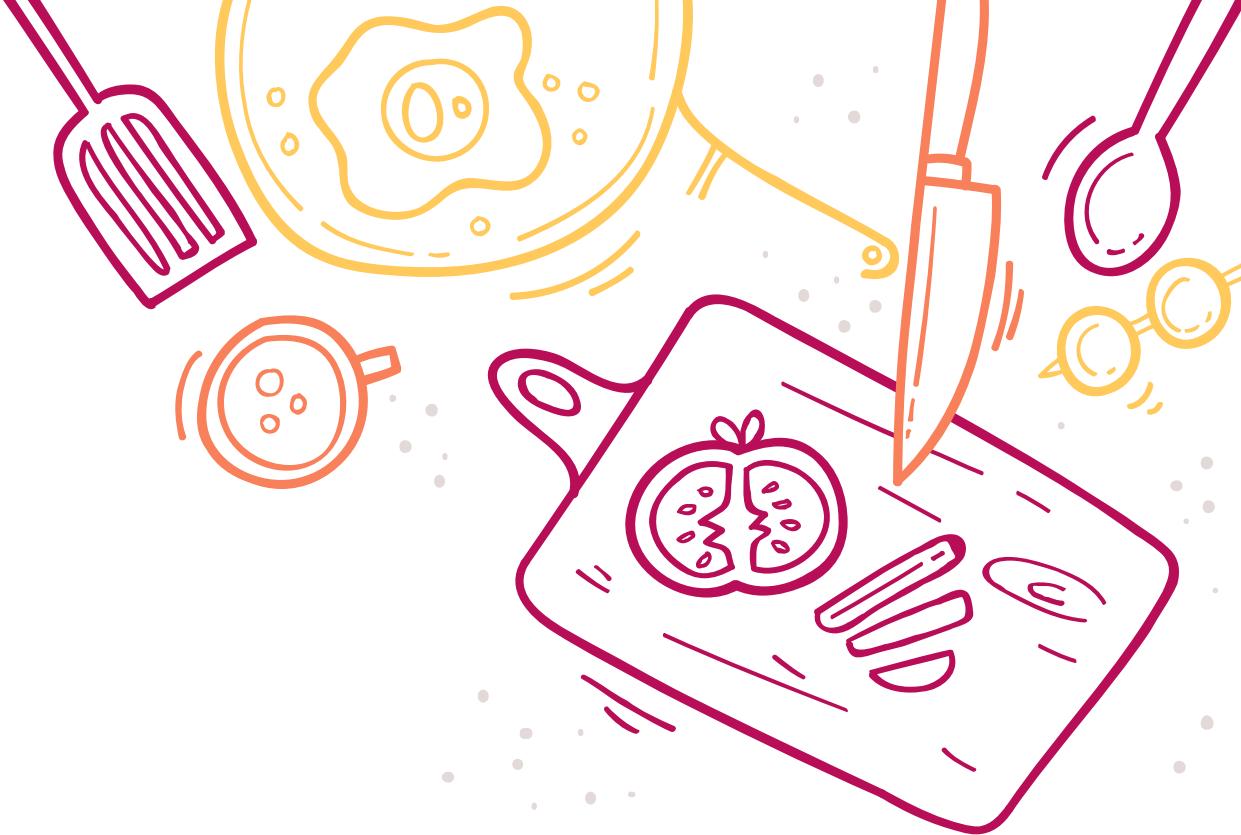


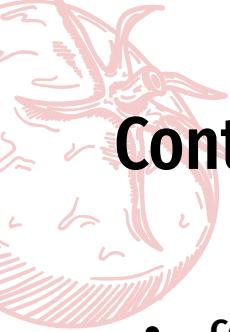
Amazon Fine Food Reviews

Raffaele Cerizza 845512



Contesto e obiettivi

- **Contesto:**
 - Recensioni su cibi e bevande pubblicate sul sito di Amazon. Lingua inglese.
 - Link: <https://nijianmo.github.io/amazon/index.html>.
- **Obiettivi:**
 - **Analisi sui prodotti:**
 - quali sono i prodotti più apprezzati dagli utenti?
 - quali sono i prodotti meno apprezzati dagli utenti?
 - **Analisi sugli aspetti:**
 - quali sono gli aspetti dei prodotti che gli utenti gradiscono di più?
 - quali sono gli aspetti dei prodotti che gli utenti gradiscono di meno?
 - **Analisi sui prezzi:**
 - quali sono le fasce di prezzo più recensite?
 - come varia il sentimento degli utenti rispetto al prezzo dei prodotti?



Descrizione del dataset

- **Composizione:**

- informazioni sulle recensioni;
- informazioni sui prodotti.

- **Attributi principali:**

Attributo	Descrizione
asin	codice alfanumerico identificativo del prodotto
reviewerID	codice alfanumerico identificativo del recensore
reviewTime	data della recensione con giorno, mese, anno
reviewText	testo della recensione
summary	sommario della recensione
price	prezzo del prodotto in dollari
overall	valutazione del prodotto come numero intero fra 1 e 5

- **Modifiche principali:**

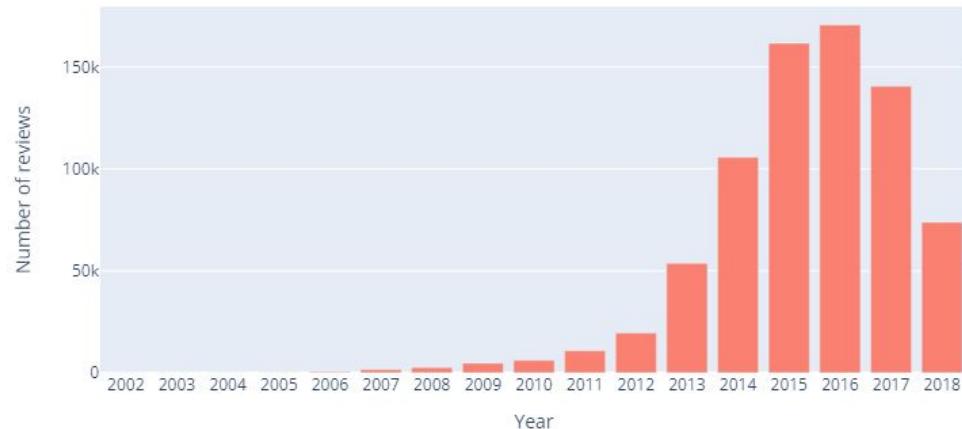
- attributo *overall* rinominato *rating*;
- introduzione attributo *opinion*;
- rimozione recensioni con informazioni importanti mancanti.

Descrizione del dataset

- Sintesi:

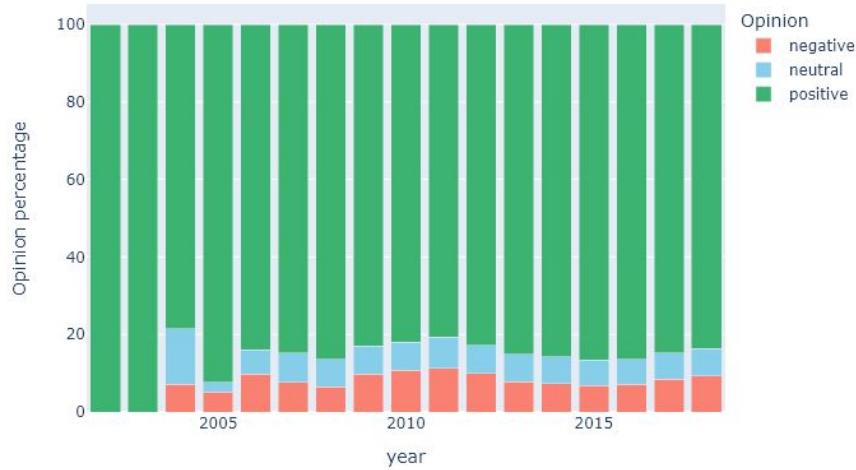
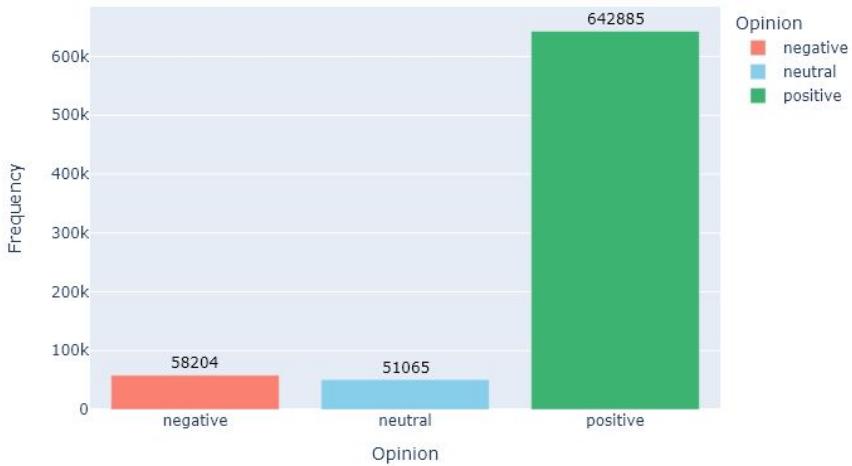
Recensioni	Prodotti	Utenti	Voto medio
752154	25218	126485	4,462

- Distribuzione delle recensioni per anno:



Distribuzione delle opinioni

- Distribuzione delle opinioni degli utenti:



Preprocessing del testo

Operazioni di preprocessing sul testo:

- espansione delle contrazioni: “don’t” -> “do not”;
- tokenizzazione;
- conversione dei caratteri in minuscolo;
- semplificazione delle ripetizioni: “tooood” -> “too”;
- rimozione della punteggiatura;
- rimozioni ulteriori: stop words e tag HTML;
- gestione delle negazioni.



Modelli di sentiment analysis

- **Approcci basati su lessici**

- **AFINN:**
 - valenza per ogni parola del lessico: da -5 a +5;
 - valenza invertita per la parola immediatamente successiva a una negazione;
 - sentimento della recensione come somma delle valenze delle sue parole:
 - sentimento negativo se somma negativa;
 - sentimento positivo se somma positiva;
 - sentimento neutrale se somma nulla.
- **VADER:**
 - sfrutta anche *(i)* punteggiatura, *(ii)* avverbi, *(iii)* congiunzioni e *(iv)* negazioni;
 - preprocessing leggero con: espansione contrazioni, semplificazione ripetizioni e rimozione tag HTML;
 - negazioni non rimosse e gestite da VADER;
 - VADER restituisce uno score di sentimento per il testo:
 - sentimento negativo se lo score è minore o uguale a -0.05;
 - sentimento positivo se lo score è maggiore o uguale a 0.05;
 - sentimento neutrale se lo score è compreso fra -0.05 e 0.05.

Modelli di sentiment analysis

- **Approccio basato sul Machine Learning**

- Modello di regressione logistica:
 - modello scelto per l'efficienza computazionale;
 - utilizzato un algoritmo di ottimizzazione e una loss compatibili con problemi multi-classe;
 - negazioni gestite aggiungendo il prefisso “NOT_” a ogni parola compresa fra una negazione e il primo simbolo di punteggiatura successivo;
 - recensioni rappresentate come Bag of Words;
 - modello addestrato con una 3-fold cross validation usando le opinioni come etichette.

Modelli di sentiment analysis

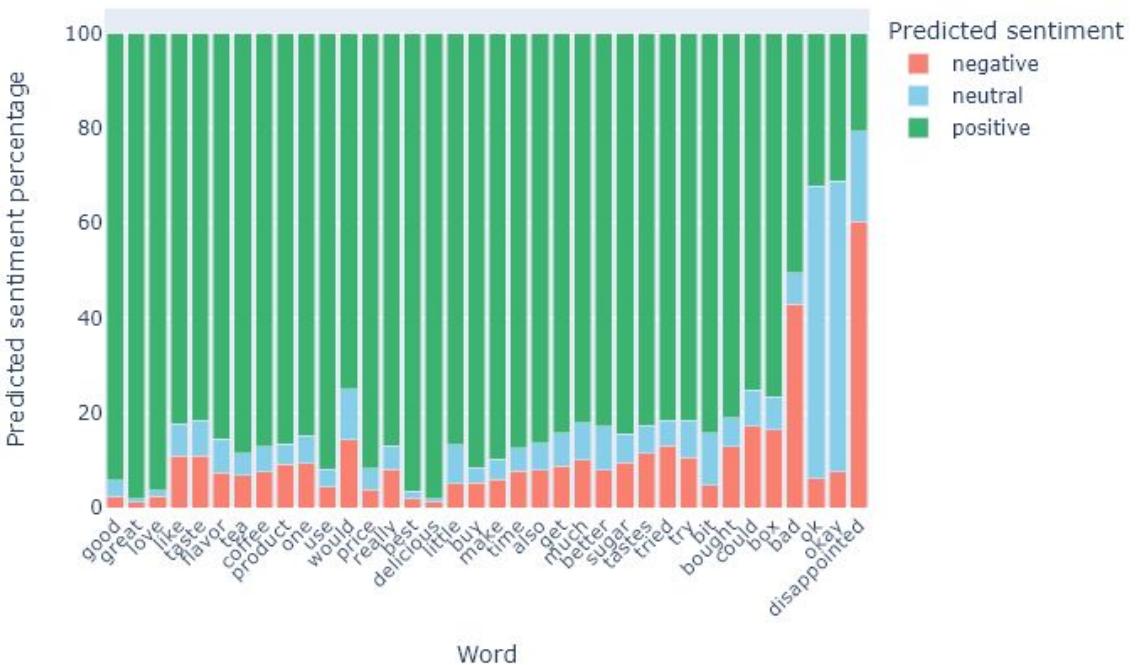
- Confronto fra gli approcci di sentiment analysis:

Modello	<i>Undersampling</i>	Accuratezza	Precisione	Recall	<i>F1-score</i>
AFINN	no	0.810	0.514	0.494	0.500
VADER	no	0.823	0.499	0.493	0.496
Reg. log.	no	0.892	0.693	0.582	0.621
Reg. log.	sì	0.789	0.560	0.692	0.591

- Metriche calcolate con approccio macro average.
- I risultati migliori per ogni metrica sono evidenziati in grassetto.

Distribuzione dei sentimenti

- Distribuzione dei sentimenti sulle recensioni che contengono le parole più frequenti:



- Sentimento positivo predominante.
- Sentimento positivo frequente anche per recensioni che contengono parole tipicamente negative come "bad".
- Sentimento neutrale prevalente per le recensioni che contengono le parole "ok" e "okay".

Esempi di misclassificazione

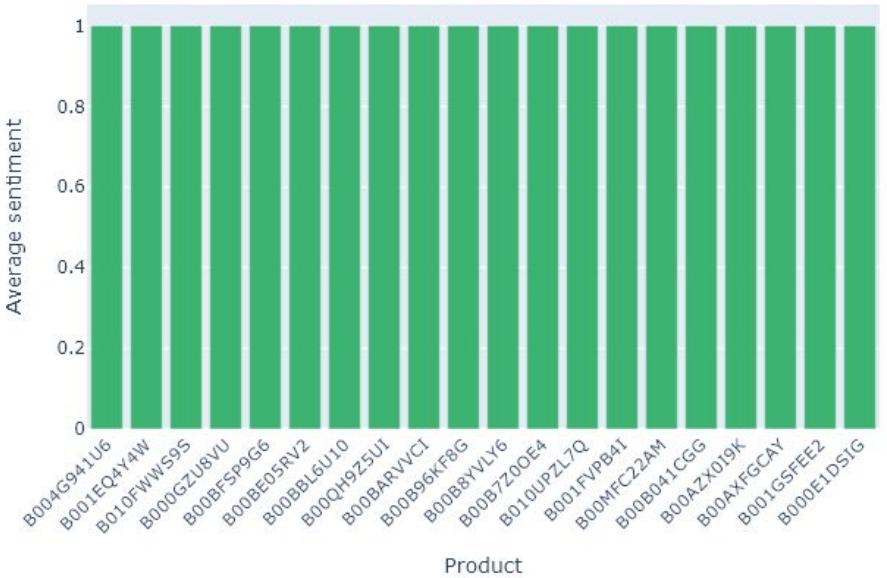
- **Sentimenti misti:**
 - Recensione: “I like the texture and taste of these wraps but they are pricey! They are a bit small when putting things inside so this makes eating a little messy. Otherwise I like them”.
 - Valutazione dell’utente: 4.
 - Sentimento predetto: neutrale.
- **Ironia:**
 - Recensione: “Let me start off by saying... ICK.. Disgusting. These are absolutely terrible. That’s exactly why I rated this product five stars. It was exactly and completely what I was expecting, and thankfully my coworkers were not expecting. It’s a great office game/prank and I will order again soon!”.
 - Valutazione dell’utente: 5.
 - Sentimento predetto: negativo.

Esempi di misclassificazione

- **Sentimenti dipendenti dal contesto:**
 - Recensione: “This product has changed. It now has very large tea leaves and tastes a bit **bitter**. It was my favorite Earl Grey tea of all time. I'm hunting for a new brand. The lid was defective as well. Sigh.....”.
 - Valutazione dell'utente: 2.
 - Sentimento predetto: neutrale.
 - Recensione: “This coffee has a very bold taste compared to the original Folgers blend. I prefer this stronger **bitter** taste and will continue to buy this particular blend.”.
 - Valutazione dell'utente: 5.
 - Sentimento predetto: neutrale.
- **La parola “ok”:**
 - Recensione: “ok”.
 - Valutazione dell'utente: 5
 - Sentimento predetto: neutrale.

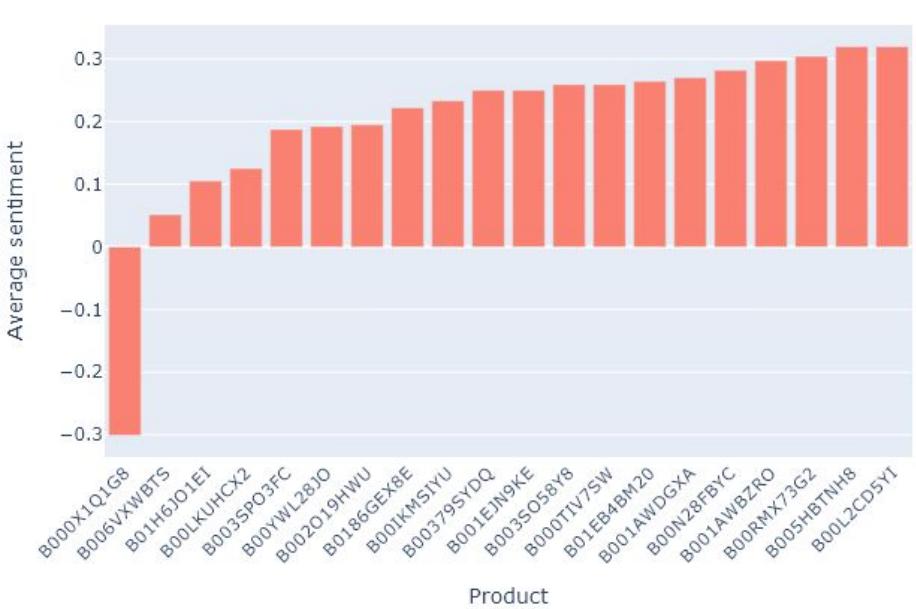
Prodotti più apprezzati

Prodotti con il sentimento medio più elevato:



ASIN	Descrizione	Sentimento medio
B004G941U6	minestrone	1.000
B001EQ4Y4W	condimento	1.000
B010FWWS9S	cracker	1.000
B000GZU8VU	sale	1.000
B00BFSP9G6	sale	1.000
B00BE05RV2	sale	1.000
B00BBL6U10	tè	1.000
B00QH9Z5UI	cannella	1.000
B00BARVVC1	bevanda energetica	1.000
B00B96KF8G	cioccolato	1.000
B00B8YVLY6	pancake	1.000
B00B7Z00E4	piante	1.000
B010UPZL7Q	cioccolato	1.000
B001FVPB4I	pepe	1.000
B00MFC22AM	polvere di carruba	1.000
B00B041CGG	pepe	1.000
B00AZX019K	cioccolato	1.000
B00AXFGCAY	burro di arachidi e cioccolato	1.000
B001GSFEE2	biscotti	1.000
B000E1DSIG	budino al caramello	1.000

Prodotti meno apprezzati



ASIN	Descrizione	Sentimento medio
B000X1Q1G8	burro di arachidi	-0.301
B006VXWBTS	burro di arachidi	0.051
B01H6J01EI	bevanda energetica	0.105
B00LKUHCX2	caffè	0.125
B003SPO3FC	riso	0.188
B00YWL28J0	bustine di avena	0.192
B002019HWU	pesche affettate	0.195
B0186GEX8E	cereali	0.222
B00IKMSIYU	succo	0.233
B00379SYDQ	cocktail di cetrioli e menta	0.250
B001EN9KE	brownie con burro di arachidi	0.250
B003S058Y8	tè	0.259
B000TIV7SW	propoli	0.259
B01EB4BM20	panna montata	0.264
B001AWDGXA	sardine	0.270
B00N28FBYC	barretta proteica	0.282
B001AWBZRO	cibo per gatti	0.297
B00RMX73G2	crema per caffè	0.304
B005HBTNH8	integratore alimentare	0.320
B00L2CD5YI	pasta	0.320

Modelli di aspect-based sentiment analysis

- **Modello ASUM [1]:**

- ASUM estende il topic model LDA incorporando i sentimenti;
- approccio non supervisionato;
- assume che ogni documento sia composto da frasi e che ogni frase sia associata a un aspetto;
- preprocessing come per la sentiment analysis con Machine Learning, con aggiunto lo stemming;
- aspetto = topic;
- limiti: (i) contesto; (ii) numero di topic; (iii) interpretazione risultati.

[1] Jo, Y., and Oh, A. H. Aspect and sentiment unification model for online review analysis. In Proceedings of the 4th ACM International Conference on Web Search and Data Mining (2011), WSDM '11, Association for Computing Machinery.

Modelli di aspect-based sentiment analysis

- **Modello LCF-ATEPC [2]:**

- modello neurale basato su BERT;
- sfrutta il contesto degli aspetti;
- preprocessing del testo leggero;
- numero di aspetti non predeterminato;
- utilizzato modello pre-addestrato su dataset analoghi a quello del presente lavoro;
- aspetto = parole che descrivono una caratteristica del prodotto;
- aspetti individuati con il formato BIO.

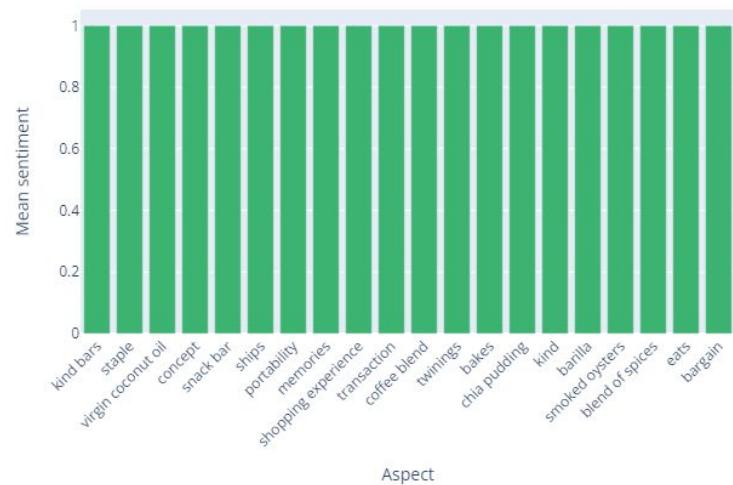
[2] Yang, H., Zeng, B., Yang, J., Song, Y., and Xu, R. A multi-task learning model for chinese-oriented aspect polarity classification and aspect term extraction. arXiv preprint arXiv:1912.07976 abs/1912.07976 (2019).

Aspetti più apprezzati

- Aspetti positivi individuati con ASUM:

Chocolate	Salt	Coffee	Sauces	Snacks	Price	Tea	Candies	Gluten-free	Quality
tast	use	coffe	flavor	great	good	tea	love	good	great
flavor	oil	flavor	use	snack	price	flavor	good	use	good
good	salt	good	good	good	great	tast	candi	tast	product
love	good	tast	like	bar	product	love	great	great	veri
great	tast	like	tast	love	buy	good	like	make	tast
use	great	cup	great	tast	amazon	like	flavor	love	price
chocol	coconut	love	sauc	eat	love	great	one	like	qualiti
like	love	roast	love	like	order	drink	tast	free	love
milk	like	great	make	veri	veri	veri	gift	gluten	excel
make	popcorn	one	soup	one	best	green	chocol	bread	fresh

- Aspetti positivi individuati con LCF-ATEPC:

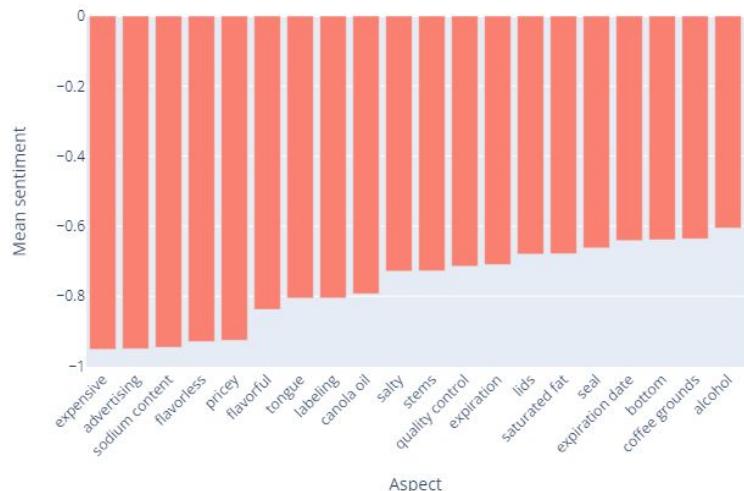


Aspetti meno apprezzati

- **Aspetti negativi individuati con ASUM:**

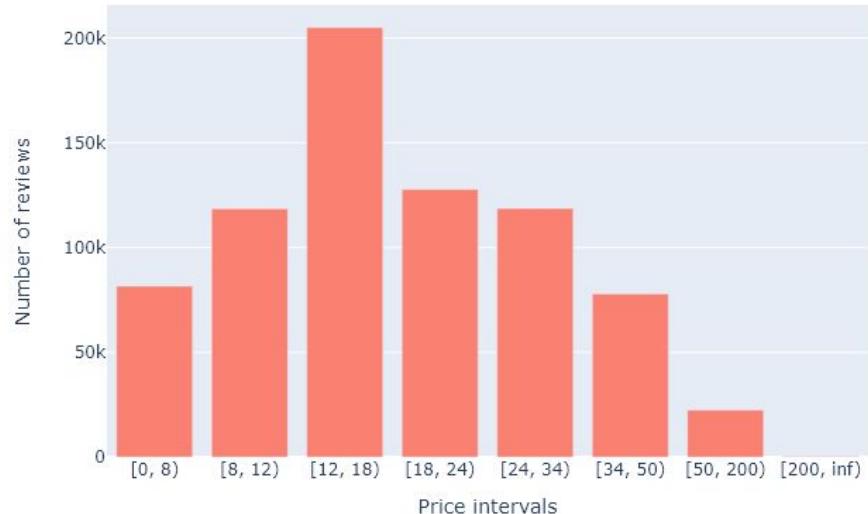
Amazon	Drinks	Cooking	Peanut	Sugar	Sweets	Price	Calories	Shipment	Packaging
product	day	water	chocol	tast	tast	store	sugar	order	bag
review	drink	use	like	like	like	buy	ingredi	time	packag
star	use	cup	tast	sugar	flavor	price	fat	buy	box
would	help	make	delici	flavor	veri	local	calori	one	open
tri	eat	add	flavor	use	littl	amazon	contain	bag	one
one	get	cook	peanut	sweet	sweet	groceri	protein	tri	like
like	like	minut	veri	water	smell	get	product	box	use
amazon	one	mix	butter	sweeteten	realli	find	serv	would	contain
compani	work	put	sweet	drink	bit	much	high	purchas	veri
becaus	time	time	nut	much	doe	better	sodium	last	plastic

- **Aspetti negativi individuati con LCF-ATEPC:**

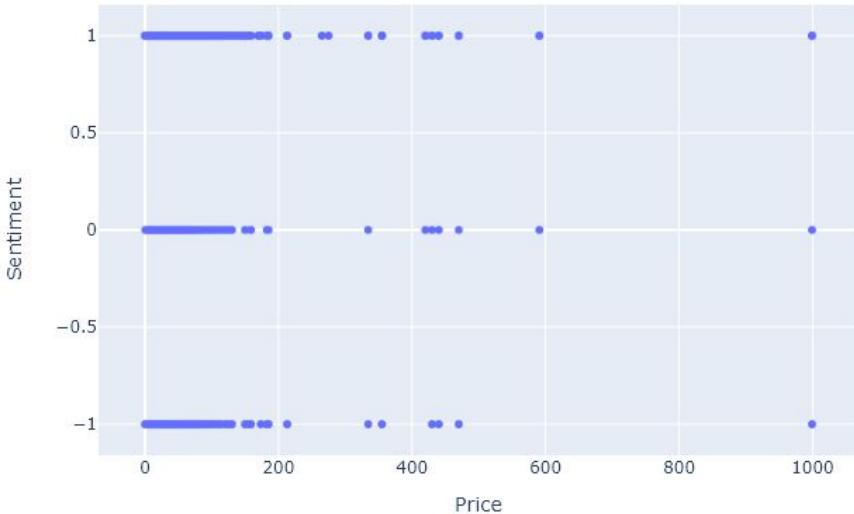


Analisi sui prezzi

Distribuzione delle recensioni per intervalli di prezzo:



Scatter plot tra prezzi e sentimenti:



- Correlazione di Pearson tra prezzi e rating: 0.012.
- Correlazione di Pearson tra prezzi e sentimenti: 0.002.

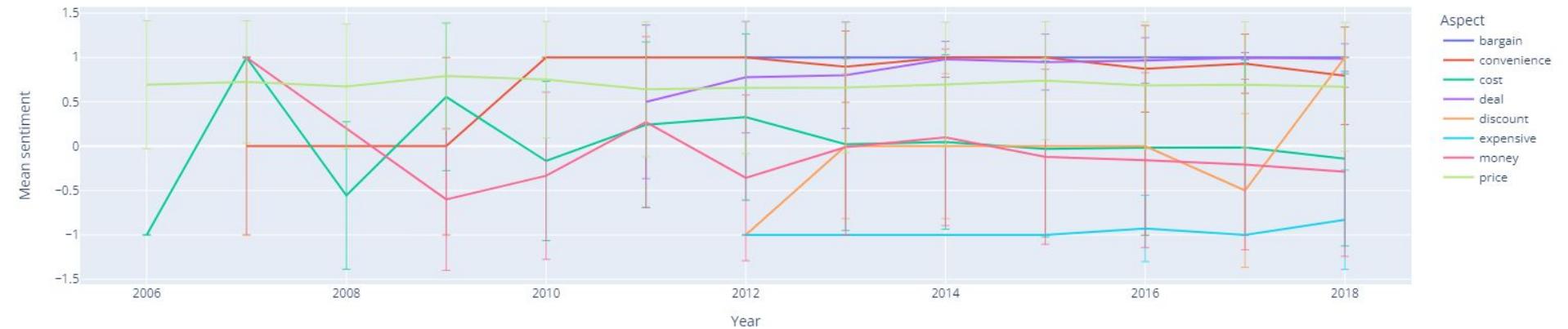
Analisi sui prezzi

- Sentimento medio degli aspetti individuati con LCF-ATEPC:

Aspetto	Media	Dev. standard	Frequenza
bargain	1.000	0.000	72
convenience	0.909	0.396	242
cost	0.004	0.987	1735
costs	-0.217	0.966	152
deal	0.967	0.253	1365
discount	-0.120	0.952	25
expensive	-0.951	0.308	206
money	-0.133	0.985	1060
price	0.695	0.705	40398
priced	0.473	0.880	2252
prices	0.254	0.952	1067
pricey	-0.926	0.378	27
pricing	0.627	0.764	572
<i>Totale</i>	0.630	0.765	49173

Analisi sui prezzi

- Distribuzione del sentimento medio sugli aspetti per ogni anno:



Conclusioni e sviluppi futuri

- **Conclusioni**
 - **Prodotti più apprezzati:** tè, snacks e sale.
 - **Prodotti meno apprezzati:** dolci e burro di arachidi.
 - **Importanza della salute.**
 - **Bassa correlazione tra prezzo e sentimento,** ma:
 - la convenienza economica è molto gradita;
 - l'eccessiva sproporzione tra prezzo e qualità è molto sgradita.
- **Sviluppi futuri**
 - **Analisi sui consumatori.**
 - **Indagine sulle recensioni più utili.**
 - **Approfondimento sull'ironia e sui sentimenti contrastanti.**



Grazie per l'attenzione