



Sentiment Analysis on Food Reviews

Pietro Colombo **793679**
Marco Fagioli **808176**

Obiettivi del progetto

- Sentiment Analysis delle reviews
 - **correlazione** polarità - voto stelle
- Classificazione con machine learning
 - **positività/negatività** delle review
- Aspect-Based Sentiment Analysis
 - divisione per **topic differenti**
 - **sentiment** per ogni topic



Dataset

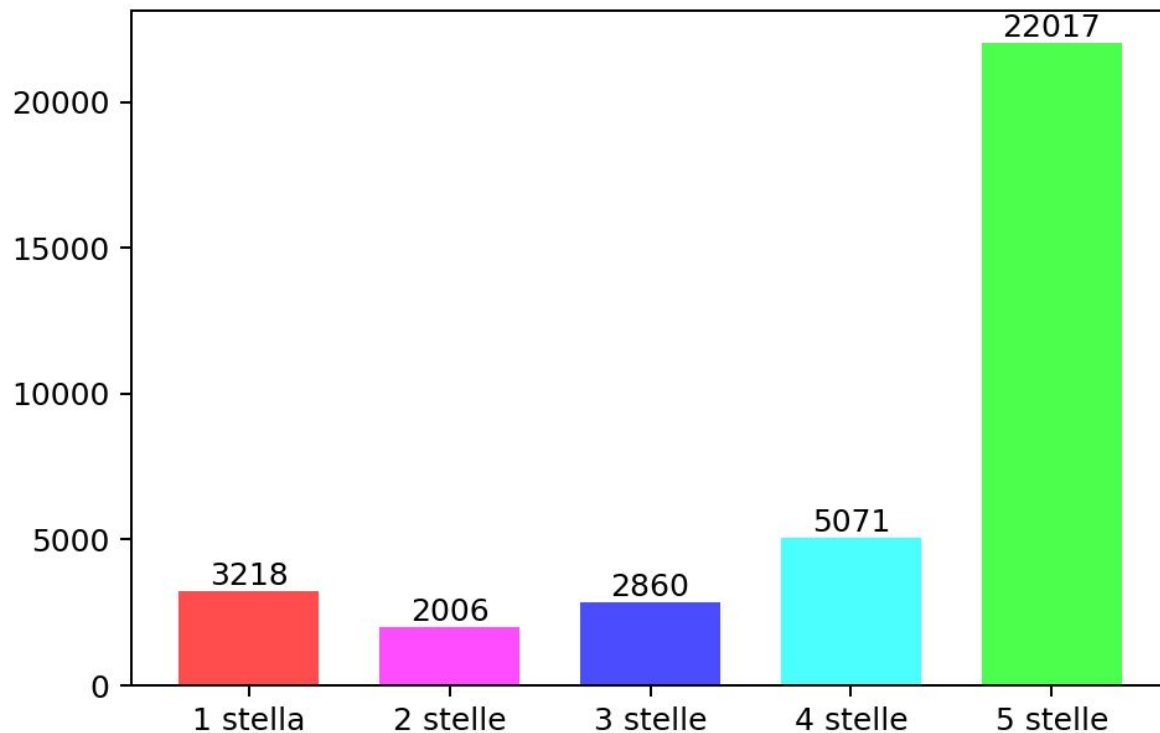
Il dataset contiene **35172 differenti review** provenienti da **amazon**

Le **feature** presenti sono:

- productid
- userid
- score (valutazione da 1 a 5)
- text

	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...

Distribuzione dei voti delle recensioni



Data preprocessing



Creazione **nuovo campo di testo** con:

- trasformazione tutto in minuscolo;
- rimozione tag HTML;
- rimozione stopword (nltk stopwords);
- rimozione punteggiatura;
- stemming (nltk LancasterStemmer).

`df['text']`  `df['clean_text']`

Creazione **sentiment** da voto stelle:

- 4 o 5 stelle → 'positive'
- 3 stelle → 'neutral'
- 1 o 2 stelle → 'negative'

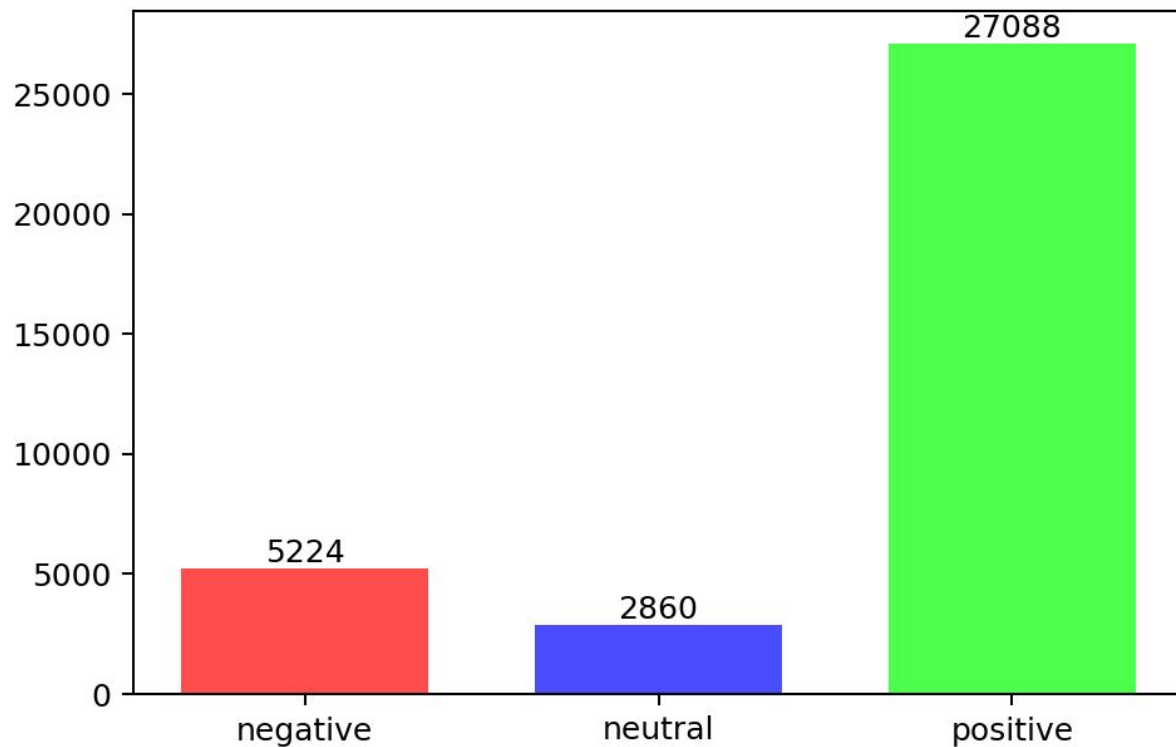
`df['score']`  `df['sentiment']`

Dataset filtrato



	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

Distribuzione del sentiment sulle recensioni



Graphical User Interface

Sentiment Analysis con TextBlob

☒ Sentiment

☐ Voto stelle

Esegui

Classificazione con sklearn

Seleziona il tipo di modello da utilizzare:

☒ Logistic Regression

☐ RandomForest

Campo sul quale applicare il modello:

☒ Sentiment

☐ Voto stelle

Matrice di confusione:

☒ Normalizzata

☐ Non normalizzata

Esegui

Based-Aspect SA con Gensim LDA


Numero di prodotti in elenco:

10

Aggiorna numero prodotti


Prodotto:

B002QWP89S

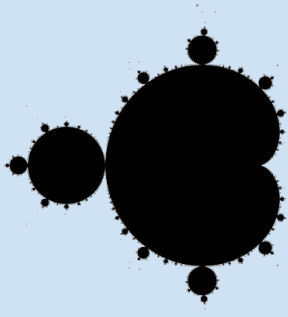


Esegui

Reset



Sentiment Analysis con TextBlob



Approccio naive tramite il metodo

TextBlob:

- calcolo **polarità** di ogni review sul testo 'filtrato'
- **correlazione** con polarità
 - **stelle votate** nelle review;
 - **sentiment** associato alle review

GUI per TextBlob

Sentiment Analysis con TextBlob

☒ Sentiment
☐ Voto stelle

Esegui

Amazon Food Reviews

Classificazione con sklearn

Seleziona il tipo di modello da utilizzare:

☒ Logistic Regression
☐ RandomForest

Campo sul quale applicare il modello:

☒ Sentiment
☐ Voto stelle

Matrice di confusione:

☒ Normalizzata
☐ Non normalizzata

Esegui

Based-Aspect SA con Gensim LDA


Numero di prodotti in elenco:

10

Aggiorna numero prodotti

Prodotto:

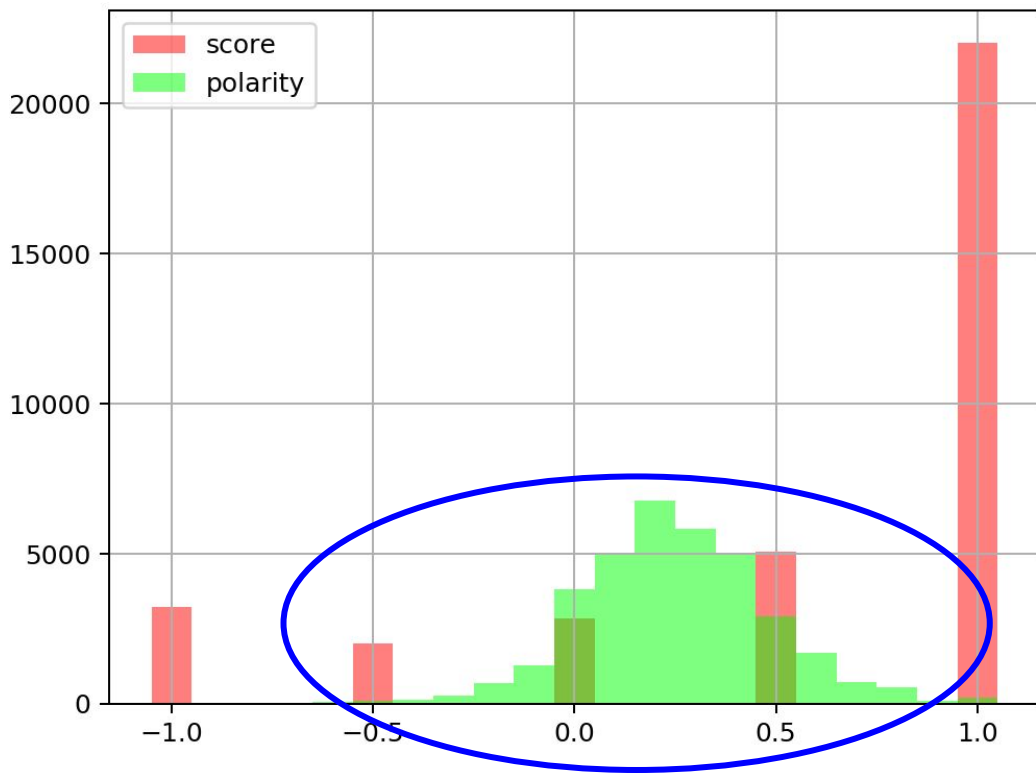
B002QWP89S



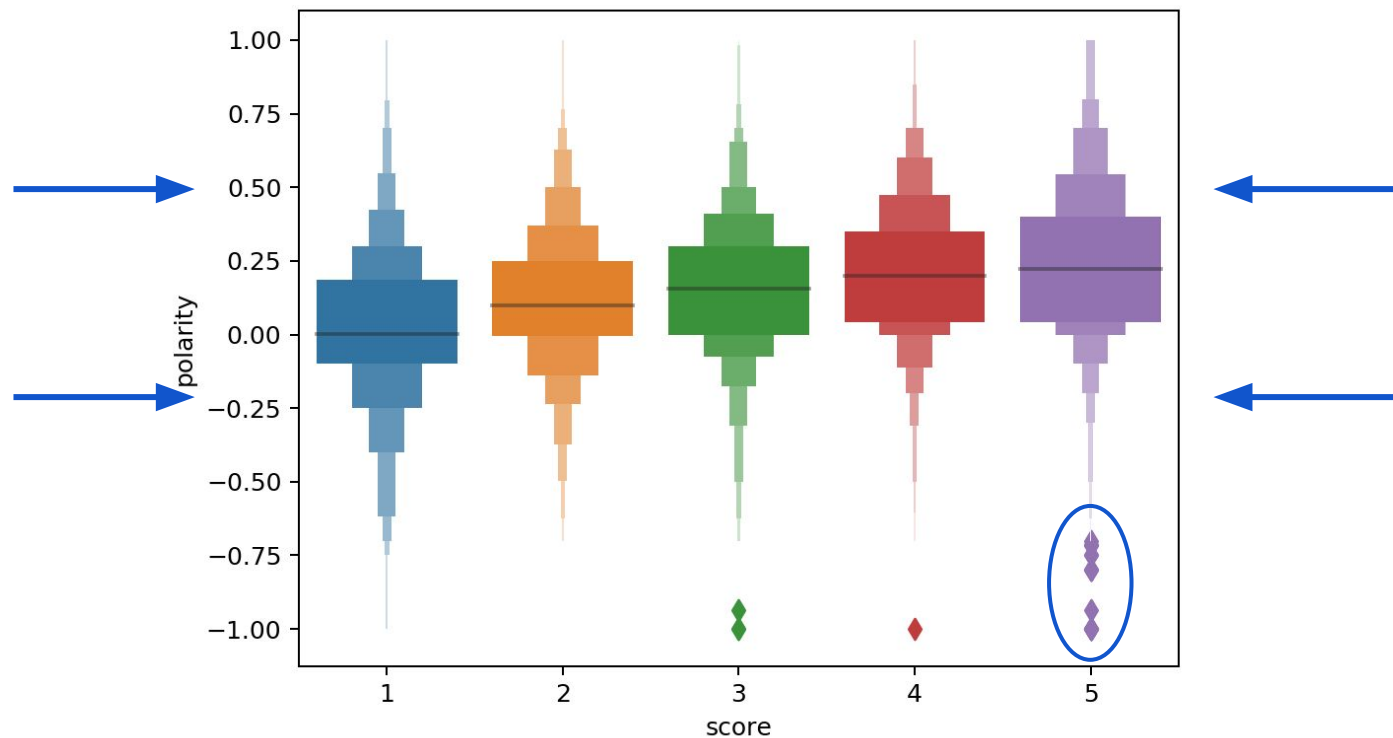
Esegui

Reset

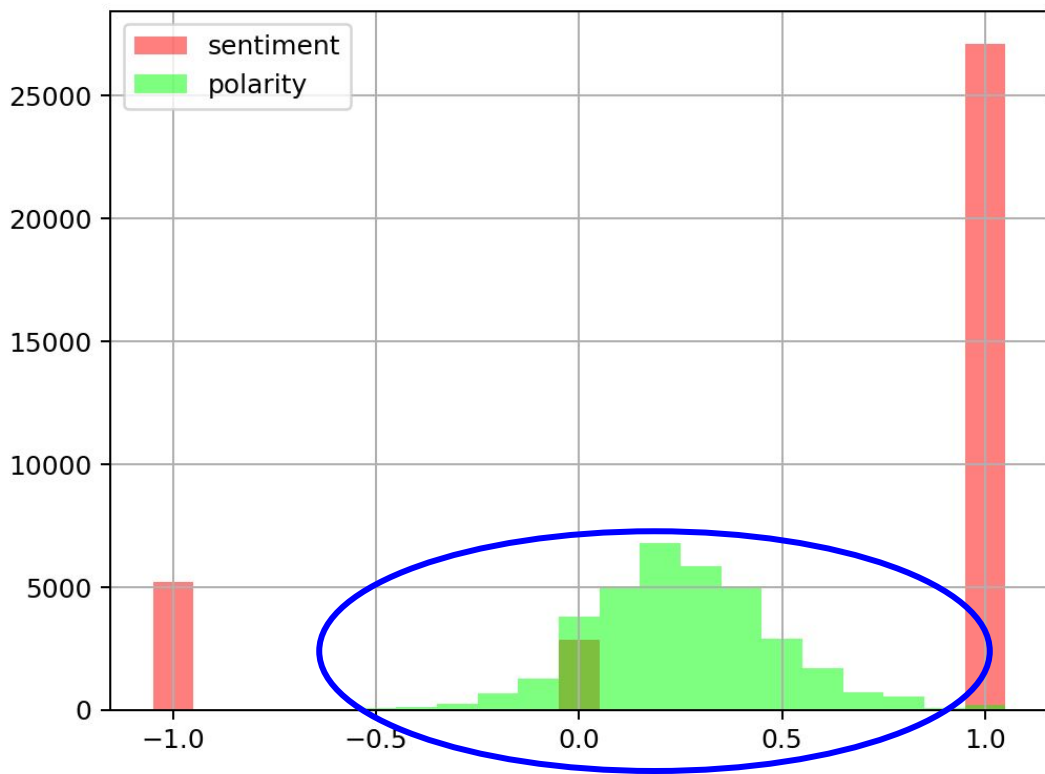
Correlazione polarità-voto stelle



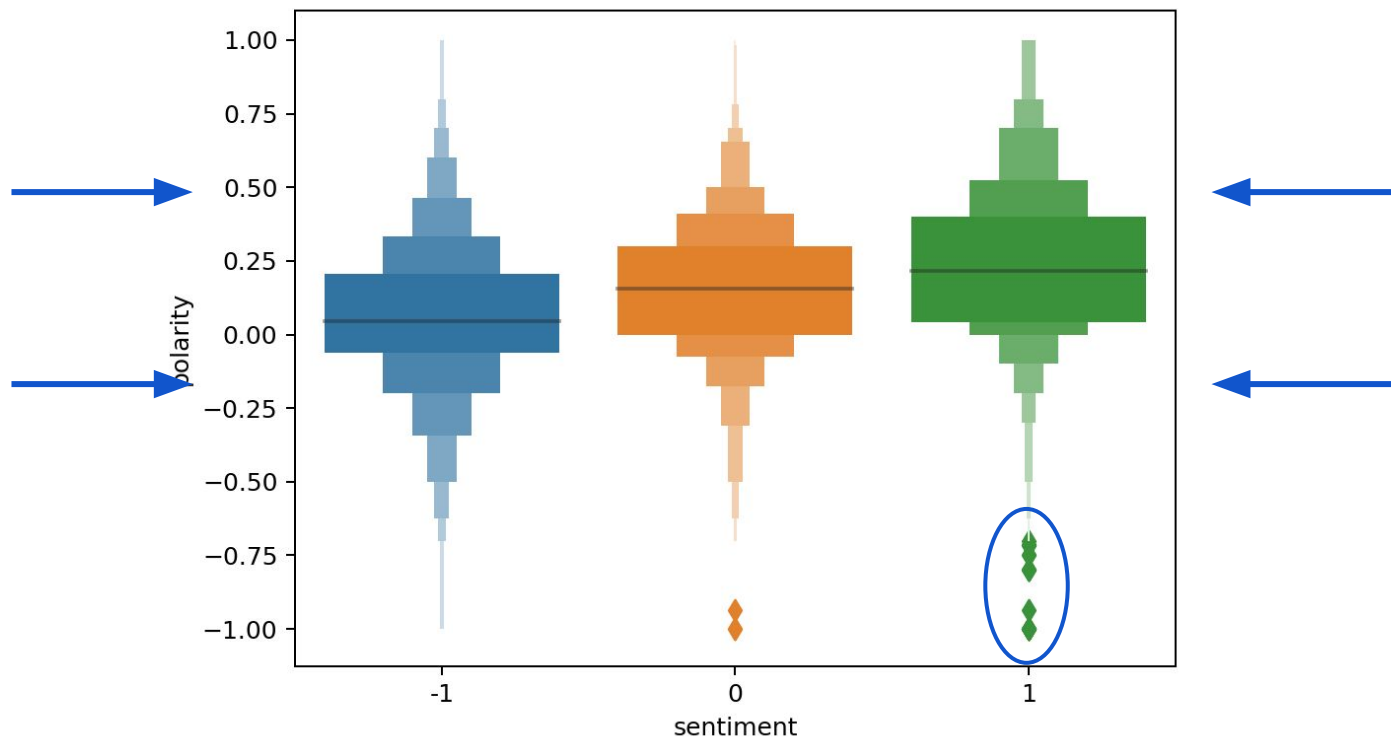
Distribuzione polarità-voto stelle



Correlazione polarità-sentiment



Distribuzione polarità-sentiment



Risultati correlazione TextBlob

Come mostrato dai grafici **non esiste correlazione** tra la **votazione** in **stelle** data dall'utente e il **sentiment** associato al corpo del testo.



```
s = "It is okay. I would not go out of my way to buy it again"
TextBlob(s).sentiment
Sentiment(polarity=0.5, subjectivity=0.5)
```



Classificazione con sklearn



Classificazione **positività** o **negatività**
delle review tramite machine learning
supervisionato sul testo:

- divisione del dataset in
 - 80% **train**
 - 20% **test**
- uso di due **differenti modelli**
 - `LogisticRegression`
 - `RandomForest`

GUI per sklearn

Sentiment Analysis con TextBlob

☒ Sentiment
☐ Voto stelle

Esegui

Amazon Food Reviews

Classificazione con sklearn

Seleziona il tipo di modello da utilizzare:

☒ Logistic Regression
☐ RandomForest

Campo sul quale applicare il modello:

☒ Sentiment
☐ Voto stelle

Matrice di confusione:

☒ Normalizzata
☐ Non normalizzata

Esegui

Based-Aspect SA con Gensim LDA


Numero di prodotti in elenco:

10

Aggiorna numero prodotti

Prodotto:

B002QWP89S



Esegui

Reset

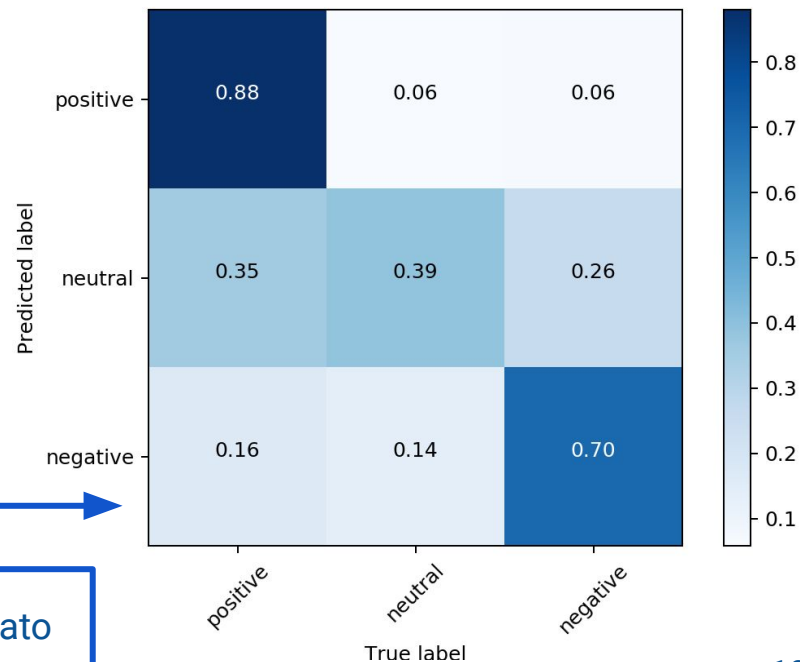
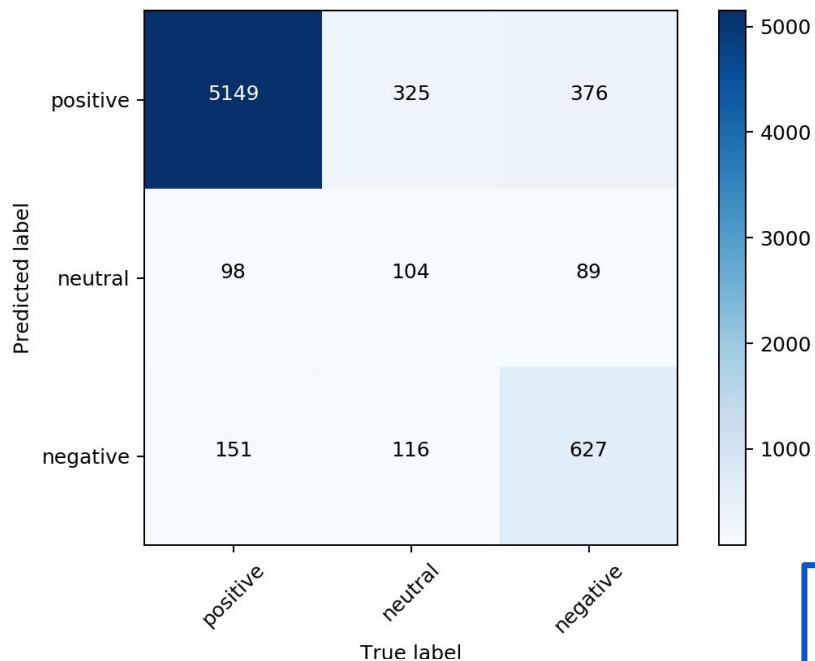
Classificazione con sklearn



Tramite l'interfaccia vengono scelti il **modello** da usare, il **campo** su cui applicarlo e la **normalizzazione** o meno delle matrici di confusione.

- Divisione del dataset nella parte di **train** e di **test**
- **Creazione e training** del modello scelto
 - Necessaria una preparazione dell'input
- **Predizione** sul test dataset

Matrici di confusione



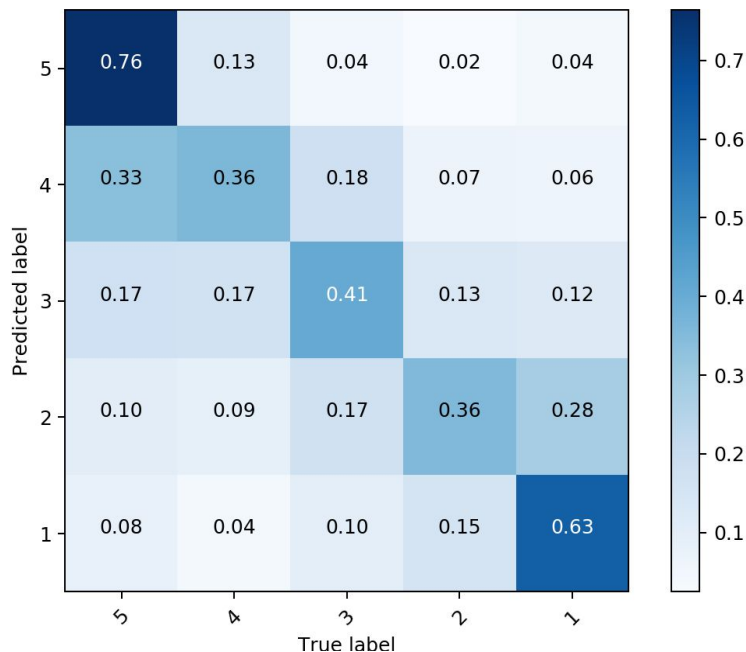
Normalizzato

Confronto classificazione voto stelle

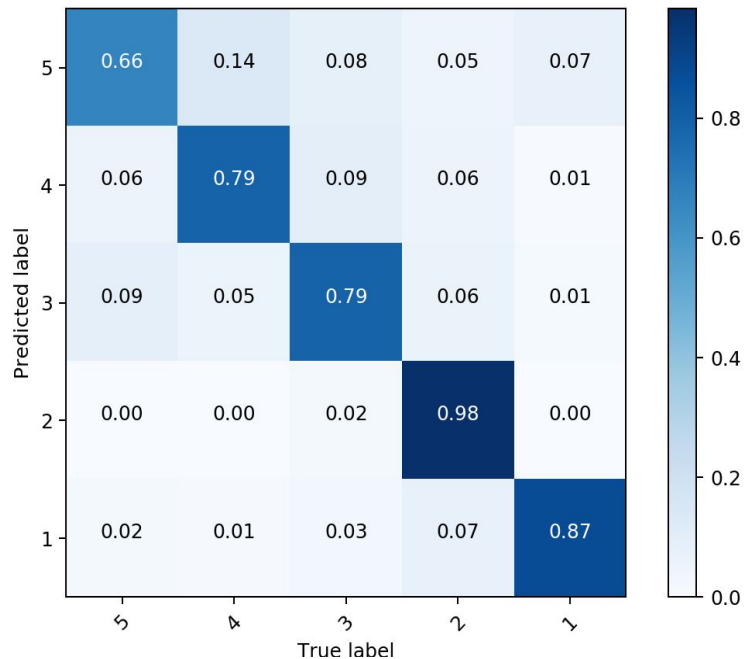
(LogisticRegression vs RandomForest)



LogisticRegression



RandomForest

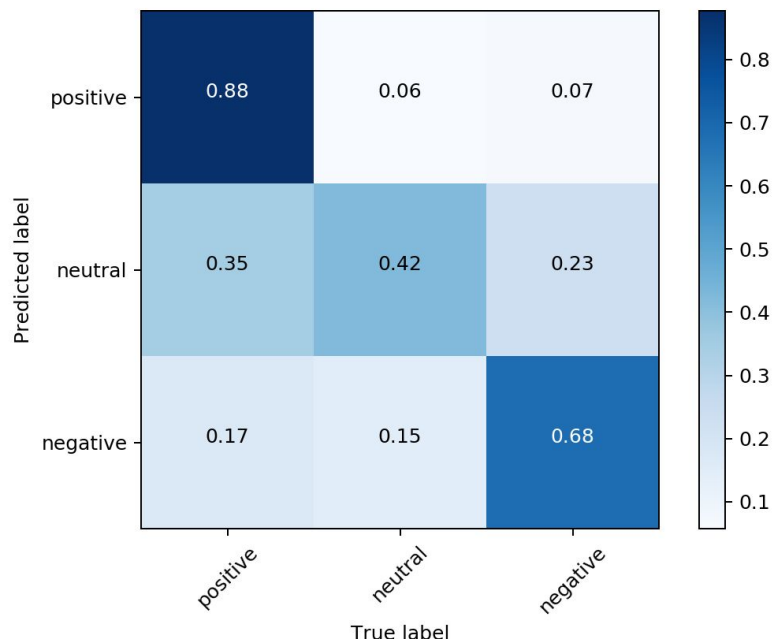


Confronto classificazione sentiment

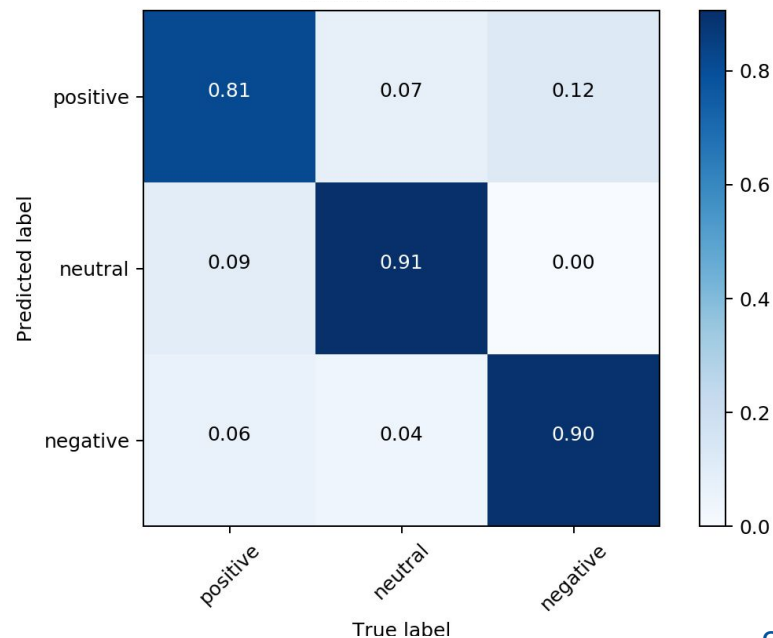
(LogisticRegression vs RandomForest)



LogisticRegression



RandomForest



Confronto risultati classificazione sentiment



Entrambi i modelli hanno una **accuracy** tra **l'80% e l'85%** in base all'esecuzione

LogisticRegression

- precision media
- recall media

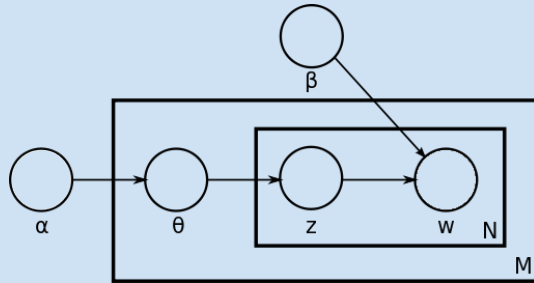
sentiment	precision	recall
positive	0.88	0.95
neutral	0.42	0.21
negative	0.68	0.57
valore medio	0.66	0.58

RandomForest

- precision alta
- recall bassa

sentiment	precision	recall
positive	0.81	1.00
neutral	0.91	0.09
negative	0.90	0.28
valore medio	0.87	0.46

Aspect-Based SA con Gensim LDA



- Applicato sui **singoli prodotti** data la grande eterogeneità del dataset
- Suddivisione delle reviews in **topic**
 - scelta numero topic ottimale
- Calcolo della **polarità** per ogni topic

GUI per gensim LDA

Sentiment Analysis con TextBlob

☒ Sentiment
☐ Voto stelle

Esegui

Classificazione con sklearn

Seleziona il tipo di modello da utilizzare:
☒ Logistic Regression
☐ RandomForest

Campo sul quale applicare il modello:
☒ Sentiment
☐ Voto stelle


Matrice di confusione:
☒ Normalizzata
☐ Non normalizzata

Esegui

Based-Aspect SA con Gensim LDA

Numero di prodotti in elenco:
10
Aggiorna numero prodotti

Prodotto:
B002QWP89S



Esegui

Reset

Aspect-Based SA con Gensim LDA



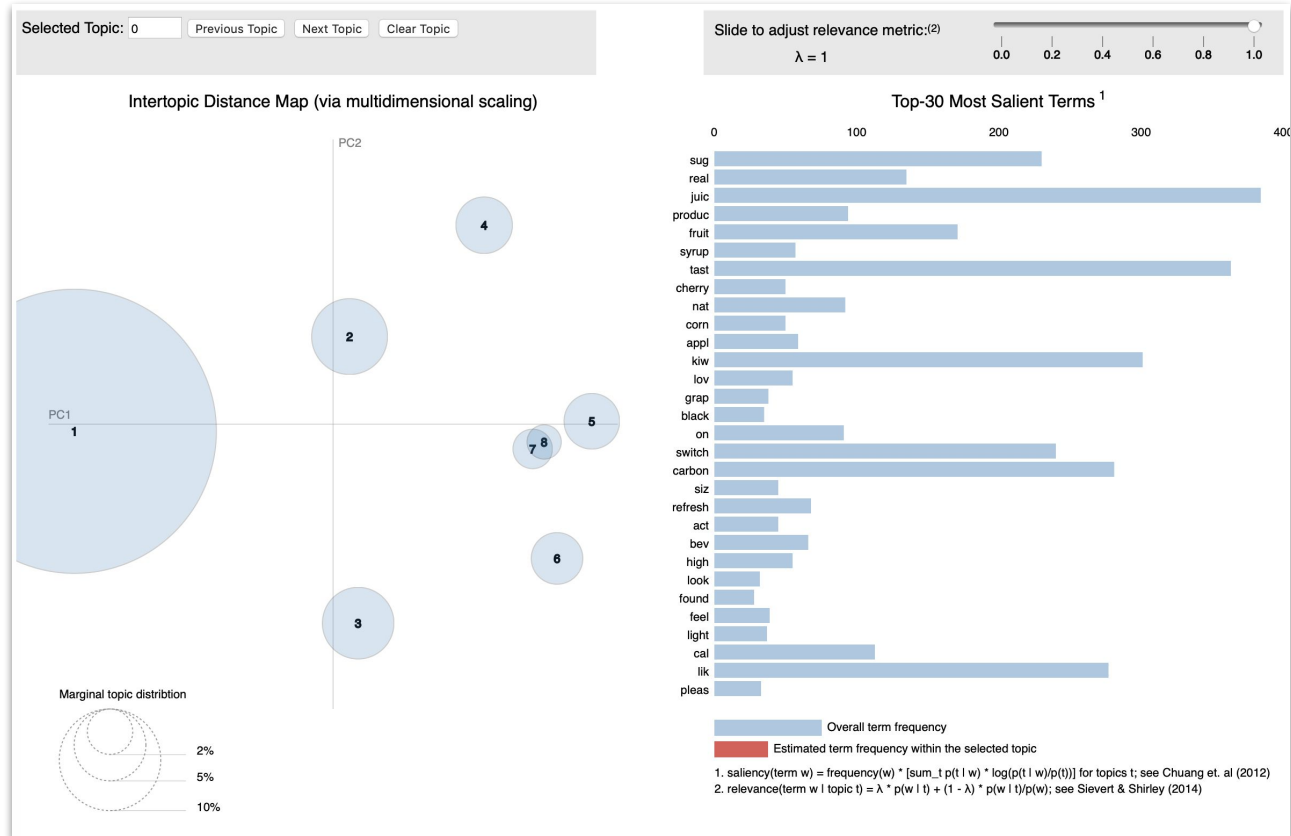
- Tramite l'interfaccia viene scelto il **prodotto** su cui fare l'analisi.
- Viene applicato il modello variando il **numero di topic** fra 2 e 10 compresi:
 - vengono salvati temporaneamente i **modelli**;
 - per ogni modello viene calcolata e memorizzata la **coerenza**.
- Il modello con il **miglior** valore di coerenza è quello scelto.

Aspect-Based SA con Gensim LDA

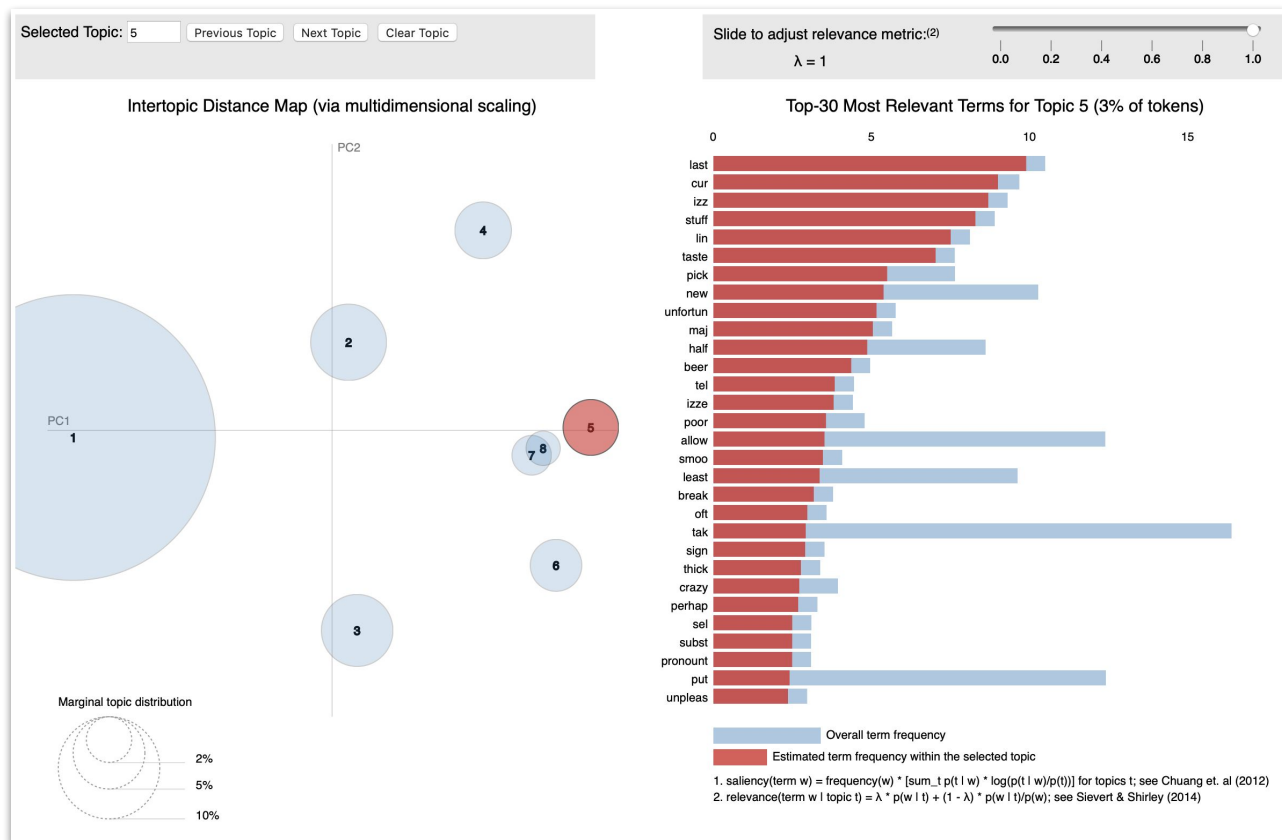


- Il modello con miglior coerenza viene **preparato** per la visualizzazione con pyLDAvis.
- Per ogni **topic** individuato viene calcolata la **polarità** associata.
- Tramite **pyLDAvis** viene mostrato, su browser, il modello generato

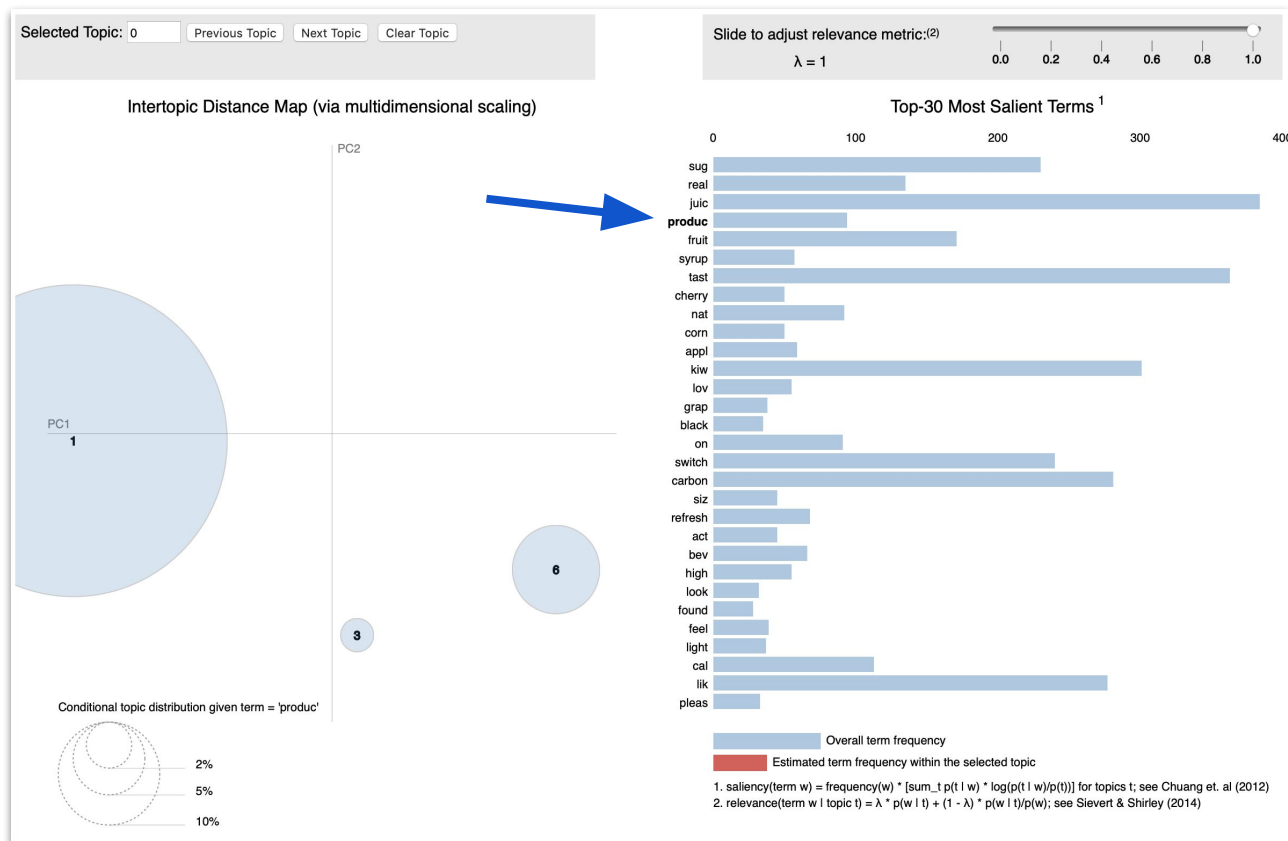
Modello LDA



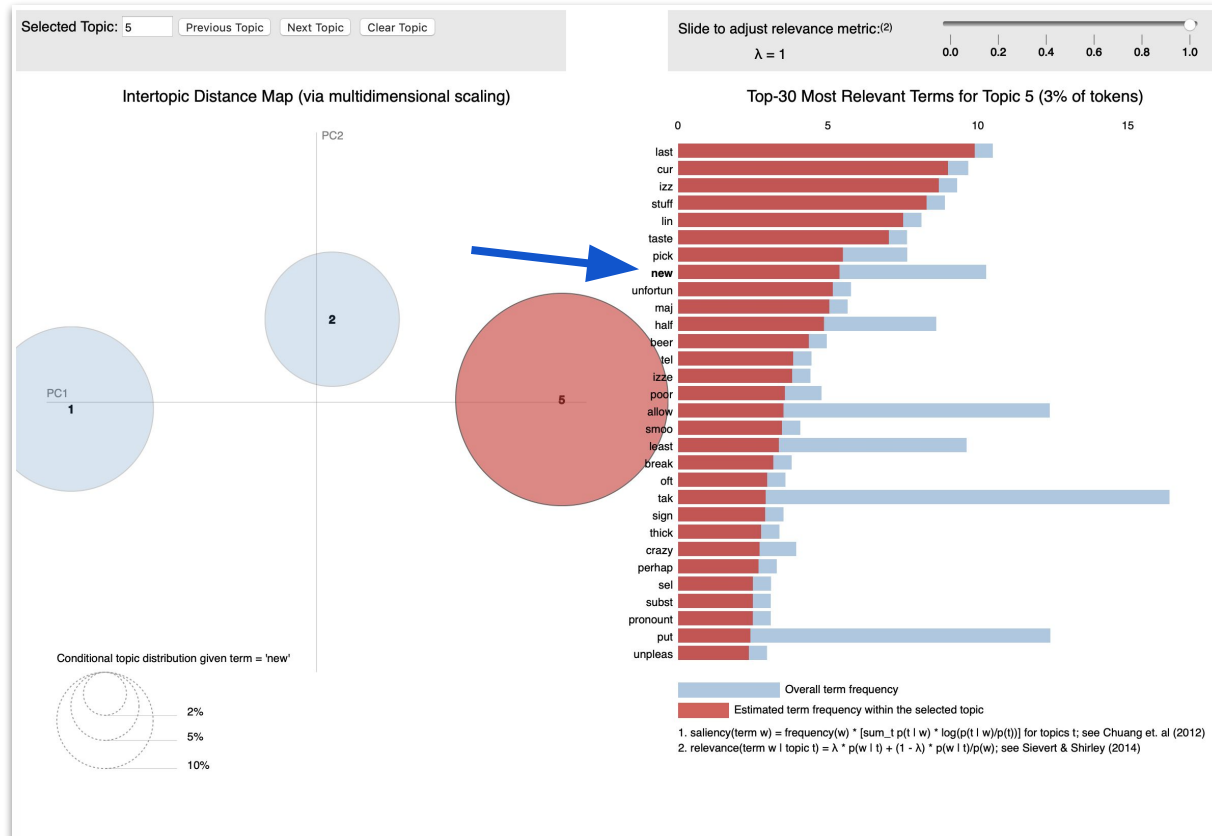
Modello LDA



Modello LDA



Modello LDA



Polarità per topic

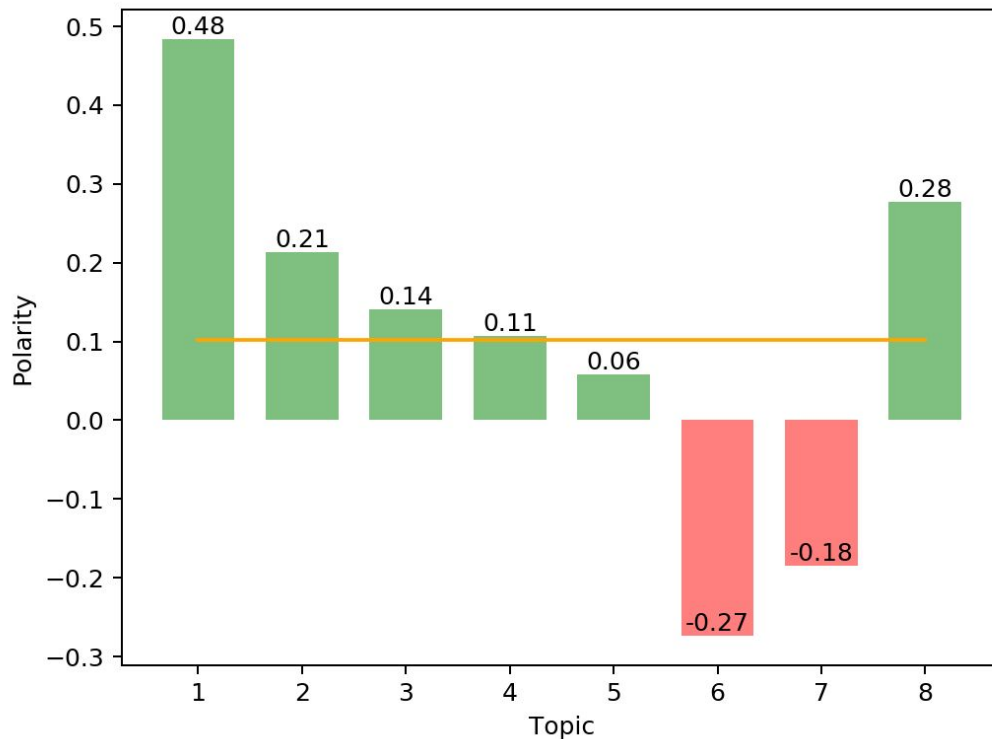


Per **topic** è stata calcolata la polarità:

- **100 parole** per ogni topic;
- per ogni parola calcolo della **polarità pesata**;
 - necessaria una normalizzazione del peso per le parole con sentiment;
- **somma** delle **polarità** delle parole con **sentiment**.



Grafico polarità per topic



Grazie per l'attenzione!

Pietro Colombo
Marco Fagioli

