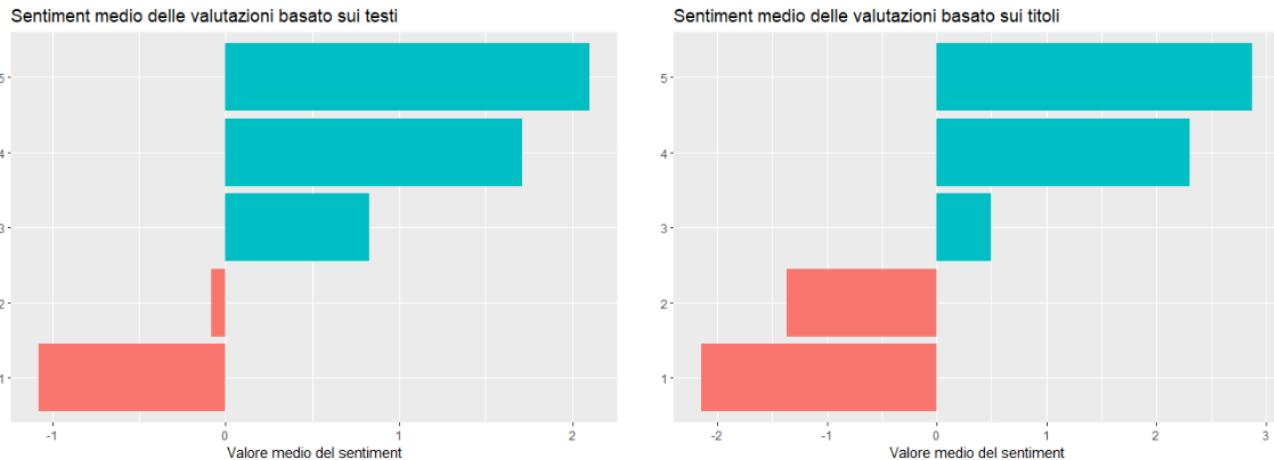
Infine, abbiamo utilizzato la funzione `slice_max()` della libreria `dplyr` per individuare i 10 termini più comuni all'interno di ciascun argomento.



Complessivamente, i modelli di argomenti ottenuti presentano notevoli somiglianze con quelli riscontrati utilizzando Python. Questi modelli ci consentono di ottenere una migliore comprensione dell'esperienza dei clienti presso lo Scholars. Emergono chiaramente l'importanza della qualità del cibo e delle bevande, dell'atmosfera accogliente e dell'ospitalità del personale, nonché la rilevanza della reputazione e delle caratteristiche peculiari del locale.

Sentiment Analysis

Utilizzando le tecniche di analisi del sentiment, possiamo esaminare la frequenza delle parole positive e negative nelle recensioni e nei relativi titoli. Nel primo studio condotto, l'obiettivo principale era analizzare le parole con un grado di positività variabile in base alle valutazioni stellari. Per questa analisi, abbiamo utilizzato il lessico dei sentimenti AFINN, che assegna un valore numerico di positività a ciascuna parola.

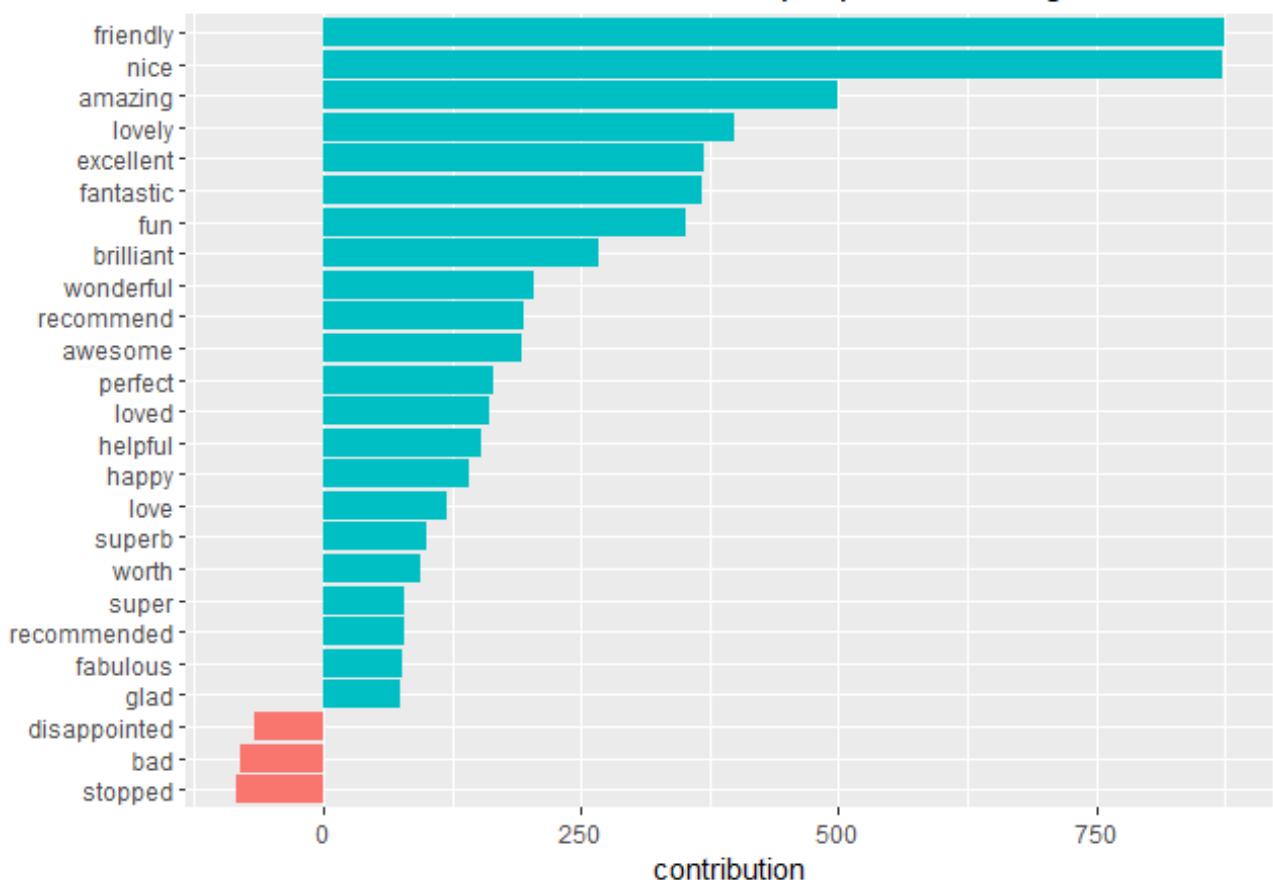


Nel primo grafico a sinistra, che rappresenta il sentimento dei testi in base alle valutazioni stellari, osserviamo una situazione che sembra essere nella norma. La positività delle parole utilizzate nelle recensioni tende ad essere più elevata nelle recensioni a 5 stelle, diminuendo gradualmente fino a diventare negativa nelle recensioni a 1 e 2 stelle.

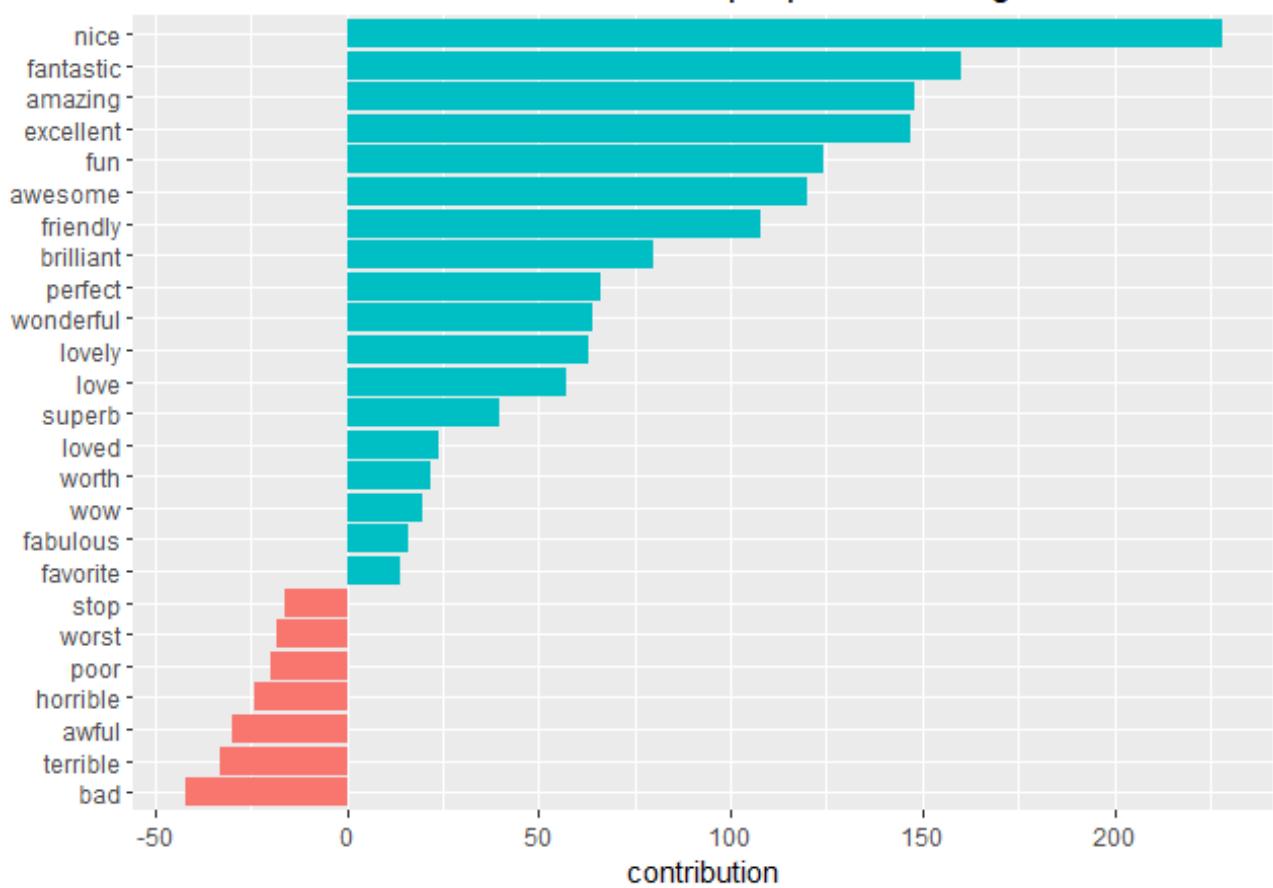
Nel grafico a destra, che descrive il sentimento dei titoli delle recensioni in base alle valutazioni stellari, notiamo una situazione simile, ma con un valore di sentimento più elevato. In generale, i risultati dell'analisi dei testi delle recensioni rimangono coerenti, poiché anche i risultati dei titoli delle recensioni rimangono simili.

Per comprendere il motivo per cui alcuni gruppi di recensioni risultano essere più positivi o negativi rispetto ad altri, abbiamo valutato il contributo complessivo delle parole in termini di positività e negatività. Successivamente, ci siamo focalizzati sull'analisi specifica delle valutazioni stellari per ottenere una visione più dettagliata.

Parole che influenzano un sentimento più positivo o negativo nei testi



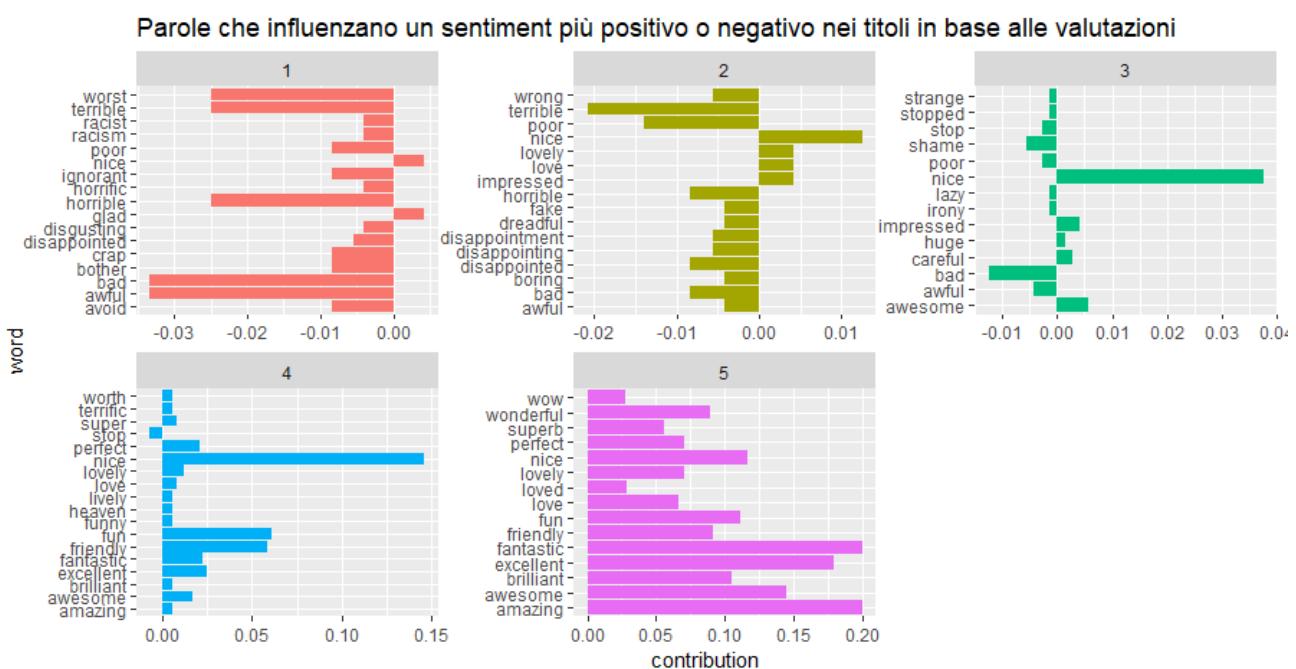
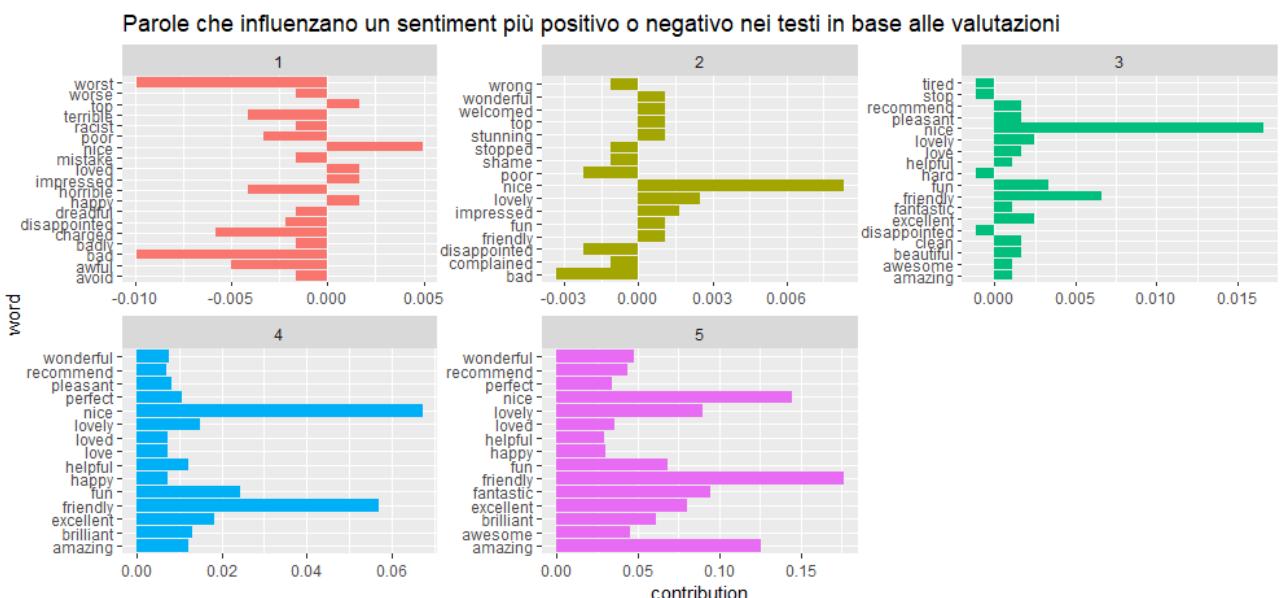
Parole che influenzano un sentimento più positivo o negativo nei titoli



Dai due grafici presentati di sopra emerge che la maggior parte delle parole utilizzate nei testi delle recensioni, così come nei titoli, sono di natura positiva. Nel primo grafico, alcune delle parole più positive comprendono "friendly", "nice", "amazing", "fantastic" e "excellent", mentre tra quelle più negative troviamo "dissatisfied", "bad", "worst", "awful" e "terrible".

È importante notare che in alcune occasioni potrebbero verificarsi degli errori, poiché alcune parole potrebbero essere utilizzate in contesti negativi nonostante la loro natura positiva. Per affrontare questa problematica, abbiamo compiuto ulteriori analisi per risolvere tali ambiguità e ottenere risultati più accurati.

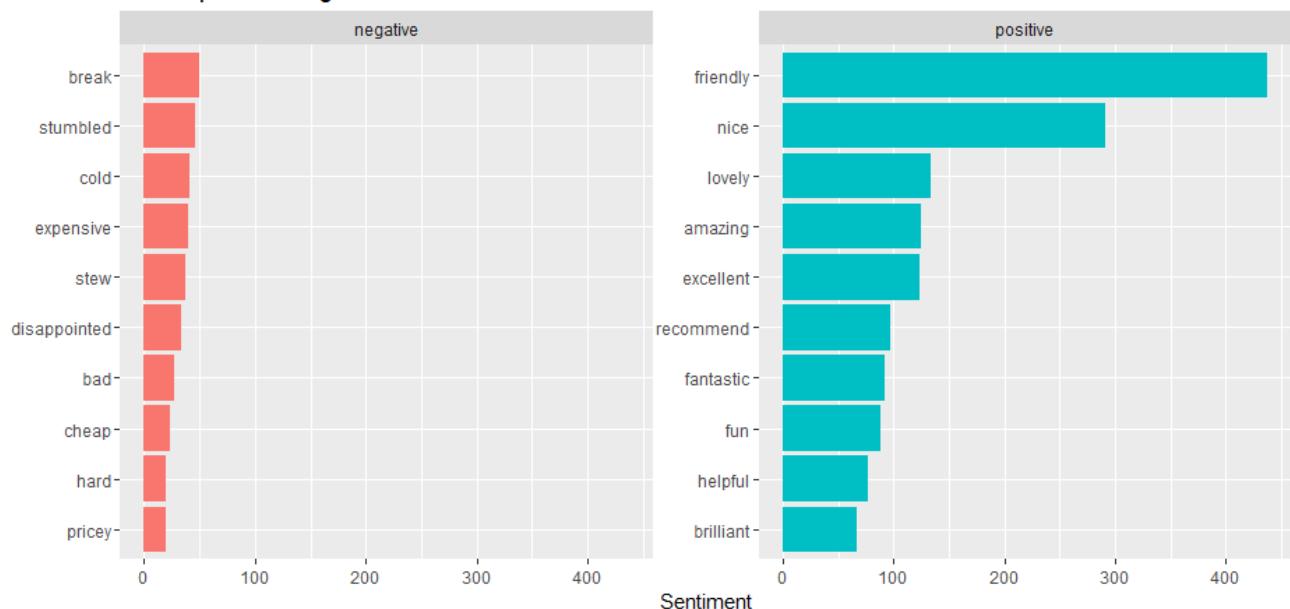
Abbiamo quindi condotto uno studio per identificare le parole che hanno maggiormente contribuito all'interno di ciascun gruppo di valutazioni stellari, al fine di individuare eventuali gruppi che potrebbero essere stati valutati in modo erroneo.



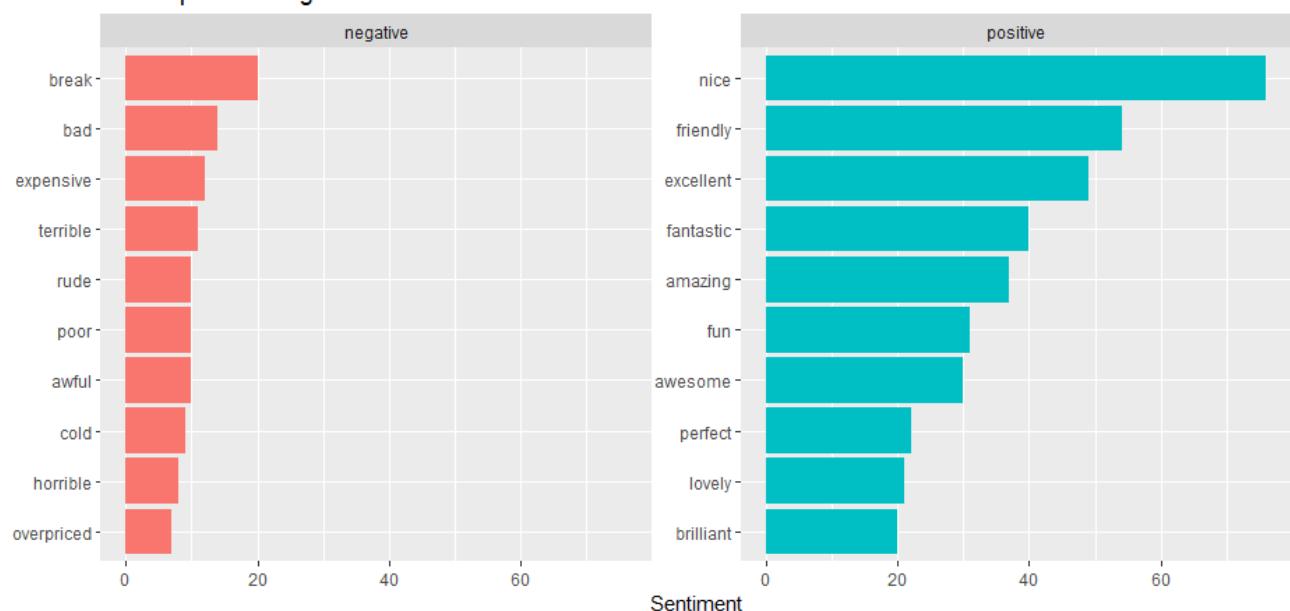
Il nostro studio delle recensioni all'interno dei diversi gruppi di valutazioni stellari ha prodotto risultati accurati ed equilibrati. Sia per i testi che per i titoli delle recensioni, come previsto, abbiamo riscontrato che nei gruppi con 1 o 2 stelle predominano le parole negative, risultando quindi classificate come tali con un punteggio di sentiment negativo. Al contrario, nei gruppi con 4 e 5 stelle, abbiamo riscontrato prevalentemente parole positive, riflettendo così un valore di sentiment positivo.

L'ultima analisi condotta riguardo il sentiment, è stata l'individuazione delle parole più comuni che esprimono sentiment positivo e negativo. Per questa analisi, abbiamo utilizzato il lessico BING, che permette la classificazione delle parole in modo binario nelle categorie positive e negative.

Parole positive/negative nei testi delle recensioni



Parole positive/negative nei titoli delle recensioni



negative



positive

Per ottenere una panoramica più completa e visivamente accattivante di queste analisi, abbiamo deciso di creare un word cloud separando le parole in due categorie: positive e negative. Tra le parole positive più comuni, riscontriamo ancora una volta "friendly", "nice", "lovely" e "amazing", che riflettono le caratteristiche positive del locale. D'altra parte, tra le parole negative, troviamo "bad", "expensive", "terrible" e "rude", che probabilmente indicano esperienze negative, costi elevati e persone scortese. Queste parole contribuiscono ad offrire una visione più completa del panorama generale.

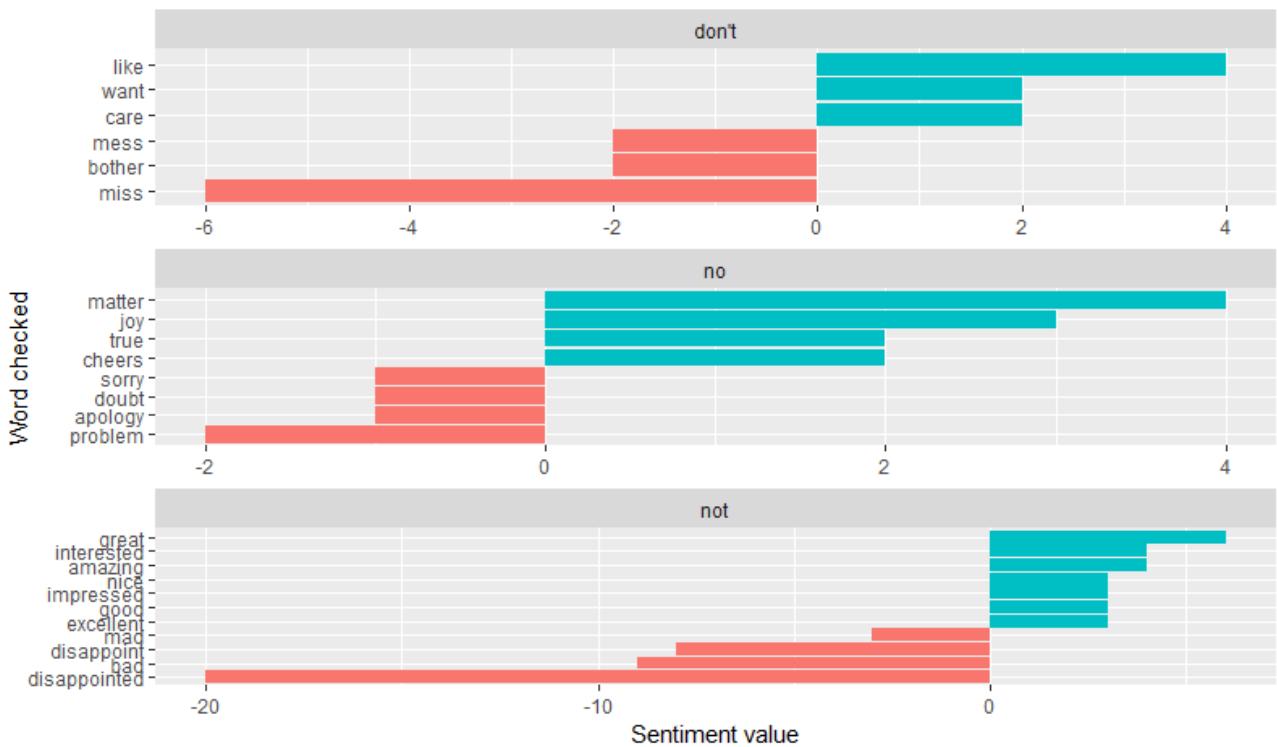
Analisi N-grammi

Per ottenere una visione più approfondita del nostro dataset, abbiamo considerato non solo le parole singolarmente, ma anche le relazioni tra di esse, analizzando quali parole tendono ad apparire vicine tra loro.

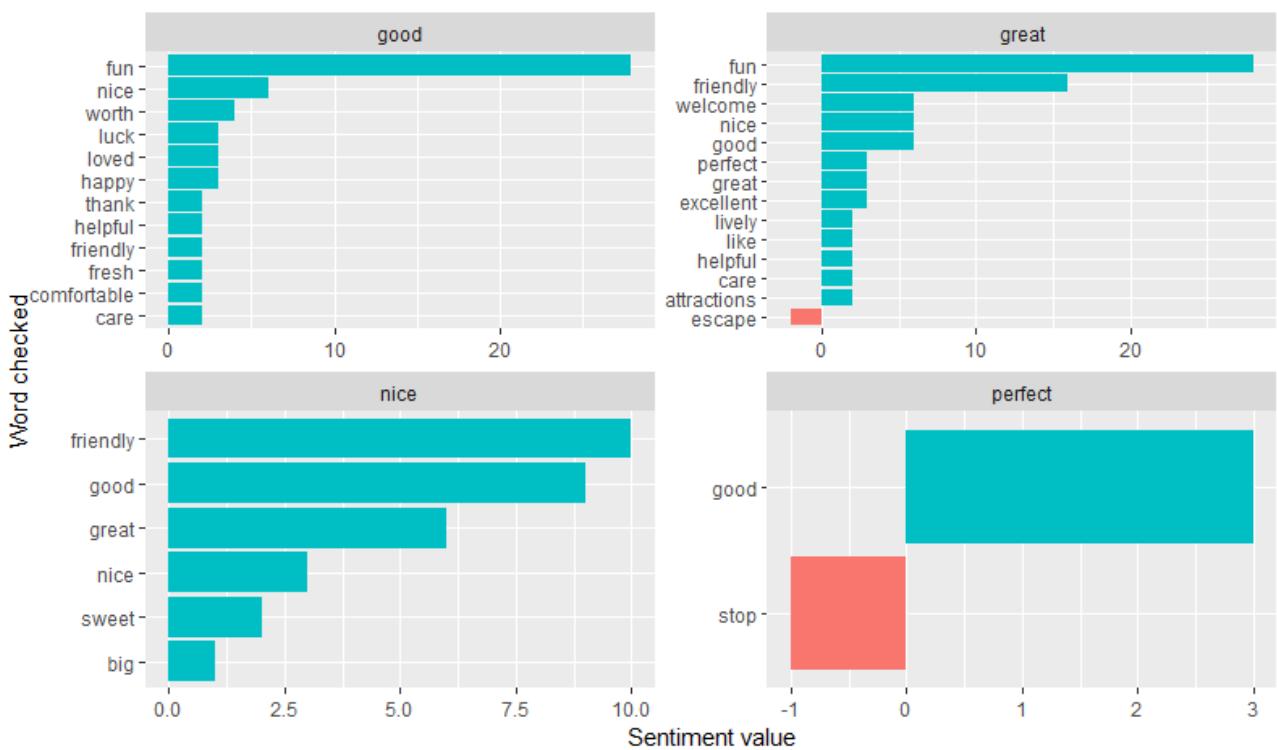
Per la tokenizzazione dei testi delle nostre recensioni in bigrammi, abbiamo utilizzato la funzione `unnest_tokens()`, specificando l'argomento "token" come "bigrams", in modo che le parole vengano suddivise in coppie di token anziché in singoli token.

Successivamente, dopo aver creato subset di dati contenenti i bigrammi, abbiamo condotto diverse analisi. Nel primo caso, abbiamo studiato l'uso di parole che possono

essere interpretate come positive o negative a seconda del contesto, se non vengono considerate insieme alla parola che le precede.

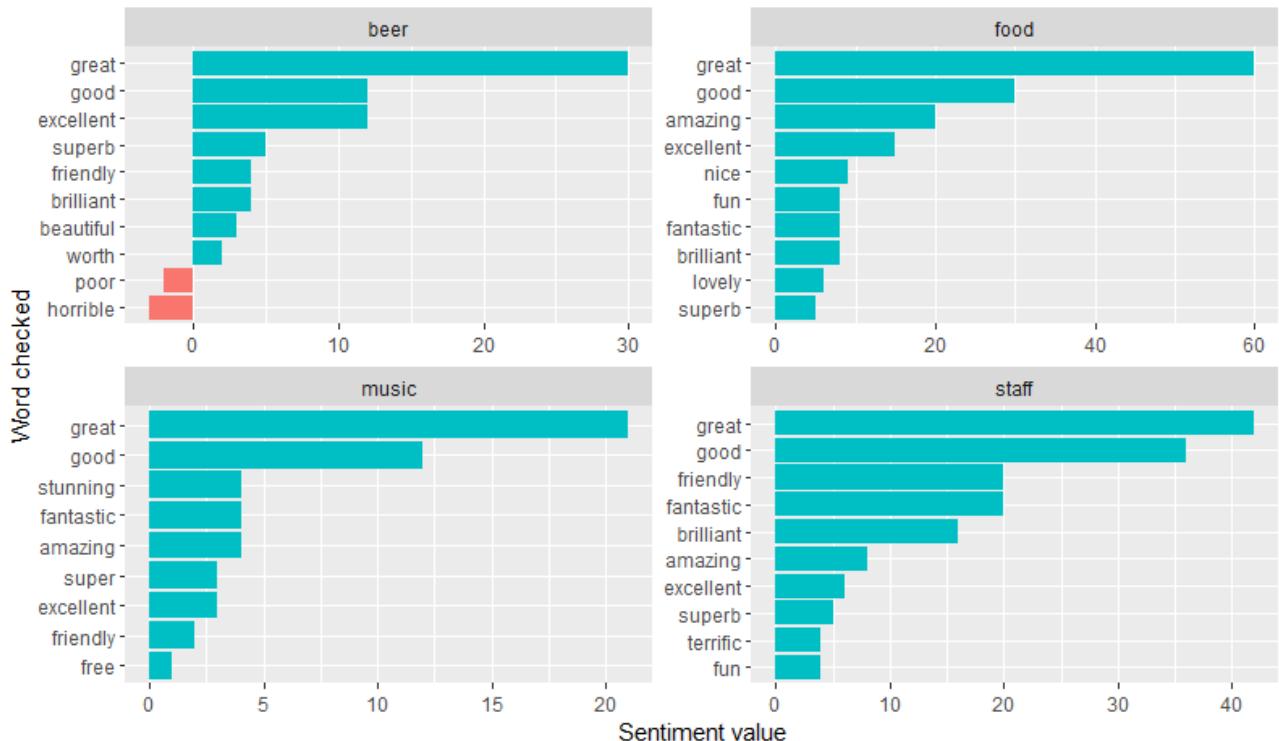


Nel primo grafico, che riguarda i lemmi di negazione, possiamo notare un interessante punto: le parole a cui vengono attribuite le negazioni non sono sempre intrinsecamente negative dal punto di vista del sentimento. Infatti, la presenza di una negazione può invertire il significato di una parola e renderla positiva nonostante la sua connotazione originaria potrebbe essere negativa.



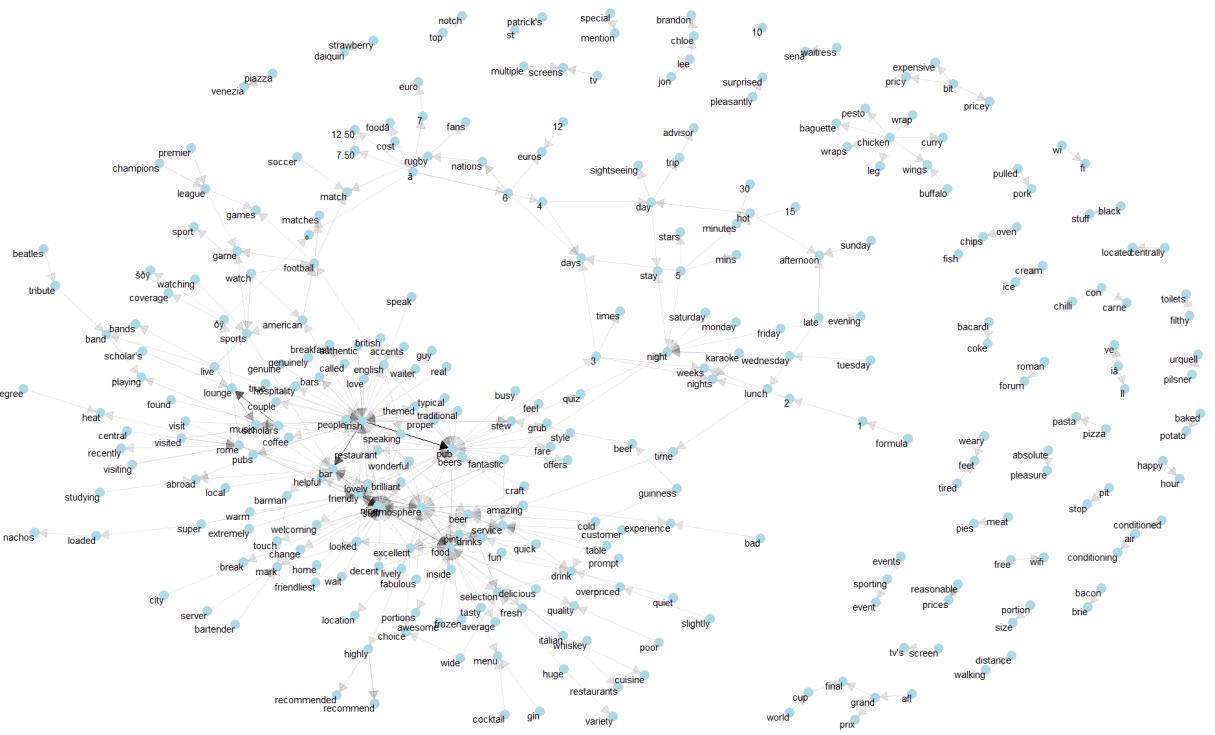
Nel caso delle parole positive, osserviamo che spesso sono associate ad altri lemmi che hanno un sentimento positivo. Ad esempio, se consideriamo il grafico delle parole associate a "perfect", notiamo che "stop" appare come un'associazione, nonostante la parola sia classificata con un sentimento negativo. Questo suggerisce che la combinazione di queste parole crea un'interpretazione positiva nonostante la presenza della parola negativa.

La seconda analisi che abbiamo condotto si è focalizzata sulle parole chiave per un pub come lo Scholars. Abbiamo preso in considerazione le parole "beer", "food", "music" e "staff", che sono rilevanti per valutare l'esperienza complessiva in un pub.



Il risultato ottenuto è esattamente ciò che ci aspettavamo da un pub di grande fama e reputazione come lo Scholars. Le uniche parole con un sentimento negativo sono associate alla parola "birra", tuttavia è importante notare che quest'ultima presenta un punteggio significativamente più alto nella parola "great". Inoltre, le altre parole prese in considerazione sono tutte accomunate dalla presenza della parola "great", a sottolineare l'eccellente qualità complessiva del locale.

L'ultimo studio che abbiamo condotto consisteva nella visualizzazione simultanea di tutte le parole.



Durante l'analisi degli n-grammi, abbiamo identificato diverse combinazioni di parole che si sono verificate frequentemente nelle recensioni. Questi n-grammi ci hanno fornito ulteriori informazioni sulla percezione dei clienti riguardo a vari aspetti del locale, del servizio e dell'esperienza complessiva.

Alcuni degli n-grammi significativi che abbiamo trovato includono:

- **"bit->pricey" e "bit->expensive"**: Questi n-grammi suggeriscono che i clienti hanno espresso l'opinione che il locale fosse un po' costoso o troppo caro. Queste espressioni indicano che il prezzo dei prodotti o dei servizi offerti potrebbe essere stato considerato elevato dai clienti.
- **"centrally->located"**: Questo n-gramma indica che il locale è situato in una posizione centrale. Questa informazione potrebbe essere stata menzionata dai clienti come un punto di forza, indicando la comodità dell'ubicazione del locale.
- **"filthy->toilets"**: Questo n-gramma suggerisce che i clienti hanno lamentato la scarsa pulizia dei servizi igienici. La parola "filthy" (sporco) suggerisce che i bagni potrebbero non essere stati mantenuti in condizioni igieniche adeguate.
- **"free->wifi"**: Questo n-gramma indica che il locale offre connessione Wi-Fi gratuita. Potrebbe essere un punto positivo menzionato dai clienti, evidenziando la comodità di poter accedere a Internet durante la visita al locale.
- **"absolute->pleasure" e "pleasantly->surprised"**: Questi n-grammi indicano un'esperienza positiva dei clienti. Esprimono soddisfazione e piacere nell'interagire con il locale o nel provare qualcosa di inaspettatamente buono. Queste espressioni suggeriscono che i clienti sono rimasti positivamente impressionati dall'esperienza complessiva.
- **"reasonable->prices" e "overpriced->drink"**: Questi n-grammi evidenziano l'opinione dei clienti riguardo ai prezzi offerti dal locale. "Reasonable->prices"

- suggerisce che i prezzi sono stati considerati ragionevoli, mentre "overpriced->drink" indica che una bevanda è stata considerata costosa rispetto al suo valore.**
- **"5->stars":** Questo n-gramma potrebbe essere associato a recensioni che assegnano una valutazione di 5 stelle al locale. Indica un'esperienza estremamente positiva, in cui il cliente elogia il locale con il massimo punteggio.
 - **"karaoke->nights":** Questo n-gramma suggerisce che il locale offre serate di karaoke. Potrebbe essere una caratteristica distintiva menzionata dai clienti che apprezzano l'opportunità di partecipare a divertenti sessioni di karaoke.
 - **"studying->abroad":** Questo n-gramma potrebbe essere associato a recensioni scritte da persone che si trovano a studiare all'estero. Indica che il locale è stato visitato da studenti che si trovano temporaneamente in un altro paese per motivi di studio.
 - **"busy->pub":** Questo n-gramma indica che il pub è frequentato e spesso affollato. Suggerisce che il locale è popolare e attrae un buon numero di clienti.
 - **"awesome->food" e "frozen->food":** Questi n-grammi indicano la qualità del cibo offerto dal locale. "Awesome->food" suggerisce che il cibo è stato considerato eccezionale, mentre "frozen->food" potrebbe indicare che alcuni clienti hanno lamentato la presenza di cibi surgelati.

Gli n-grammi identificati ci hanno fornito un'ulteriore comprensione delle opinioni dei clienti e delle loro esperienze all'interno del locale. Questi pattern linguistici ci aiutano a delineare una visione più completa e dettagliata dei diversi aspetti del locale che sono stati menzionati nelle recensioni.

Conclusioni

Attraverso tutte queste analisi, siamo stati in grado di scrutare a fondo i pensieri e le opinioni dei clienti riguardo al locale. I risultati ottenuti forniscono una preziosa prospettiva sull'esperienza dei clienti, consentendoci di comprenderne meglio le dinamiche e i fattori che influenzano la loro soddisfazione complessiva.

Nel complesso, i risultati ottenuti evidenziano una buona coerenza tra le aspettative e l'esperienza effettiva dei clienti presso lo Scholars. La predominanza di parole positive nei testi e nei titoli delle recensioni con un alto punteggio di stelle sottolinea la qualità eccellente offerta dal locale. Tuttavia, ciò non significa che non vi siano spazi per miglioramenti. Le parole negative rilevate nelle recensioni rappresentano importanti spunti per indirizzare interventi mirati al fine di ottimizzare l'esperienza complessiva dei clienti.