
# Sentiment Analysis on Food Reviews


Pietro Colombo 793679

Marco Fagioli 808176


Progetto Data Analytics 2018-1019

## Obiettivi del progetto

- **Sentiment Analysis** delle reviews, correlazione tra la polarità del testo delle recensioni e le stelle votate per la review.
- **Classificazione** positività o negatività delle review tramite machine learning supervisionato sul testo.
- **Aspect Based Sentiment Analysis** per ogni prodotto con individuazione dei topic e polarità associata.

## Dataset

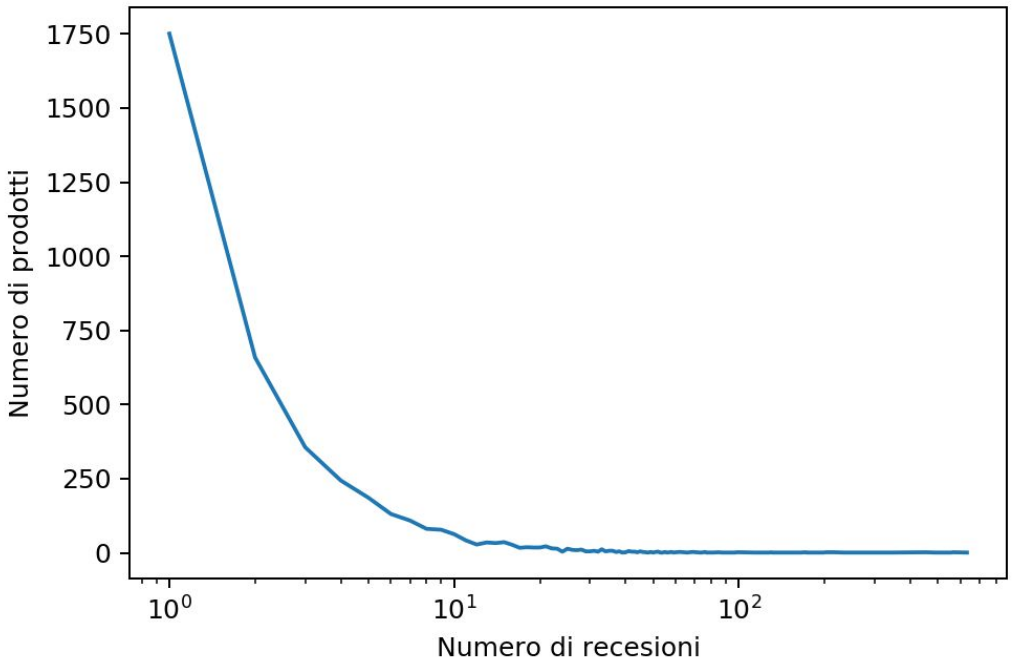
Il dataset è costituito da 35.172 differenti reviews su prodotti alimentari provenienti da Amazon. Le recensioni riguardano un elevato numero di prodotti e sono effettuate da molteplici utenti, le feature presenti all'interno del dataset sono:

- **productId**: identifica in modo univoco il prodotto alimentare e può essere usato per accedere al prodotto (https://www.amazon.com/dp/**productId**).
- **userid**: identifica univocamente l'utente che ha effettuato la review.
- **score**: esprime il punteggio dato dal recensore al prodotto ed è una votazione che va da 1 stella, il minimo, fino a 5 stelle il massimo.
- **text**: contiene il testo in forma integrale della recensione.

| | productid | userid | score | text |
|---|---|---|---|---|
| 0 | B001E4KFG0 | A3SGXH7AUHU8GW | 5.0 | I have bought several of the Vitality canned d... |
| 1 | B00813GRG4 | A1D87F6ZCVE5NK | 1.0 | Product arrived labeled as Jumbo Salted Peanut... |
| 2 | B000LQOCH0 | ABXLMWJIXXAIN | 4.0 | This is a confection that has been around a fe... |
| 3 | B000UA0QIQ | A395BORC6FGVXV | 2.0 | If you are looking for the secret ingredient i... |
| 4 | B006K2ZZ7K | A1UQRSCLF8GW1T | 5.0 | Great taffy at a great price. There was a wid... |

## Esplorazione Dataset

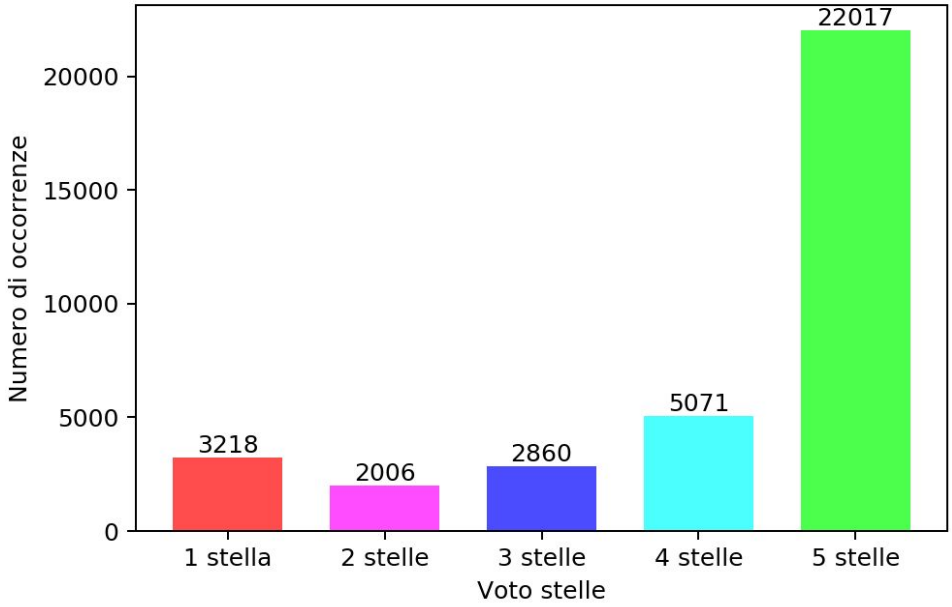
La prima operazione eseguita per analizzare il dataset è stata quella di andare a vedere come sono distribuite le recensioni sui prodotti.




Dal grafico è evidente che esista un **elevato** numero di **prodotti** (4224) all'interno del dataset ma di cui la maggior parte possiede solamente poche recensioni. Risulta infatti che:

- **circa il 70%** dei prodotti alimentari possiede 5 o meno recensioni;
- circa il 85% dei prodotti alimentari possiede 10 o meno recensioni;
- circa il 95% dei prodotti alimentari possiede 25 o meno recensioni;
- solamente poco più dell'1%  dei prodotti ha più di 100 recensioni.

L’analisi si è poi spostata sulle votazioni e stelle assegnate per ciascuna review. Da una prima occhiata è facile notare come la **distribuzione** dei voti assegnati alle recensioni all’interno sia molto **sbilanciata**.



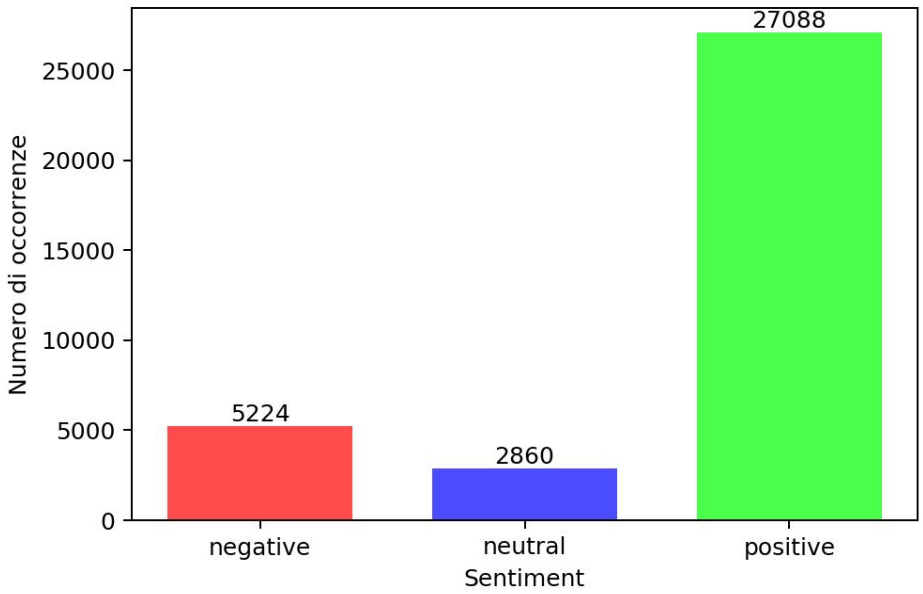

Le votazioni con 4 e 5 stelle rappresentano rispettivamente circa il 14,5% e 62,5% mentre per 1, 2 e 3 stelle la percentuale non supera nemmeno il 10%.

## Data preprocessing

Per poter utilizzare al meglio le informazioni a nostra disposizione è stato necessario un **preprocessing** dei dati stessi. In particolare abbiamo deciso di creare un’ulteriore campo per passare dal valore dei voti in stelle al valore di sentiment e la costruzione di un nuovo testo filtrato a partire dall’originale.

Il campo aggiunto, denominato ***sentiment***, è stato creato effettuando una conversione in:

- **positive**, se i voti avevano valore di 4 o 5 stelle;
- **neutral**, se i voti avevano valore di 3 stelle;
- **negative**, se i voti avevano valore di 1 o 2 stelle.



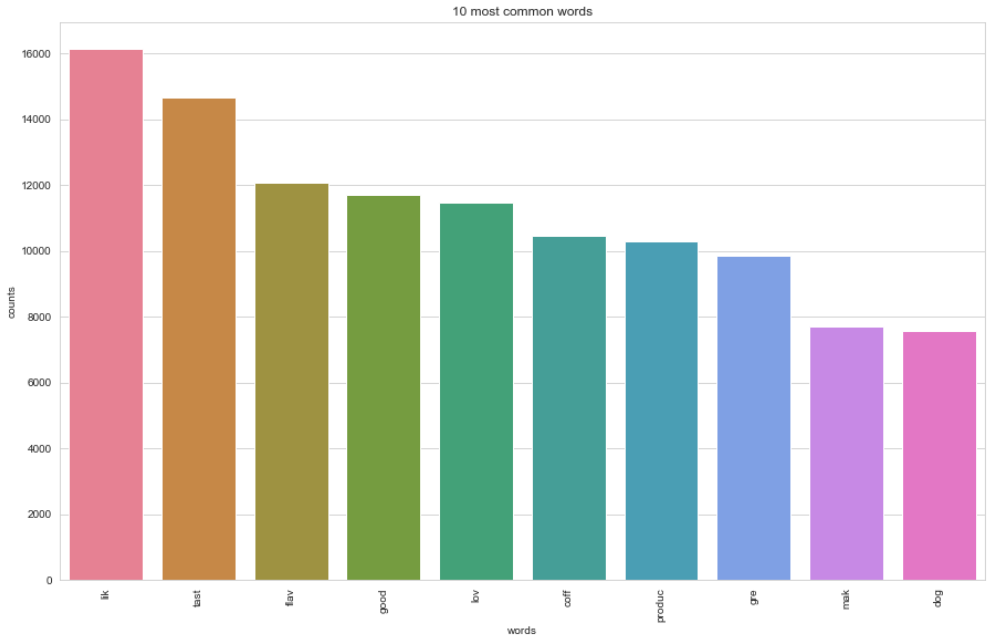

La distribuzione secondo della nuova feature è di circa 15% per negative, 10% per neutral e 75% per positive, rimanendo comunque sbilanciata ma in maniera inferiore e permette di avere una maggiore accuratezza delle predizioni.

Il nuovo campo testo filtrato, denominato ***clean_text***, è stato ottenuto dal testo originale su cui sono state fatte in sequenza le operazioni di:

- trasformazione di tutte le parole in minuscolo;
- rimozione dei tag HTML;

- rimozione degli stopwords, forniti dalla libreria *nltk stopword*;
- rimozione della punteggiatura;
- stemming, secondo il *LancasterStemmer* di *nltk*.

Dopo le modifiche effettuate al testo  è stato fatto un counting delle **parole più frequenti** ottenendo:



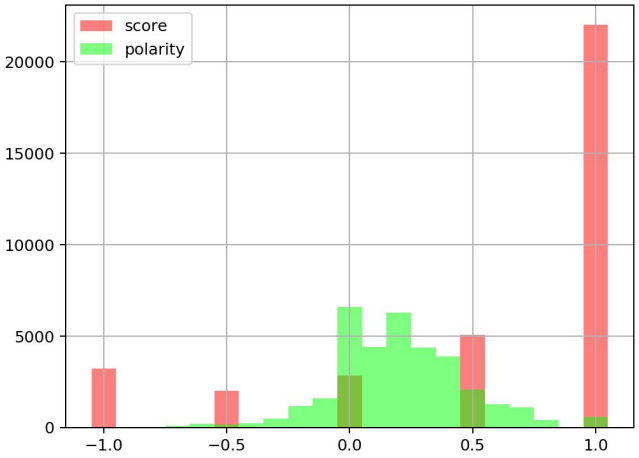

Il dataset finale ottenuto dopo tutte le modifiche, si presenta nella forma:

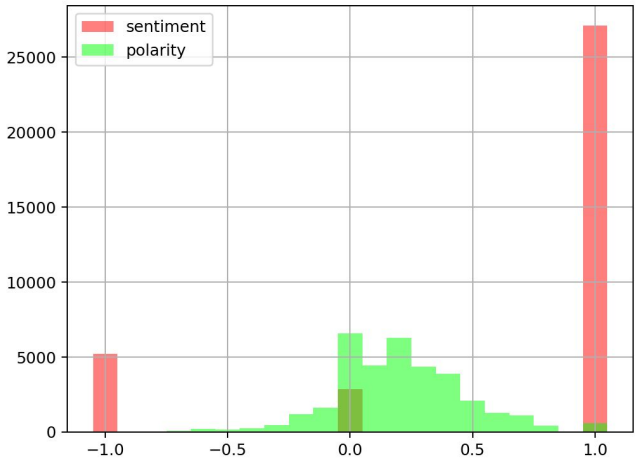
| | productid | userid | score | text | clean_text | sentiment |
|---|---|---|---|---|---|---|
| **0** | B001E4KFG0 | A3SGXH7AUHU8GW | 5 | i have bought sev... | bought several vitalit... | positive |
| **1** | B00813GRG4 | A1D87F6ZCVE5NK | 1 | product arrived labe... | product arrived labe... | negative |
| **2** | B000LQOCH0 | ABXLMWJIXXAIN | 4 | this is a confectio ... | confection around... | positive |
| **3** | B000UA0QIQ | A395BORC6FGVXV | 2 | if you are lookin ... | looking secret ing... | negative |
| **4** | B006K2ZZ7K | A1UQRSCLF8GW1T | 5 | great taffy at a g... | great taffy great pric... | positive |

## Sentiment Analysis con TextBlob

*TextBlob* è una semplice, comoda e intuitiva libreria Python che permette il processing dei dati testuali. Nel nostro caso è stato deciso di utilizzarla per calcolare la polarità di una frase/testo.

Per ogni reviews si è perciò deciso di calcolare la **polarità**, espressa tra **-1** (negativa) e **1** (positiva), per poi andarla a confrontare con i voti in stelle o il sentiment, alla ricerca di una eventuale correlazione.



Come evidenziano i grafici **non esiste una correlazione diretta** tra la polarità calcolata e le valutazioni disponibili.

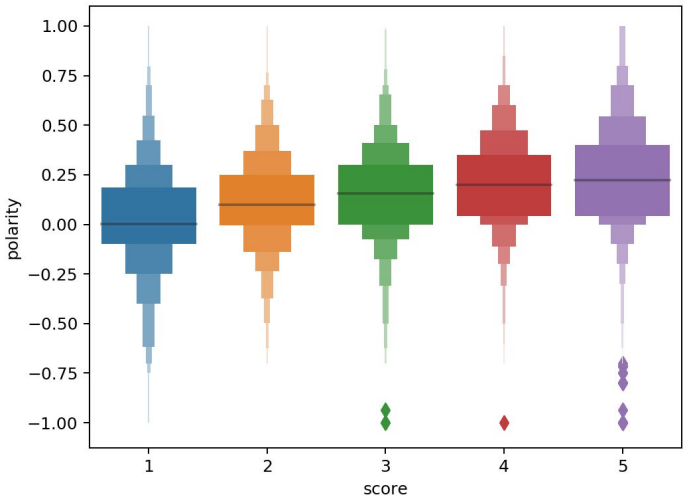
Un recensione che lo prova è la seguente:

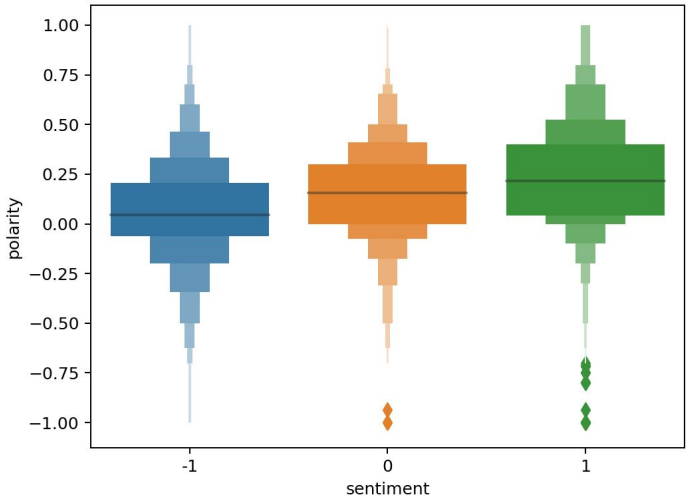
★☆☆☆☆ **Way TOO Salty!**
October 6, 2014
Flavor: Organic Salt & Pepper

I love salty snacks but these Salt and Pepper potato chips are so salty they're inedible. I got a variety pack and had to throw all the Salt and Pepper bags away, no one would eat them! The pepper seasoning was great, nice and spicy but too bad they ruined them with the salt.

nonostante la review sia molto bassa, 1 stella, e quindi negativa, la polarità calcolata risulta essere di 0.33, perciò associata ad un testo con maggioranza di parole con sentiment positivo.

Per confermare la mancanza di correlazione sono state infine guardate le **distribuzioni** delle **polarità** in base al numero di stelle votate e al sentiment.






Possiamo notare che tutte le polarità per ogni classe delle stelle o del sentiment si aggirano tra -0,25 e 0,5. Da questo concludiamo che indipendentemente dal voto in stelle dato dall'utente il testo non rispecchia questa valutazione. Perchè?

- La maggior parte delle parole della recensione non porta sentiment;
- TextBlob applica il sentiment solo a una ristretta classe di parole, principalmente aggettivi.

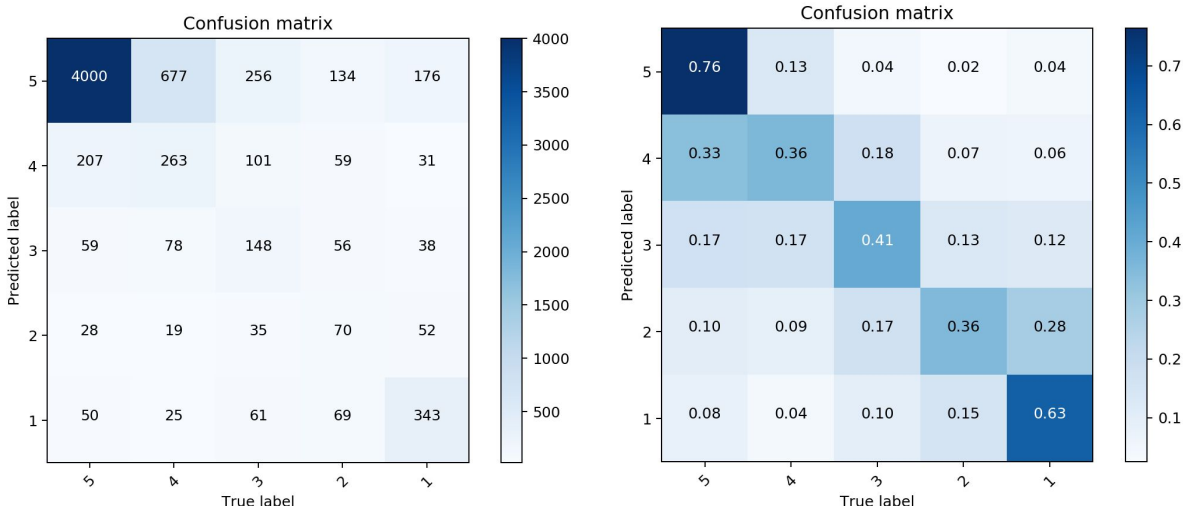
## Classificazione con sklearn

*Sklearn* è una libreria open source di apprendimento automatico per il linguaggio di programmazione Python. Contiene algoritmi di classificazione, regressione e clustering, progettata per operare con le librerie NumPy e SciPy.

Per ogni reviews si è deciso di effettuare la **previsione** per la positività o negatività tramite **machine learning supervisionato** sul testo. Per la classificazione sono stati utilizzati due differenti modelli *LogisticRegression* e *RandomForest*.

La prima operazione da effettuare è la divisione del dataset in 80% train e 20% test, realizzata mantenendo le classi bilanciate. Successivamente è necessario preparare l'input del training set per i classificatori, creando una matrice la quale rappresenta il vocabolario contenente tutte parole all'interno delle reviews e un intero che rappresenta il numero di occorrenze di ognuna di esse nelle recensioni.

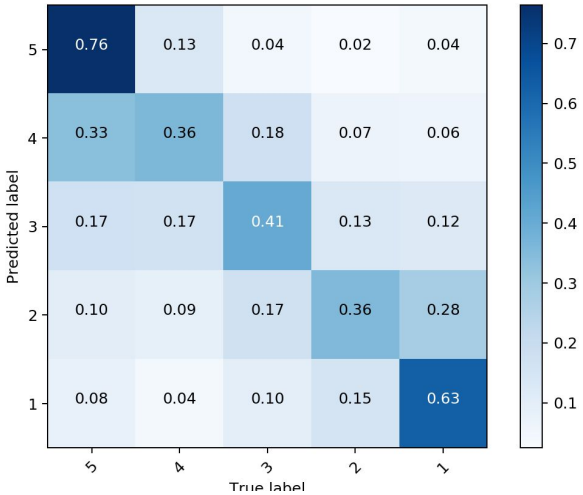
Viene poi applicato il modello scelto e valutata la sua efficienza in termini di accuracy e della matrice di confusione con le metriche associate.



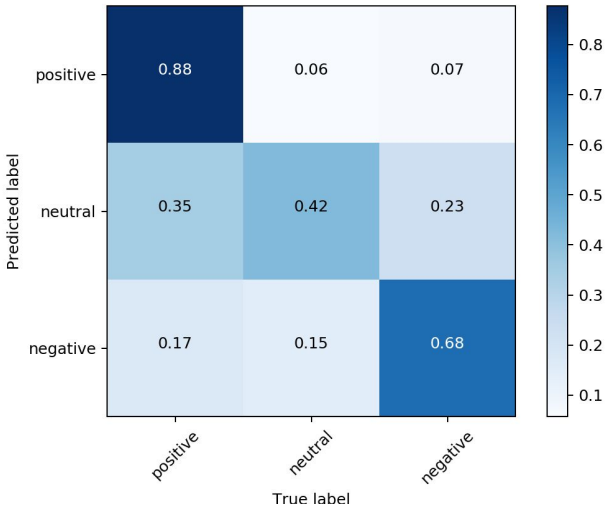

Per una maggiore chiarezza e lettura dei dati, si è deciso di stampare le matrici di confusione normalizzate secondo le righe, operazione che permette di visualizzare direttamente la precision per ogni classe.

## Analisi dati per LogisticRegression

Matrice di confusione per i voti in **stelle**:

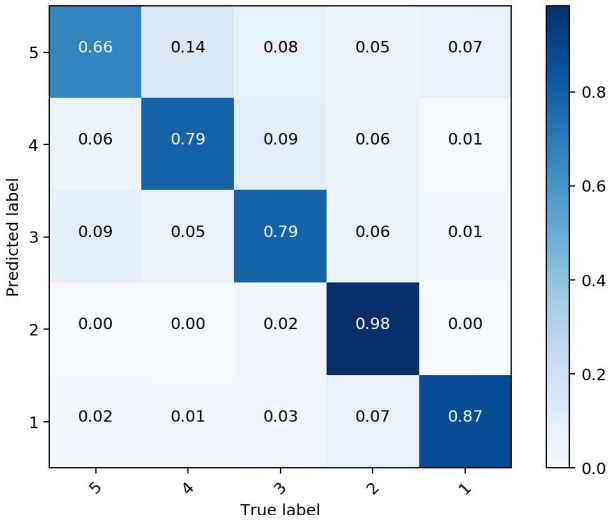



Matrice di confusione per il **sentiment**:

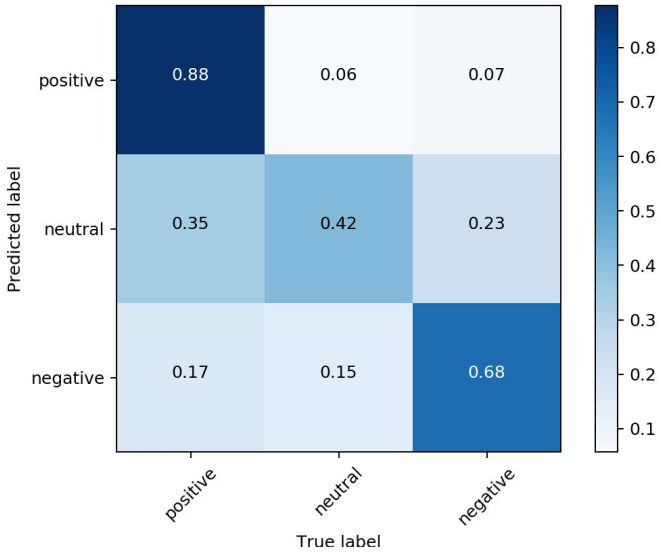


## Analisi dei dati per RandomForest

Matrice di confusione per i voti in **stelle**:




Matrice di confusione per il **sentiment**:



### LogisticRegression vs RandomForest

Abbiamo eseguito la previsione in successione con i due differenti modelli sul sentiment e confrontato i risultati ottenuti in una tabella di confronto.

| | **LogisticRegression** | | **RandomForest** | |
|---|---|---|---|---|
| **sentiment** | precision | recall | precision | recall |
| *positive* | 0.88 | 0.95 | 0.81 | 1.00 |
| *neutral* | 0.42 | 0.21 | 0.91 | 0.09 |
| *negative* | 0.68 | 0.57 | 0.90 | 0.28 |
| *valore medio* | **0.66** | **0.58** | **0.87** | **0.46** |

Da questo confronto concludiamo che generalmente:

- il modello *LogisticRegression* ha una precision più bassa ma recall più alta;
- il modello RandomForest viceversa.

## Aspect-Based SA con Gensim LDA (Latent Dirichlet Allocation)

*LDA (Latent Dirichlet Allocation)* è un modello di analisi e studio del linguaggio naturale che consente di capire la semantica del testo analizzando la similarità tra i termini presenti e di estrarre argomenti dall'insieme di parole del testo. Ciascun topic è caratterizzato da una particolare distribuzione di termini.

Avendo il dataset una grande eterogeneità tra i prodotti che lo costituiscono, è stato deciso di creare i modelli lavorando singolarmente su di essi e preferendo l'utilizzo di quelli che possiedono il maggior numero di recensioni.

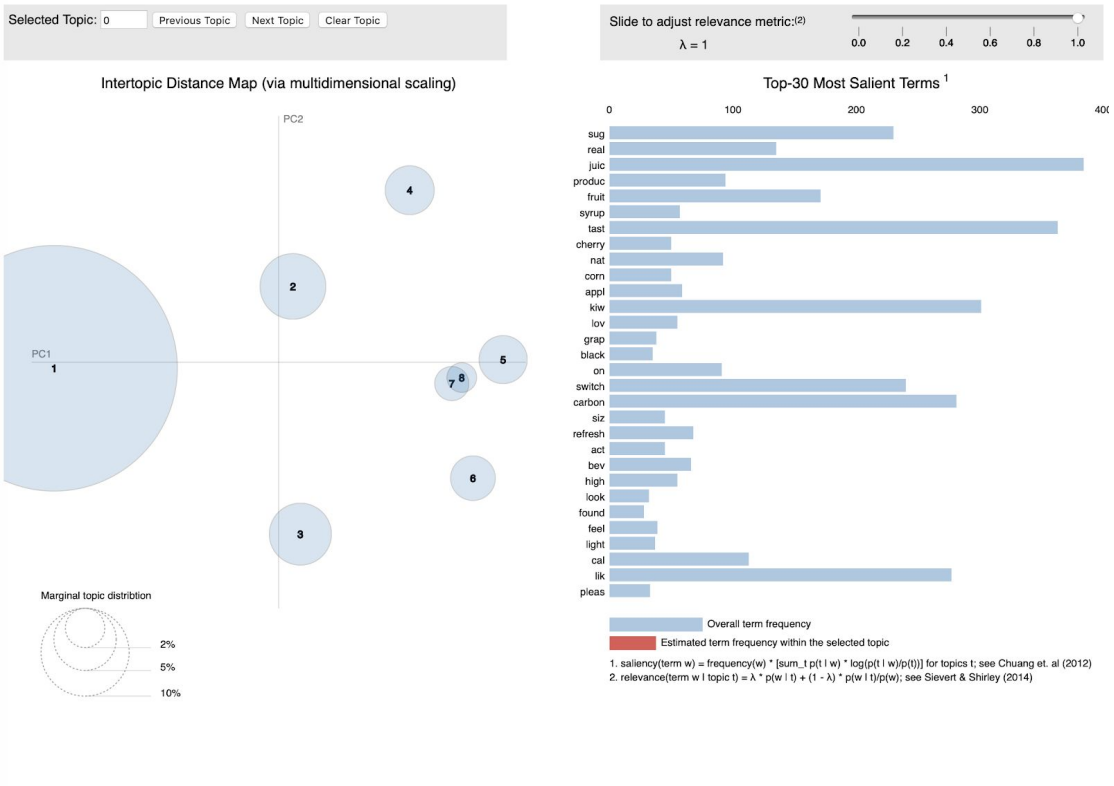
Per l'identificazione degli argomenti a partire da una recensione viene usata la libreria genism che implementa LDA.

## Scelta numero di Topic

Per l'estrazione ottimale dei topic è necessaria un'operazione preliminare in cui vengono generati diversi modelli con numero crescente, da 2 a 10, di topic impostato. Ognuno dei modelli creati viene temporaneamente salvato e per ciascuno di essi viene calcolato il valore della coerenza, per infine utilizzare il modello con questo valore ottimale.

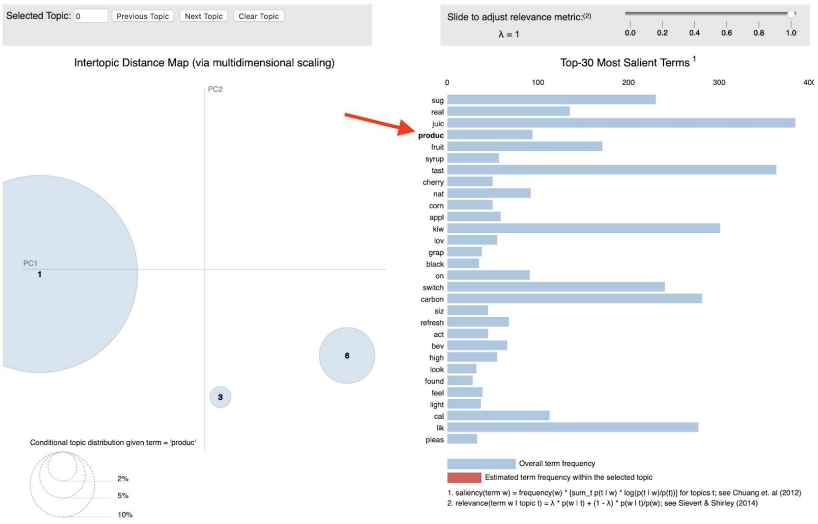
## Visualizzazione dei Topic

Il modello scelto viene visualizzato tramite il package *pyLDAvis* che crea una pagina web interattiva per vedere i differenti topic estratti.



Posizionando il cursore su di un argomento e selezionandolo si possono vedere le frequenze delle diverse parole, con maggior occorrenze, del suddetto topic.

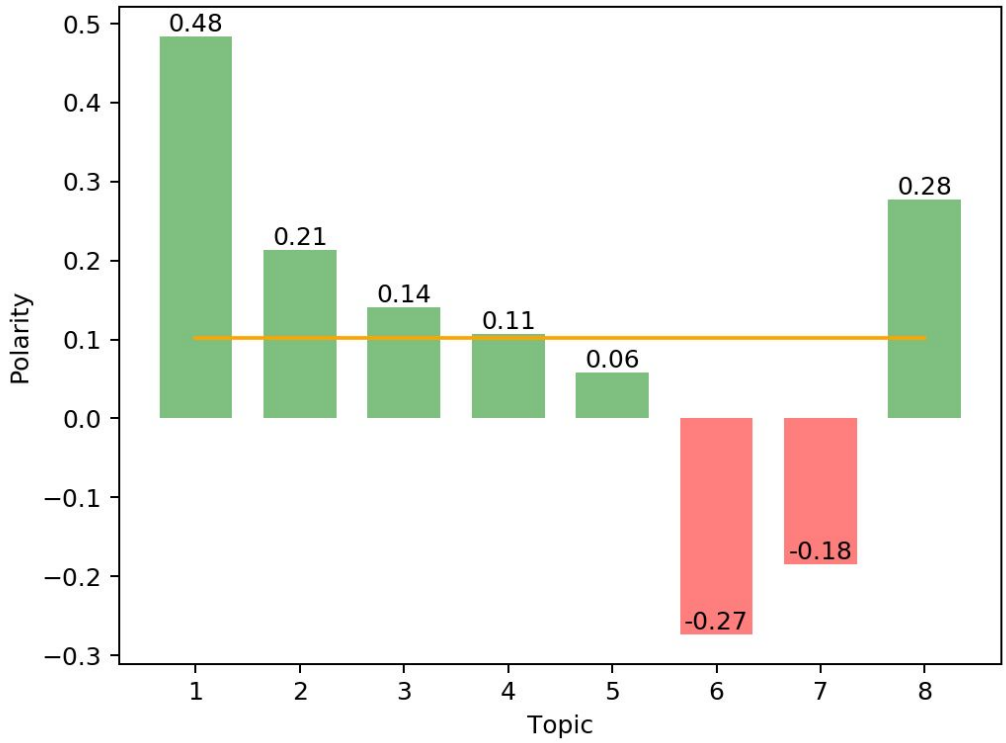

È possibile selezionare una parola nella parte destra dell'interfaccia per visualizzare i topic in cui essa è presente. Questa operazione può essere effettuata sia con un argomento selezionato che non.



## Polarità per ogni topic

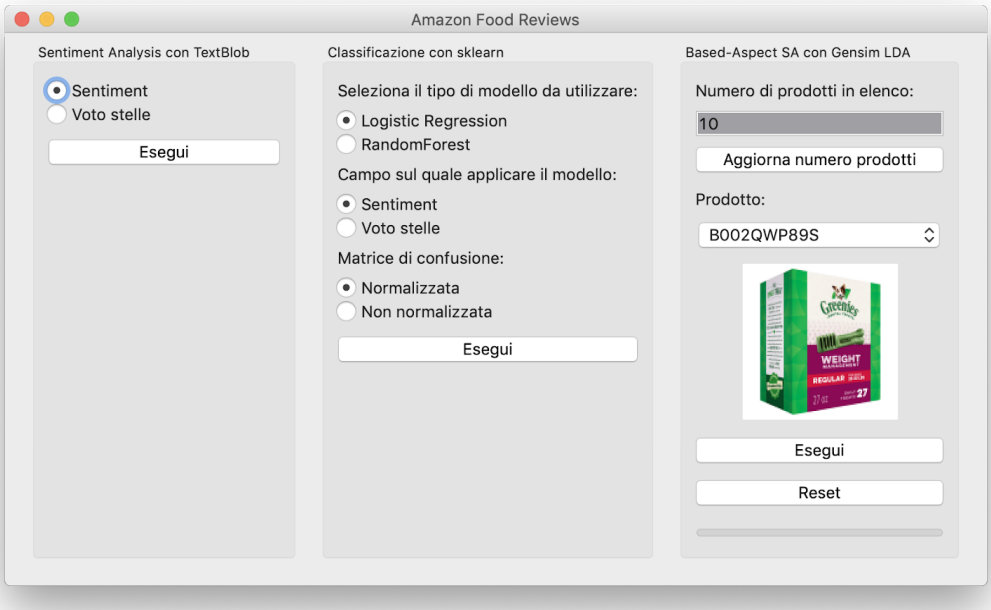
Su ogni topic estratto dal modello è stata calcolata la polarità tramite TextBlob. Il calcolo della polarità è stato effettuato sulle prime 100 parole più occorrenti per l'argomento ed ha richiesto:

- per ogni parola il calcolo della polarità;
- visto l'alto numero di parole senza sentiment il valore dei pesi dei termini è stato normalizzazione e distribuito solo sulle parole con sentiment;
- somma delle polarità con il peso normalizzato delle parole portatrici di sentiment;
- visualizzazione dei grafici relativi a una prodotto.



# Graphical User Interface

Per l'esecuzione del codice è stata implementata una GUI con la libreria Python PyQt5. L'interfaccia realizzata è molto semplice e basilare, di seguito viene comunque riportata una breve guida che illustra i vari componenti e funzionalità.




## Sentiment Analysis con TextBlob


È possibile selezionare il campo sul quale fare il confronto con la polarità per visualizzare i grafici.

Alla prima esecuzione vengono calcolate le polarità associate ai testi ma successivamente viene solo mostrato i grafici comparatori, senza aver bisogno del ricalcolo della polarità e impiega molto meno tempo.

## Classificazione con sklearn


Tramite l'interfaccia e i pulsanti di selezione è possibile scegliere:

- il tipo di modello da selezionare;
- il campo su cui applicare il modello;
- la modalità di visualizzazione della matrice di confusione.

## Based-Aspect Sentiment Analysis con gensim LDA


Tramite l'interfaccia è possibile scegliere il numero di prodotti da visualizzare nel menù a tendina ed incrementarlo o diminuirlo.

Un volta effettuato la scelta, tramite menù a tendina, del prodotto è possibile far partire l'analisi con il pulsante *Esegui*, che effettuerà l'aspect based sul prodotto selezionato e stamperà i vari grafici. I modelli migliori generati ad ogni esecuzione vengono salvati per potervi accedere. Il pulsante *Reset* riporta il menù a tendina allo stato iniziale e cancella i modelli memorizzati.