DOI: 10.26314/GARR-Conf18-proceedings-21

Customer Satisfaction from Booking

Maurizio Romano, Luca Frigau, Giulia Contu, Francesco Mola, Claudio Conversano

Università di Cagliari, Dipartimento di Science Economiche ed Aziendali

Abstract. Nel presente studio, analizzando i commenti dei clienti presenti su Booking, si è creato un modello affidabile al 91% che: a) supporta le strutture ricettive nell'individuazione di pregi e difetti, identificando i fattori su cui operare per ottenere un miglioramento del servizio offerto; b) consente d'identificare i punti di forza e debolezza della destinazione in cui opera la struttura ricettiva; c) può essere applicato in considerazione di differenti ambiti territoriali; d) consente di prevedere lo score di una recensione ed i relativi punti di forza/debolezza. Tale modello sfrutta l'affidabilità, la capacità e il filtraggio della Rete per effettuare le operazioni d'acquisizione periodica senza restrizioni e per rendere fruibile l'applicazione con il correlato Data Warehouse.

Keywords. Customer, Sentiment, Booking, WoM, Bayes

1. Introduzione

Questo studio è stato realizzato nell'ambito del progetto P.I.A. "Realizzazione di una piattaforma ICT a supporto del settore turistico" (RAS, 2007/2013), il cui obbiettivo è sviluppare una piattaforma ICT che consenta di valutare la Customer Satisfaction per il settore turistico sardo.

Sono state analizzate le recensioni dei clienti per ciascuna struttura ricettiva operante in Sardegna e presente su Booking. Considerando che nel settore turistico si ricorre all'applicazione automatica della sentiment analysis sia tramite machine learning che tramite metodi basati sul lessico per prevedere quando una recensione sia positiva o negativa (Schmunk et al., 2013), si è proceduto implementando un programma Python che estraesse tutte le informazioni utili fornite da Booking relative alle strutture d'interesse. Successivamente, le recensioni sono elaborate in linguaggio naturale al fine di comprendere per quali aspetti il cliente sia soddisfatto o meno e per fornire uno strumento di benchmarking delle strutture ricettive.

2. Obiettivi

La scelta della piattaforma Booking è dipesa dal fatto che le recensioni in esso presenti implicano che il recensore abbia soggiornato in una struttura. Booking utilizza una valutazione in scala [2.5; 10] e suddivide ogni recensione in due parti: un commento positivo ed uno negativo.

Gli scopi dell'analisi sono quindi molteplici:

- creare un classificatore che, dato un commento, lo classifichi in negativo/positivo;
- misurare quantitativamente l'incidenza (negativa o positiva) di una data parola all'interno del commento;

Cagliari, 3-5 ottobre 2018 - ISBN: 978-88-905077-8-6

- sviluppare un modello di previsione per lo Score di Booking in funzione delle recensioni;
- produrre uno strumento di benchmarking.

Per raggiungere gli scopi prefissati si è reso necessario attuare due fasi: 1) data collection: i dati sono stati estrapolati da Booking e trattati in preparazione dell'analisi statistica; 2) analisi delle recensioni: è stato implementato un classificatore Naive Bayes* ad-hoc e un modello di previsione dello score di una recensione su Booking in funzione delle parole in essa contenute.

3. Data Collection

E' stato implementato un estrattore Python che, mediante web scraping, ha estrapolato tutte le possibili informazioni d'interesse offerte pubblicamente da Booking e le ha inserite in tabelle create ad hoc.

Fig. 1
Esempio di una pagina di Booking contenente i giudizi dei clienti

I dati di Booking riguardano due tipologie d'informazione (Fig. 1):

- le 619 strutture alberghiere operanti in Sardegna;
- le 66237 recensioni per ciascuna struttura, in italiano o in inglese, scisse nei due commenti (positivi e negativi) che le compongono, per un totale di 106800 commenti.

I dati delle strutture alberghiere sono organizzati in una tabella (Fig. 2) ove ciascuna riga rappresenta una struttura e le cui colonne sono il risultato dell'accorpamento delle informazioni reperibili sulle stesse su Booking (tipologia di struttura, CAP, comune, valutazioni sulla pulizia, comfort, ecc.)

Nome Struttura	Tipologia	Cap	Comune/Località	...
Struttura 1	Extralberghiere	09044	Sant' Isidoro	...
Struttura 2	3 Stelle	09049	Villasimius	...
Struttura 3	3 Stelle	07013	Mores	...
Struttura 4	4 Stelle	09123	Cagliari	...
...	...	...	...	...

Fig. 2
Estratto della struttura tabellare contenente i dati delle strutture alberghiere

I dati di ciascuna recensione successivamente vengono organizzati in un'altra tabella (Fig. 3) contenente, oltre al commento, le informazioni utili per risalire alla struttura e alla recensione originale, nonché i dati sul recensore e sulla tipologia di cliente (viaggio di lavoro, viaggio di piacere ecc.).

Nome Struttura	ID Commento	ID Review	Commento	Neg-Pos	Score	...
Struttura 1	1	1	christina was the best...	Pos	10.0	...
Struttura 1	2	2	we booked an appartment...	Pos	10.0	...
Struttura 1	3	3	we travelled into cagliari...	Pos	9.2	...
Struttura 1	4	4	it was fantastic	Pos	10.0	...
...	...	...	...	...	...	...

Fig. 3
Estratto della struttura tabellare contenente i dati dei commenti/recensioni

4. Riorganizzazione e Data Cleaning

I dati sono stati riorganizzati raggruppando le parole con il medesimo significato e ripulendoli da congiunzioni, punteggiatura, numeri e altre stopwords.
Affinché sia possibile trattare nello stesso modo parole dal medesimo significato e, in funzione di questo, calcolare opportunamente le frequenze di ciascuna di esse, si è reso necessario procedere a un accorpamento. Il modo più semplice per implementarlo consiste nel sostituire nei commenti originali ogni parola accorpabile con una "più comune" per significato.
Successivamente, tutte le parole di ogni commento sono state estrapolate e raggruppate in un insieme unico (Bag Of Words – BOW, Fig. 4). Questo raggruppamento ha consentito di calcolare le frequenze per ciascuna parola.

Fig. 4
Schema riassuntivo della logica di produzione del BOW

L'analisi delle frequenze delle parole ha consentito di definire delle macro-categorie che andassero a raggrupparle (Fig. 5). Nello specifico, sono state individuate le seguenti categorie: bar, cleaning, comfort, food, hotel, position, pricequalityrate, room, services, sleepquality, staff, wifi, other.

La categorizzazione è stata utilizzata esclusivamente per trattare i dati in maniera aggregata. Ciò consente di analizzare i pregi e i difetti relativi ai servizi offerti per le singole categorie. Il classificatore Naive Bayes* non le sfrutta poiché un tale accorpamento ridur-

rebbe sensibilmente la variabilità determinando un peggioramento generale della capacità predittiva del modello. Tuttavia, l'accorpamento viene utilizzato per sostituire temporaneamente le parole presenti in un commento con la categoria rilevata, al fine di comprendere in maniera immediata a quali categorie lo stesso appartenga o meno.

Fig. 5
Estratto della struttura tabellare a supporto per la categorizzazione

Word	Category
Colazione	Food
Ristorante	Food
Bread	Food
Cakes	Food
Mangiare	Food
Conto	Price-quality rate
Caro	Price-quality rate
Pagamento	Price-quality rate
Pay	Price-quality rate
Gestore	Staff
Stintino	Position
Orosei	Position
…	…

5. Analisi delle recensioni

L'analisi delle recensioni consiste nel creare un classificatore in grado di prevedere accuratamente se un dato commento sia positivo o negativo in funzione delle parole che lo compongono. Sono stati applicati i classificatori più comuni suggeriti dalla letteratura e, osservando la qualità dei risultati da loro prodotti (Fig. 8), il modello Naive Bayes* è risultato il migliore in base alla capacità di generalizzazione.
Successivamente, considerando l'unione del commento positivo e negativo di ogni recensione, si è applicato il classificatore anche alle recensioni nel loro insieme, e quindi in contesti in cui la separazione fra aspetti positivi e negativi non è netta.

*5.1 Il modello Naiva Bayes**

Sfruttando un approccio frequentista, si calcola la probabilità che un commento sia negativo in rapporto alla probabilità che esso non lo sia (Rischio Relativo o Likelihood Ratio, LR). Il teorema di Bayes semplifica i calcoli e consente di stimare il LR in funzione di numero e tipologia di parola, o del commento stesso. Si descrive di seguito formalmente il modello.
Sia W l' insieme di tutte le parole comparse nelle recensioni.
Siano $C_{pos} \subseteq W$, $C_{neg} \subseteq W$ due commenti.
Sia $R = C_{pos} \cup C_{neg}$ una recensione.
Il Likelihood Ratio è dato da con $LR_C = \frac{P(neg|C)}{P(pos|C)}$ $con\ C \subseteq W$.

Per il teorema di Bayes, $\log(LR_C) \cong pres_C + abs_C + \log\left(\frac{P(neg)}{P(pos)}\right)$.
in cui:

$$pres_C = \sum_{w \in C} \log\left(\frac{P(w|neg)}{P(w|pos)}\right)$$

$$abs_C = \sum_{w \in W \setminus C} \log\left(\frac{P(\bar{w}|neg)}{P(\bar{w}|pos)}\right)$$

$pres_C$ e abs_C rappresentano, rispettivamente, le intensità delle parole presenti e non presenti nel commento.

Il *LR* può essere calcolato sia per un commento, sia per una recensione che per una singola parola od un insieme ben preciso delle stesse (ad es. una categoria). Ogni parola ha dunque un valore di *log(LR)* nel caso in cui essa sia presente nel commento ed un altro valore di *log(LR)* in cui non sia presente. Il *log(LR)* del commento sarà dato dalla somma di $pres_C$ e abs_C.

Tale valore viene calcolato per ogni singola parola presente nel BOW. I risultati vengono inseriti in una tabella (rappresentata in Fig. 6): nelle colonne sono indicate le parole e, nelle righe, i valori di *log(LR)* nel caso in cui la parola sia presente o assente e la proporzione di commenti negativi/positivi che contengono la parola stessa.

	"Assolutamente"	*"Mare"*	...
P(neg)	0,011	0,026	...
P(pos)	0,007	0,075	...
Log(LR) Parola presente	0,411	-1,077	...
Log(LR) Parola assente	-0,004	0,052	...

Fig. 6
Estratto della tabella in output dopo il calcolo dei log(LR)

Ad ogni commento (positivo o negativo) è dunque assegnato un valore di *log(LR)*. Si ricorre successivamente alla Cross Validation per stimare il valore soglia (τ) tale che:

se *log(LR)* > τ → il commento è classificato come "Negativo",

se *log(LR)* ≤ τ → il commento è classificato come "Positivo".

La scelta di τ, così come si osserva in Fig. 7, è stata effettuata minimizzando l'indice di errata classificazione (Misclassification Error) per entrambi gli errori che il classificatore Naive Bayes* può compiere al variare dei valori assumibili da τ.

Il valore stimato di τ è: τ=1,138.

La somma dei *log(LR)* dei commenti positivi e negativi di una recensione indica, in funzione

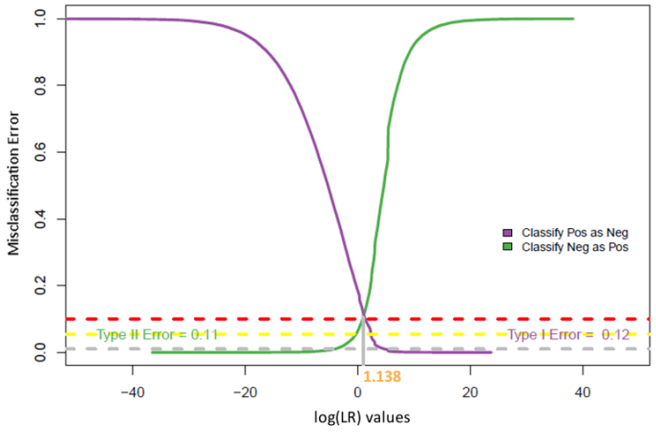


Fig. 7
Le due tipologie di errore verificabili nella classificazione dei commenti e gli andamenti del valore soglia τ

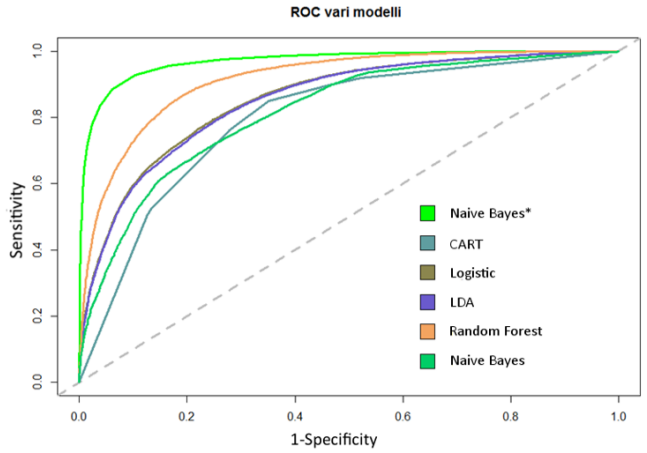


Fig. 8
Curve ROC del classificatore Naive Bayes* e degli altri classificatori utilizzati nell'analisi

del τ, quale sia l'andamento della stessa, ovvero se la recensione è da considerarsi, nel suo insieme, come negativa/positiva e di quanto questa lo sia (a seconda della distanza da τ).

In Fig. 8 e nella Tab. 1 è possibile confrontare le performance del modello Naive Bayes* e degli altri classificatori. Il modello qui proposto performa meglio rispetto a ciascuna misura di benchmarking.

Tab. 1
Tabella delle più comuni misure di benchmarking per i vari classificatori

Model	Misclassification Error	Accuracy	Sensitivity	Fall-out	F1 score	Matthews Correlation Coefficient
Naive Bayes*	**0,089**	**0,911**	**0,929**	**0,117**	**0,926**	**0,813**
Logistic	0,150	0,850	0,884	0,532	0,877	0,361
Random Forest	0,189	0,811	0,873	0,591	0,849	0,303
Naive Bayes (e1071)	0,194	0,806	0,804	0,389	0,834	0,390
Naive Bayes (klaR)	0,194	0,806	0,804	0,389	0,834	0,390
CART	0,232	0,768	0,842	0,587	0,815	0,272
LDA	0,236	0,764	0,860	0,641	0,816	0,246

5.2 Previsione score di Booking

Al fine di produrre una previsione dello score di una recensione su Booking in funzione delle parole in essa contenute, sono stati addestrati diversi modelli. Il miglior modello di previsione dello score di una recensione è risultato essere il random forest (errore quadratico medio: 0,6704). Come predittori sono state utilizzate le seguenti variabili create mediante l'uso di Naive Bayes*:

- review_position, review_bar, review_cleaning, ..., review_services, che rappresentano i valori di *log(LR)* per ciascuna delle 13 categorie definite;
- polarità, ossia la classificazione in pos/neg del *log(LR)* generale della recensione tramite τ.

6. Conclusioni

Il presente lavoro ha consentito di definire un modello (affidabile al 91%) che, sulla base

del *LR* calcolato sulle recensioni dei clienti su Booking, è capace di:

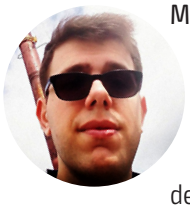
- supportare le strutture ricettive nell'individuazione dei propri punti di forza e debolezza identificando gli aspetti sui quali operare per ottenere un miglioramento del servizio offerto;
- consentire l'identificazione di pregi e difetti della destinazione in cui opera la struttura ricettiva;
- essere applicato in differenti ambiti territoriali;
- prevedere lo score di una recensione su Booking e, per ciascuna classe definita, i punti di forza/debolezza. In questo modo è possibile definire gli aspetti su cui intervenire per migliorare il risultato ottenuto.

Cardine del Progetto è l'impiego della rete scientifica GARR in quanto consente di effettuare sia le operazioni di acquisizione periodica senza restrizioni, sia di rendere fruibile l'applicazione del modello secondo la modalità nativa as-a-Service per il relativo Data Warehouse (DW) offrendo al contempo ampi spazi di sviluppo in settori scientifici limitrofi.

Riferimenti bibliografici

Schmunk Sergej, Wolfram Höpken, Matthias Fuchs, and Maria Lexhagen. (2013). Sentiment Analysis: Extracting Decision Relevant Knowledge from UGC. In Information and Communication Technologies in Tourism 2014, Zheng Xiang and Iis Tussyadiah, Cham.

Autori



Maurizio Romano - romano.maurizio@unica.it

Maurizio Romano è dottorando in Scienze Economiche ed Aziendali presso l'Università degli Studi di Cagliari. Laureatosi in Informatica con una tesi afferente al Semantic Web, ha vinto una borsa di ricerca per la "Creazione di algoritmi per l'analisi di dati del settore turistisco". Tutor didattico nel corso "Metodi di Apprendimento Statistico per il Data Science" della Laurea Magistrale in Data Science, Business Analytics e Innovazione, incentra la sua ricerca su Statistical Learning e Big Data.

Luca Frigau - frigau@unica.it

Luca Frigau è ricercatore presso il dipartimento di Scienze Economiche ed Aziendali dell'Università degli Studi di Cagliari, dove insegna Statistica e Analisi di Mercato. Ha conseguito il dottorato di ricerca in Probability and Mathematical Statistics presso la Charles University (Praga) e i suoi principali interessi di ricerca sono il Machine Learning, il Classification assessment, il Data Mining e il Data Science.

Giulia Contu - giulia.contu@unica.it

Giulia Contu è un dottore di ricerca in "Scienze del turismo: metodi, modelli e politiche", Università di Palermo, Italia), è attualmente dottoranda presso il Dipartimento di Scienze economiche e Aziendali dell'Università degli Studi di Cagliari. I suoi interessi includono la statistica del turismo, il revenue management, il turismo sommerso e il fenomeno di Airbnb. È membro della Società Italiana di Scienze del Turismo (SISTUR).

Francesco Mola - mola@unica.it

Francesco Mola è professore ordinario di Statistica Metodologica, si occupa di Statistica Computazionale e Data Science. È Prorettore Vicario dell'Università degli Studi di Cagliari, dove tiene corsi di Statistica e Data Science ed è stato Direttore del Dipartimento di Scienze Economiche ed Aziendali. I principali temi di ricerca riguardano la statistica multivariata, computazionale e il Data Mining. È stato membro del board dell'International Association for Statistical Computing.

Claudio Conversano - conversa@unica.it

Claudio Conversano è professore associato di Statistica metodologica e docente di Statistica, Modelli statistici per l'asset allocation, Analisi di big data e Quantitative Methods for Management presso l'Università degli Studi di Cagliari. I suoi principali interessi di ricerca riguardano i metodi di apprendimento statistico (statistical learning), i Big Data, la finanza computazionale, l'inferenza causale e la qualità dei dati statistici.