TripAdvisor Review Classification

FABIO CIMMINO 807070
ROBERTO LOTTERIO 807500
SAMUELE VENTURA 793060

Introduzione e obiettivi del progetto

- Dallo studio statistico effettuato da PhoCusWright [1] è emerso che il 77% delle persone non sono disposte a prenotare un hotel prima di aver letto le recensioni online su di esso.
- Spesso la valutazione complessiva di un hotel non è coerente con quanto riportato nella recensione. La causa di questo effetto è la soggettività intrinseca nel giudizio di ogni valutatore nei confronti dell'hotel.
- Risulta quindi evidente che chiunque voglia informarsi su una struttura alberghiera sia in difficoltà a capire quale sia l'effettiva qualità dell'albergo.

[1] https://www.phocuswright.com/About-Us

Introduzione e obiettivi del progetto

- Abbiamo quindi deciso di sviluppare un modello basato su una rete bayesiana che consente di classificare una recensione sulla base del commento e dei metadati inseriti dagli utenti di TripAdvisor.
- Al fine di dare una dimostrazione pratica della BN è stata sviluppata una applicazione che consente di:
  - Assegnare una valutazione ad una recensione inserita dall'utente con i relativi metadati
  - Visualizzare un istogramma ed un wordcloud delle parole più rilevanti associate ad un hotel

Descrizione del dataset iniziale

- Dopo aver effettuato il parsing dei file .dat, il dataset iniziale risulta composto da 13 attributi e 240 mila righe. Analizzando il dataset abbiamo riscontrato la presenza di features non rilevanti; le restanti risultano le seguenti:

Feature	Descrizione	Range
Value	Rapporto qualità/prezzo	0 - 5
Rooms	Qualità camere	0 - 5
Location	Qualità struttura alberighera	0 - 5
Cleanliness	Pulizia dell'hotel	0 - 5
Check-in	Accoglienza	0 - 5
Service	Qualità servizi hotel	0 - 5
Business	Qualità servizi business	0 - 5
Overall	Valutazione complessiva	0 - 5

Fase di preprocessing commento

- Questo processo consente di trasformare i dati grezzi (commenti) in un formato comprensibile per i modelli di Natural Language Processing. Sono stati effettuati i seguenti passi:

1. Tokenizzazione
2. Rimozione delle stopwords
3. Rimozione punteggiatura, numeri, caratteri non ASCII
4. Standardizzazione dei caratteri
5. Lemmatizzazione

Selezione dei termini per la rete bayesiana

- La selezione dei termini è stata effettuata utilizzando la funzione TF-IDF (Term Frequency – Inverse Document Frequency), che risulta il prodotto tra:

$$TF = \frac{N° \; di \; volte \; in \; cui \; la \; parola \; appare \; nel \; documento}{N° \; di \; parole \; totali \; nel \; documento}$$

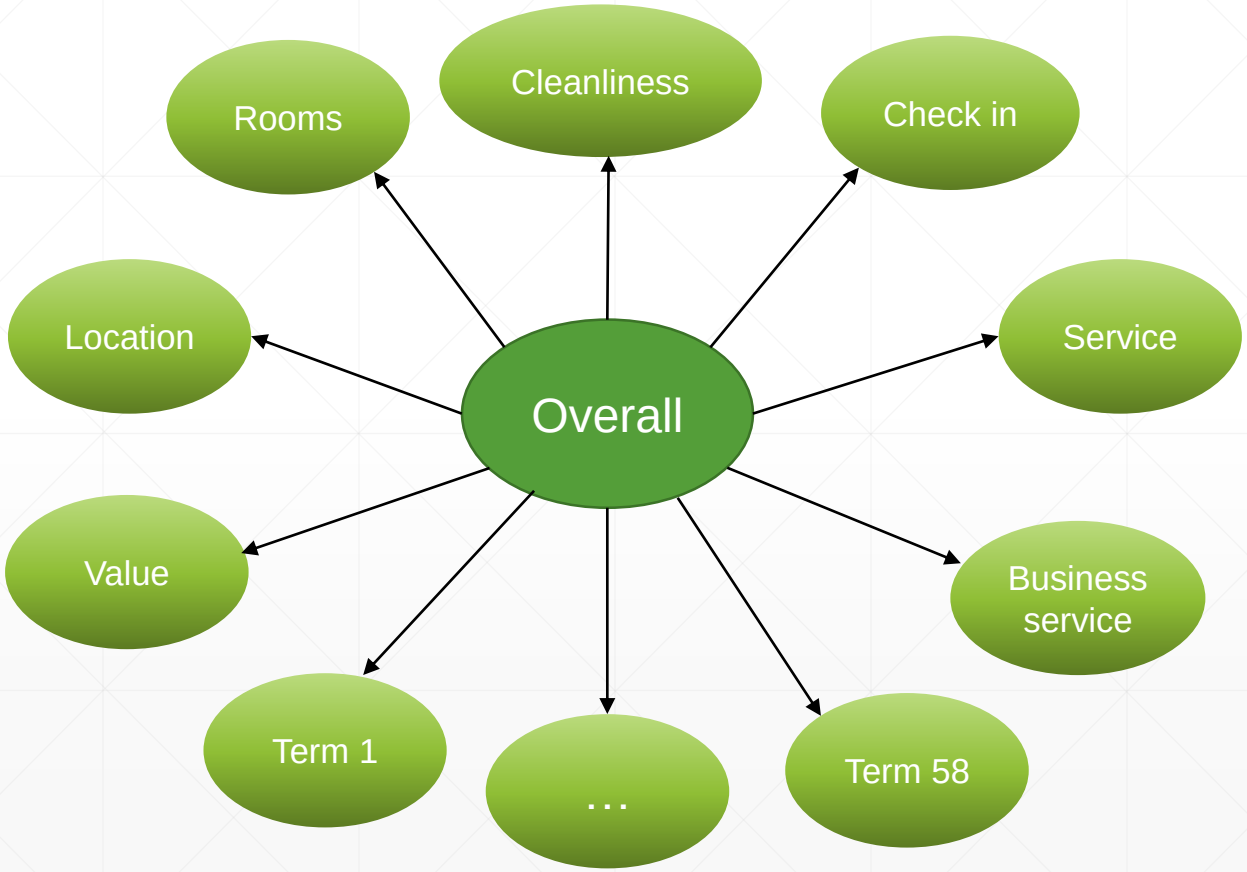
$$IDF = log_{10} \frac{N° \; di \; docuenti}{N° \; di \; documenti \; in \; cui \; la \; parola \; appare}$$

- In questo modo viene misurata l'importanza di un termine rispetto ad un documento. L'idea alla base è quella di dare più importanza ai termini che compaiono nel documento ma che in generale sono poco frequenti.

Selezione dei termini per la rete bayesiana

- E’ stato utilizzato il dataset *Affin contenente circa 2500 termini con uno score compreso tra -5 e 5 indicante il grado di polarità della parola.*
- Abbiamo quindi rimosso da ogni commento tutti i termini non presenti nel dataset o con score in valore assoluto minore di 3, perché considerati neutri.
- Quindi come nodi della rete bayesiana sono stati inseriti i primi 58 termini con TF-IDF più alta.

Modello creato



- In questo modello la presenza o l'assenza di una particolare feature in una recensione non è correlata alla presenza o assenza di altre features.
- I nodi relativi ai termini possono assumere valore 0 o 1, 1 quando il termine è presente nel commento, 0 quando non è presente.

Esperimenti e valutazione dei risultati

- Il dataset è stato diviso in Training e Test set (rispettivamente 80% e 20%) ed è stata stimata la probabilità $P(Overall|metadati, termini)$ per ogni riga del Test set ottenendo un'accuracy del 63%.
- Il valore non elevato dell'accuracy è dovuto alla presenza di valutazioni non coerenti all'interno del dataset. Perciò sono state rimosse tutte le tuple aventi una delle seguenti caratteristiche:
  - Il valore della variabile target è pari a 0
  - Tutti i metadati hanno valore -1 e l'Overall è maggiore di 2

Esperimenti e valutazione dei risultati

- E’ stato quindi condotto un nuovo esperimento ottenendo i seguenti risultati:

Accuracy	Precision	Recall	F-measure
68%	63%	68%	65%

- Abbiamo investigato ulteriormente sulla presenza di recensioni non coerenti eliminando dal dataset tutte le righe che presentavano 5 e più metadati con il valore -1. I risultati sono i seguenti:

Accuracy	Precision	Recall	F-measure
70%	67%	67%	66%

Esperimenti e valutazione dei risultati

- Infine si è deciso di aumentare i termini utilizzati nel modello, passando da 58 termini a 74 con i seguenti risultati:

Accuracy	Precision	Recall	F-measure
70%	67%	67%	66%

- Come si può notare queste performance si presentano in linea con il precedente modello. Questo è dovuto al fatto che aumentando il numero di termini vengono inclusi nel modello sempre più termini meno influenti.

Esperimenti e valutazione dei risultati

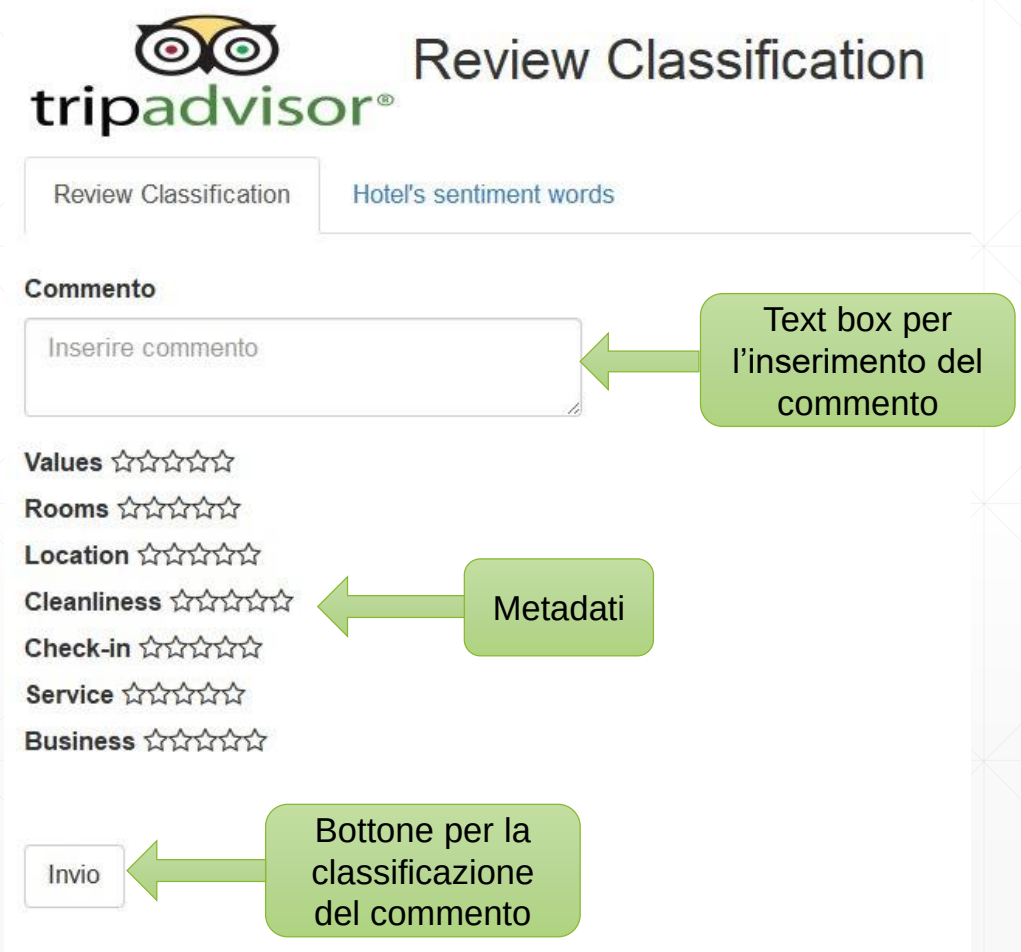
- Come ultimo esperimento è stato preso in considerazione il problema di Multi-label classification assegnando ad ogni istanza del Test set i due Overall più probabili. Le performance ottenute sono le seguenti:

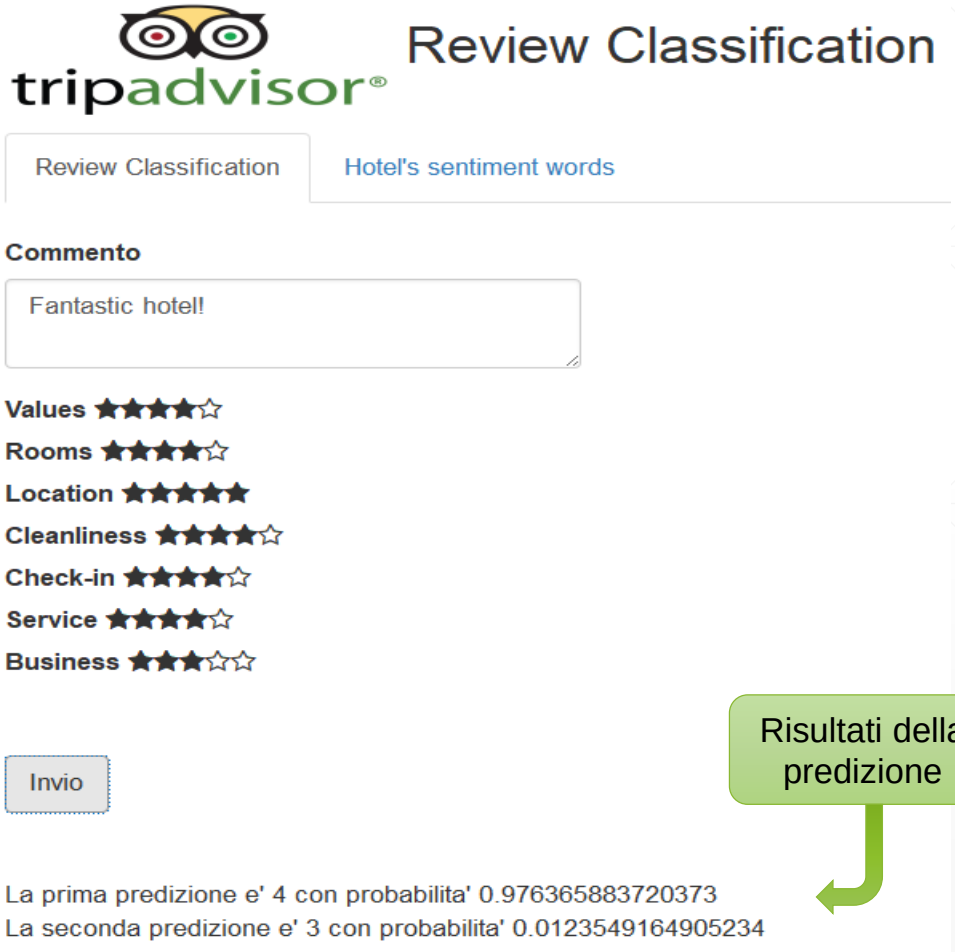
Accuracy	Precision	Recall	F-measure
94%	93%	94%	93%

Conclusioni

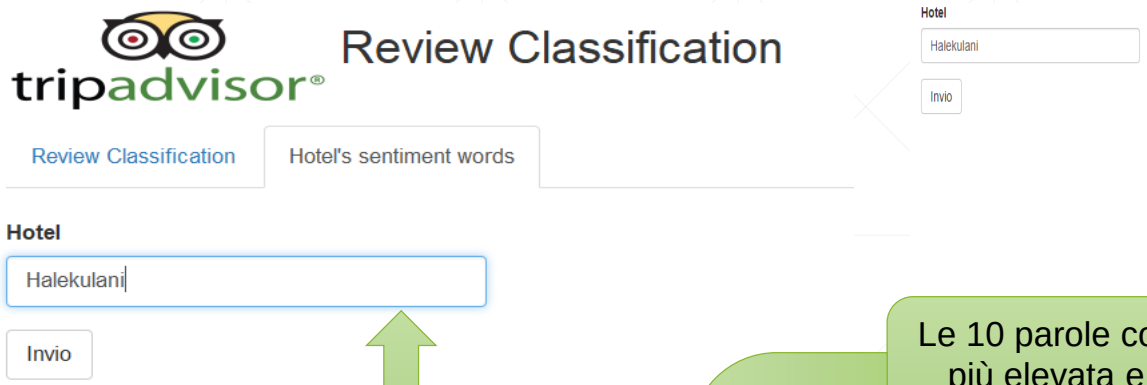
- A seguito dei vari esperimenti e analisi che sono stati condotti possiamo concludere che abbiamo ottenuto risultati discreti in rapporto alla qualità del dataset iniziale a nostra disposizione.
- Il contributo dei termini per la predizione dell'Overall resta limitato anche con l'aumentare dei termini scelti.
- Data la soggettività delle recensioni risulta difficile stabilire in modo univoco una sola classe di appartenenza. Per questo è stato considerato il problema di Multi-label classification con un incremento notevole dell'accuracy.
- Al fine di migliorare il peso dei commenti sul modello si potrebbe pensare di selezionare non token formati da n-grammi e non da unigrammi.

Demo – Sezione 1





Demo – Sezione 2



Text box per inserire il nome dell'hotel

Le 10 parole con TF-IDF più elevata e relativa frequenza

Wordcloud relativo all'hotel inserito