AMAZON FINE FOOD REVIEWS

CONFALONIERI RICCARDO (830404) |
RANIERI SILVIA (878067)

SOMMARIO

INTRODUZIONE e OBIETTIVI

Amazon è una dei siti di e-commerce più popolari al Mondo. Uno dei suoi punti di forza è il *sistema di recensioni* che permette a tutti i clienti di esprimere dei giudizi sui prodotti acquistati. Questo sistema è cresciuto molto negli anni e ad oggi è molto organizzato ed è un forte fattore di influenza in fase di acquisto. Alcuni studi riportano che oltre la metà dei clienti si basa proprio sulle recensioni per decidere quale prodotto acquistare.

L'obiettivo del progetto è quindi creare un sistema che permetta, in automatico, di verificare lo score assegnato alle recensione. Così facendo è possibile prevenire tecniche, da parte dei venditori, di sistemi automatici per ottenere più recensioni positivi e conseguente maggior visibilità. Per questo motivo sono stati sviluppati due diverse tecniche:

CLASSIFICATION

CLUSTERING

1 IL DATASET

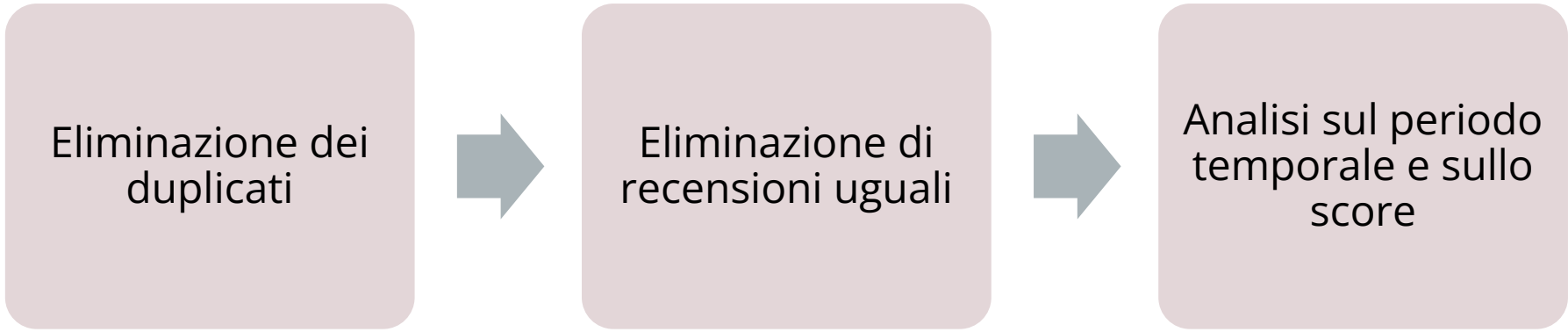
Il dataset utilizzato per lo sviluppo del progetto è reso disponibile dalla piattaforma Kaggle. Contiene i dati di recensioni di cibi raffinati ed è caratterizzato da diverse features tra cui quelle di interesse per il progetto:

- *ProductId*. Id univoco del prodotto.
- *UserId*. Id univoco per gli utenti.
- *Score*. Voto assegnato alla recensione, compreso tra [1,5]
- *Time*. Data della recensione in formato UNIX.
- *Text*. Testo della recensione.

In totale il dataset contiene oltre 500.000 recensioni di più di 200.000 utenti, su 74.528 prodotti.

1.2 PULIZIA DEL DATASET

Prima di iniziare con l'effettivo sviluppo del progetto si è esplorato il dataset a disposizione. Inoltre sono state effettuate le seguenti correzioni:



1.3 RIMOZIONE DUPLICATI

Si sono considerate come righe duplicate quelle contenenti la stessa coppia di *(UserId, ProductId)*. Infatti su Amazon ogni utente può recensire una sola volta un prodotto, per questo motivo si sono tenute le recensioni più recenti.

Inoltre si sono eliminate le righe con stesso *(UserId, Score, Time, Text).* Sembra infatti che la recensione sia duplicata in automatico su più prodotti simili, ma ovviamente è anomalo che un utente compri prodotti equivalenti lo stesso giorno e dia recensioni uguali.

ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Summary	Text
B000PMJLJO	AZYMD9P9F9UZ6	W. Coombe	0	0	5	1239148800	Good Jerky	I like the peppered flavor a lot better than t...
B000GW46D4	AZYMD9P9F9UZ6	W. Coombe	0	0	5	1239148800	Good Jerky	I like the peppered flavor a lot better than t...
B000GW6786	AZYMD9P9F9UZ6	W. Coombe	0	0	5	1239148800	Good Jerky	I like the peppered flavor a lot better than t...

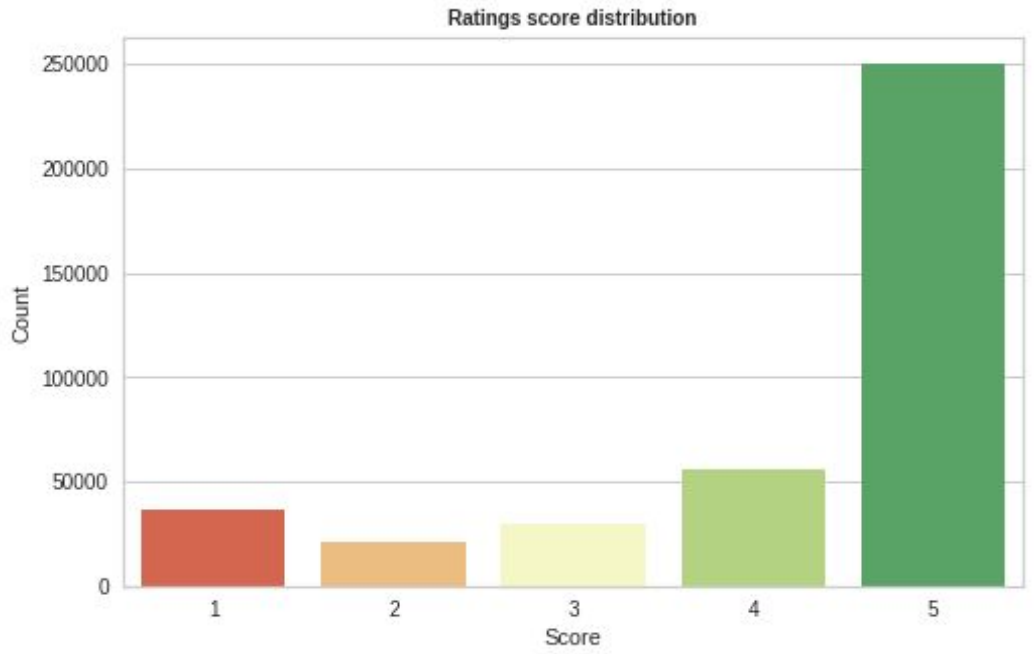
In totale si è eliminato il 30% dei dati, avendo quindi a disposizione 392.969 recensioni.

1.4 ANALISI TEMPORALE e SCORE

Dall'analisi temporale è emerso che le recensioni si riferiscono al periodo temporale 1999-2012. Tuttavia la maggior parte si riferisce agli anni dal 2006 in avanti e non si è notata una differenza testuale sostanziale tra le recensioni dei primi anni.

Per quanto riguarda la variabili score si è invece notato che:

1. È fortemente sbilanciata
2. Dal 2006 in poi si osserva un trend crescente del numero di recensioni, tuttavia le recensioni a 5 stelle hanno un andamento anomalo. Questo può essere dovuto a recensioni non verificate o false.



2. PRE PROCESSING

Lo step di preprocessing ha permesso di uniformare la rappresentazione degli indici del testo, in particolare si sono effettuati i seguenti step.

Normalization

- Lowercase
- Gestione delle abbreviazioni (not)
- Accenti
- Casi particolari (html, emoji)

Stopwords removal

- Lista predefinita
- Aggiunte parole di contesto (Amazon/Order)
- Preservato il not

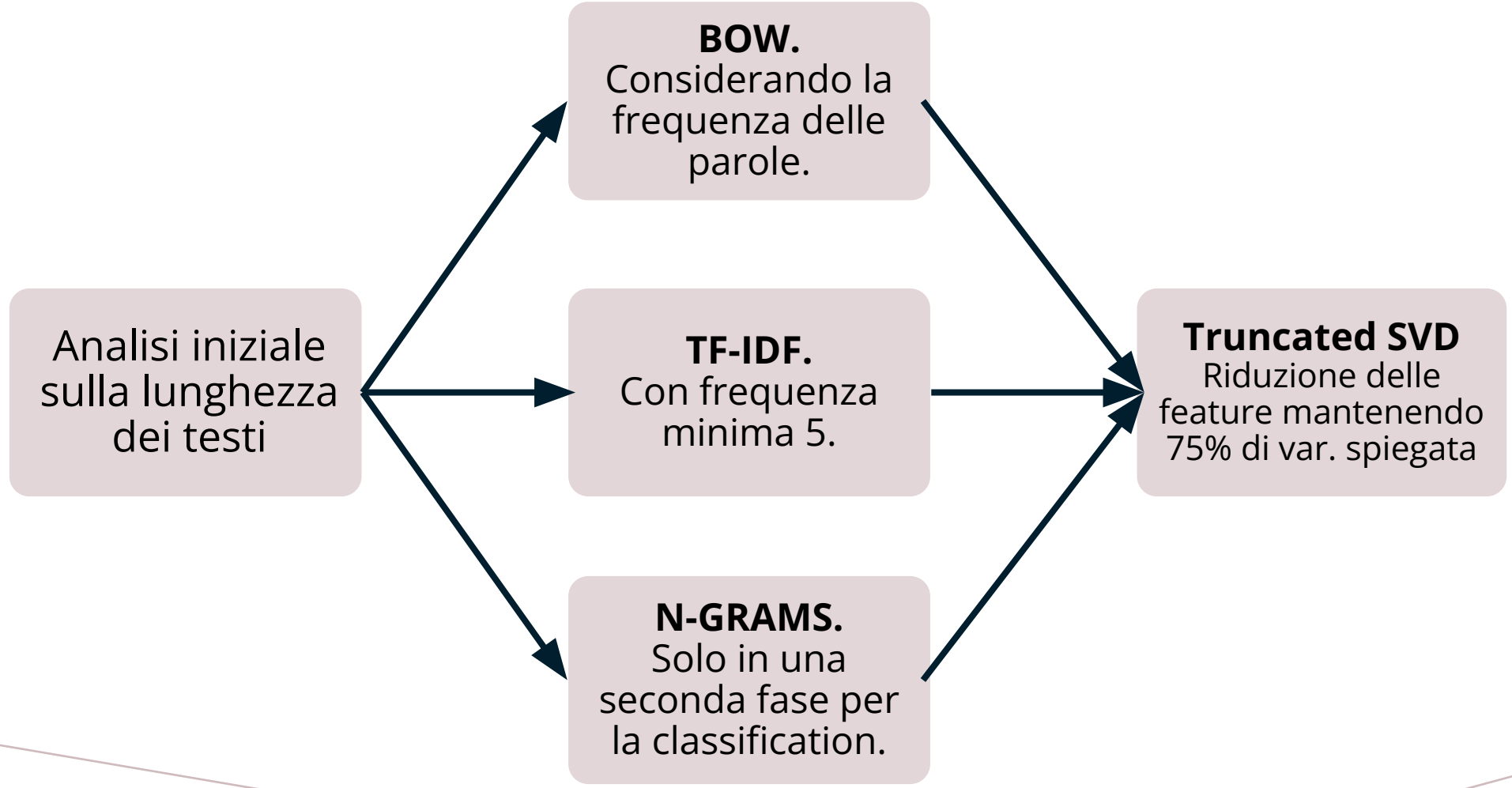
Tokenization e stemming

- Porter stemmer
- Oltre 90.000 token

```
<-------- Before remove stopwords -------->
Example1:  I do not like sour taste and this has a sour kind of taste which i don't like. The smell isn't that great either
Example2:  I just love it, and I am Not a major Indian cooking fan--just enough. Really, it mixes with anything you are doing like <a href="http://www.amazon.com/gp/product/B000FIXT2I"> ...

<-------- After remove stopwords -------->
Example1:  not like sour taste sour kind taste not like smell not great either
Example2:  love not major indian cooking fan enough really mixes anything like steamed brown rice bowl organic microwaveable ounce bowls pack use convenience not ...
```

3 TEXT REPRESENTATION



4. CLASSIFICATION

Il primo task considerato è stato *classification*, nello specifico l'obiettivo è classificare le recensioni in due macro categorie: positive (score ≥ 4) o negative (score ≤ 2). Per farlo i dati sono stati ulteriormente manipolati per:

1. **Convertire lo score** in binario. Si sono eliminate tutte le recensioni neutre (Score = 3) e si sono binarizzate le restanti.
2. Si è **bilanciato** il dataset eliminando recensioni dalla classe maggioritaria (Positive)

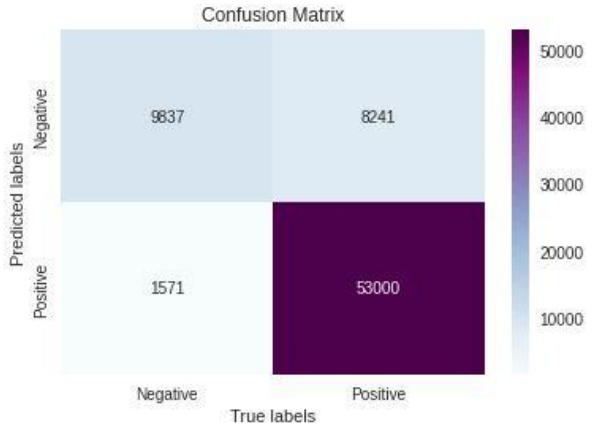
Il task di classificazione binaria è stato quindi svolto applicando diversi modelli:

- Regressione logistica.
- Light SVM.
- LGBM leggero.

4.1 REGRESSIONE LOGISTICA

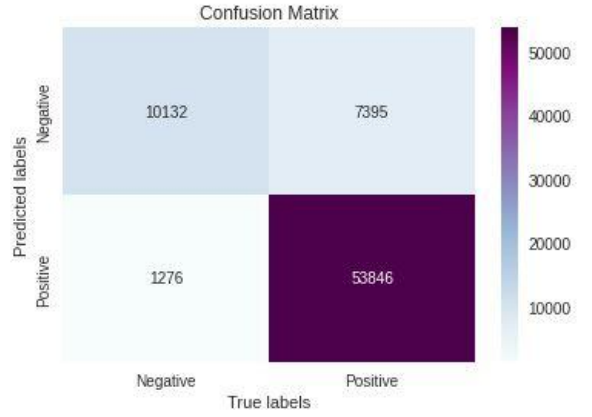
Classificatore lineare che assegna una probabilità compresa tra 0 e 1 per ogni classe, con somma di uno. Il valore di soglia predefinito, che è stato utilizzato in questo progetto, è ≥ 0,5. Impiega solo 13s nel caso di TF-IDF.

	precision	recall	f1-score	support
Negative	0.86	0.54	0.67	18078
Positive	0.87	0.97	0.92	54571
accuracy			0.86	72649
macro avg	0.86	0.76	0.79	72649
weighted avg	0.86	0.86	0.85	72649



Risultati con BOWs

	precision	recall	f1-score	support
Negative	0.89	0.58	0.70	17527
Positive	0.88	0.98	0.93	55122
accuracy			0.88	72649
macro avg	0.88	0.78	0.81	72649
weighted avg	0.88	0.88	0.87	72649

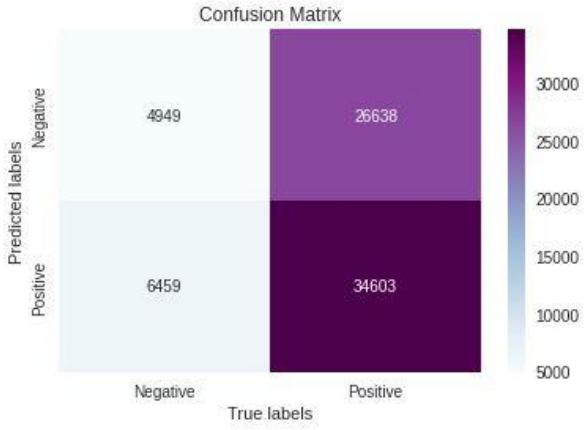


Risultati con TF-IDF

4.2 SVM

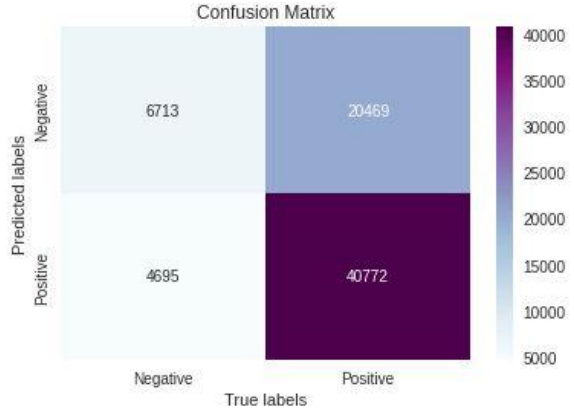
Il classico classificatore SVM non può essere applicato a causa della complessità dei dati, quindi è stata utilizzata una *versione approssimata* che consente tempi di esecuzione molto rapidi. Impiega infatti solo 3s per TF-IDF

	precision	recall	f1-score	support
Negative	0.43	0.16	0.23	31587
Positive	0.57	0.84	0.68	41062
accuracy			0.54	72649
macro avg	0.50	0.50	0.45	72649
weighted avg	0.51	0.54	0.48	72649



Risultati con BOWs

	precision	recall	f1-score	support
Negative	0.59	0.25	0.35	27182
Positive	0.67	0.90	0.76	45467
accuracy			0.65	72649
macro avg	0.63	0.57	0.56	72649
weighted avg	0.64	0.65	0.61	72649

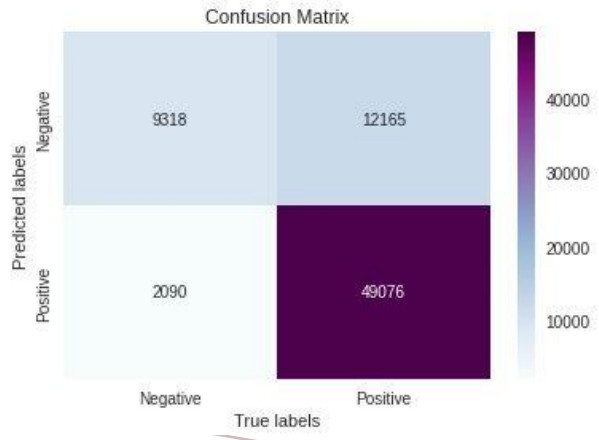


Risultati con TF-IDF

4.3 LGBM

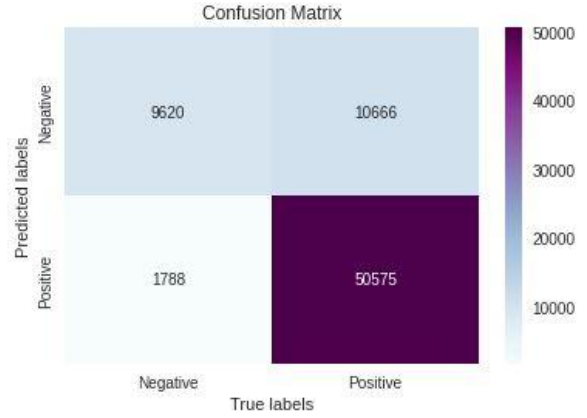
Questo è un framework di potenziamento del gradiente molto veloce, distribuito e ad alte prestazioni basato su algoritmi dell'albero decisionale. I risultati sono simili alla regressione logistica ma richiede *5min* su TF-IDF.

	precision	recall	f1-score	support
Negative	0.82	0.43	0.57	21483
Positive	0.80	0.96	0.87	51166
accuracy			0.80	72649
macro avg	0.81	0.70	0.72	72649
weighted avg	0.81	0.80	0.78	72649



Risultati con BOWs

	precision	recall	f1-score	support
Negative	0.84	0.47	0.61	20286
Positive	0.83	0.97	0.89	52363
accuracy			0.83	72649
macro avg	0.83	0.72	0.75	72649
weighted avg	0.83	0.83	0.81	72649



Risultati con TF-IDF

ISPEZIONE RISULTATI

Analizzando le classificazioni errate dei diversi classificatori precedenti è emerso che molte frasi *contenevano la parola 'not'*. Inoltre le parole più significative per disambiguare le recensioni sembrano essere randomiche e poco utili.

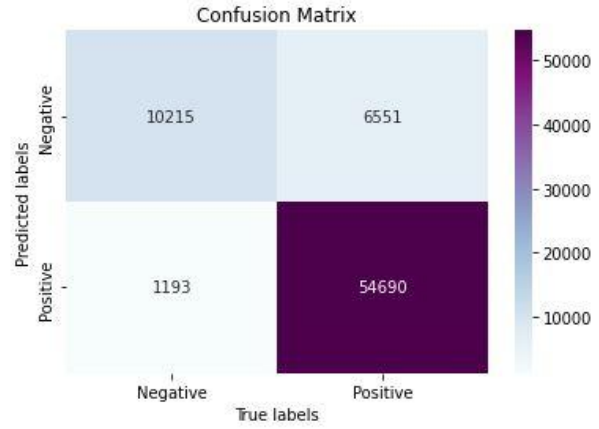
-1.8676	abl	1.2743	antisept
-1.5903	advertis	1.2472	acut
-1.4381	abhor	1.1295	aesthet
-1.2205	abund	1.1274	achiev
-1.0414	apiec	1.0715	aafco
-0.9961	apex	1.0427	aggrav
-0.9757	arginin	1.0349	abid
-0.9682	acidophilu	1.0152	adventuresom
-0.9650	accessori	1.0073	anti
-0.9465	apart	0.9997	alik
-0.9286	appet	0.9401	aerogarden
-0.9209	ambul	0.9266	aback
-0.9126	antidot	0.9168	absinth
-0.8698	asterisk	0.9008	adren

Si è quindi provato a utilizzare TF-IDF considerando coppie di token, 2-grams.

4.4 REGRESSIONE LOGISTICA (2-GRAMS)

I risultati sono più incoraggianti, aumentano tutti gli score e migliorano notevolmente le parole più significative.

	precision	recall	f1-score	support
Negative	0.90	0.61	0.73	16766
Positive	0.89	0.98	0.93	55883
accuracy			0.89	72649
macro avg	0.89	0.79	0.83	72649
weighted avg	0.89	0.89	0.89	72649



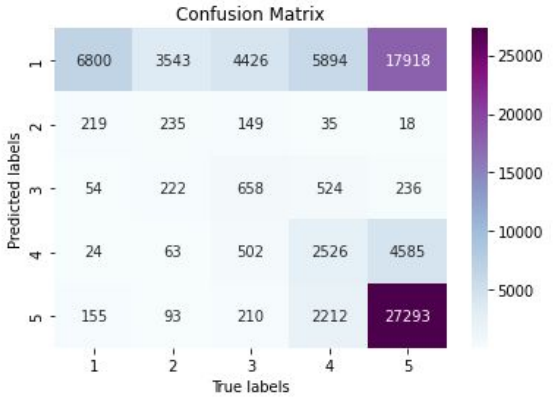
Risultati con TF-IDF

-13.8783	disappoint		13.5828 great
-10.3715	not		11.4268 delici
-9.7754	not recommend	11.3074	best
-8.8912	worst	10.6514	love
-8.6430	not good	9.8481	perfect
-8.5169	not buy	9.2003	good
-8.0209	not worth	8.8540	not disappoint
-7.6340	terribl	8.1196	excel
-7.4375	unfortun	7.2719	favorit
-7.3566	aw	7.1035	nice
-6.7330	horribl	6.9527	amaz
-6.5910	return	6.6659	happi

4.5 MULTICLASSE

Si è successivamente provato ad approcciare la problematica della classificazione multiclasse, mantenendo quindi il *punteggio reale* nell'intervallo [1,5]. Si è stato tentato un approccio moderno utilizzando una *rete neurale ricorrente*, in particolare LSTM con uno strato di embedding che mappa a vettori di dimensione 100.

	precision	recall	f1-score	support
1	0.18	0.94	0.30	7252
2	0.36	0.06	0.10	4156
3	0.39	0.11	0.17	5945
4	0.33	0.23	0.27	11191
5	0.91	0.55	0.68	50050
accuracy			0.48	78594
macro avg	0.43	0.38	0.30	78594
weighted avg	0.69	0.48	0.52	78594



Risultati con LSTM

I risultati non sono molto incoraggianti soprattutto per la classe 1 classificata quasi in modo randomico!

5. CLUSTERING

Il second task considerato è stato il *clustering*, nello specifico l'obiettivo iniziale è quello di raggruppare le recensioni in 5 diversi gruppi. L'idea è quindi trovare dei cluster rappresentativi per i diversi score. Successivamente, analizzando semanticamente i risultati, si è passati a clusterizzare in un numero maggiore di cluster cercando di massimizzare determinate metriche.

Gli algoritmi presi in considerazione per il clustering sono:

- K-means.
- Gerarchico agglomerativo.

5.1 RISULTATI

Analizzando i risultati dei due diversi cluster non si ottengono performance valide

```
No. of reviews in Cluster-0: 4841
No. of reviews in Cluster-1: 6215
No. of reviews in Cluster-2: 40903
No. of reviews in Cluster-3: 23890
No. of reviews in Cluster-4: 7266

Rand index           : 0.5981333172465243
Adjusted Mutual Info : 0.008051656398375396
Homogeneity          : 0.007286547678001394
Completeness         : 0.009163022529309783
V measure            : 0.00811775623228642
Fowlkes Mallows      : 0.2670176345603583
Silhouette           : 0.011880372905302599
```

Risultati k-means (k=5)

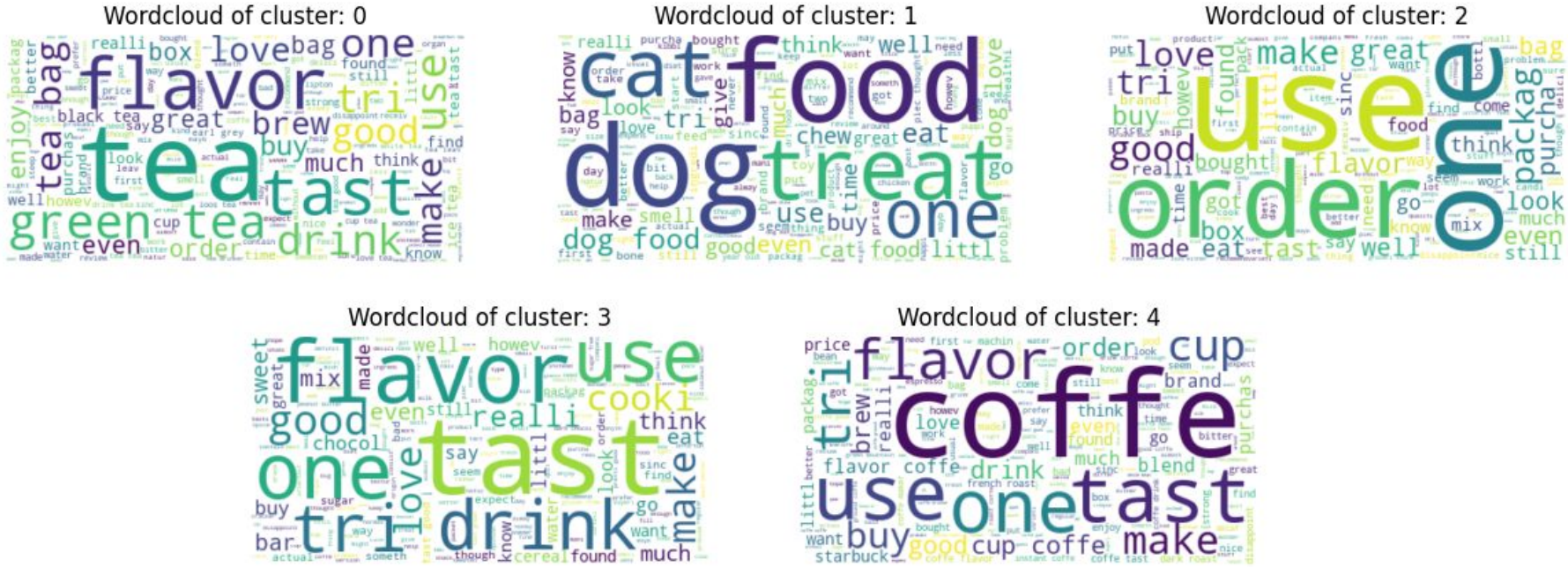
```
No. of reviews in Cluster-0: 11747
No. of reviews in Cluster-1: 641
No. of reviews in Cluster-2: 1534
No. of reviews in Cluster-3: 798
No. of reviews in Cluster-4: 280

Rand index           : 0.4230088583683357
Adjusted Mutual Info : 0.0011221649188717187
Homogeneity          : 0.0011680343902362516
Completeness         : 0.0023802389639944535
V measure            : 0.0015670725952446239
Fowlkes Mallows      : 0.3549546141524873
Silhouette           : 0.005335375457491175
```

Risultati hierarchical (k=5)

5.2 SEMANTICA

Analizzando la semantica dei cluster si è notato che esiste una possibile suddivisione in base agli argomenti ma anche in questo caso ci sono diverse ripetizioni e non si ottengono cluster unici.



Wordcloud k-means (k=5)

5.3 K-MEANS (K=9)

Si è dunque tentato di massimizzare la metrica della silhouette per trovare il numero ottimale di cluster che è risultato essere 10. L'obiettivo è formare dei cluster che dividano il contenuto delle recensioni.

Wordcloud k-means (k=10)

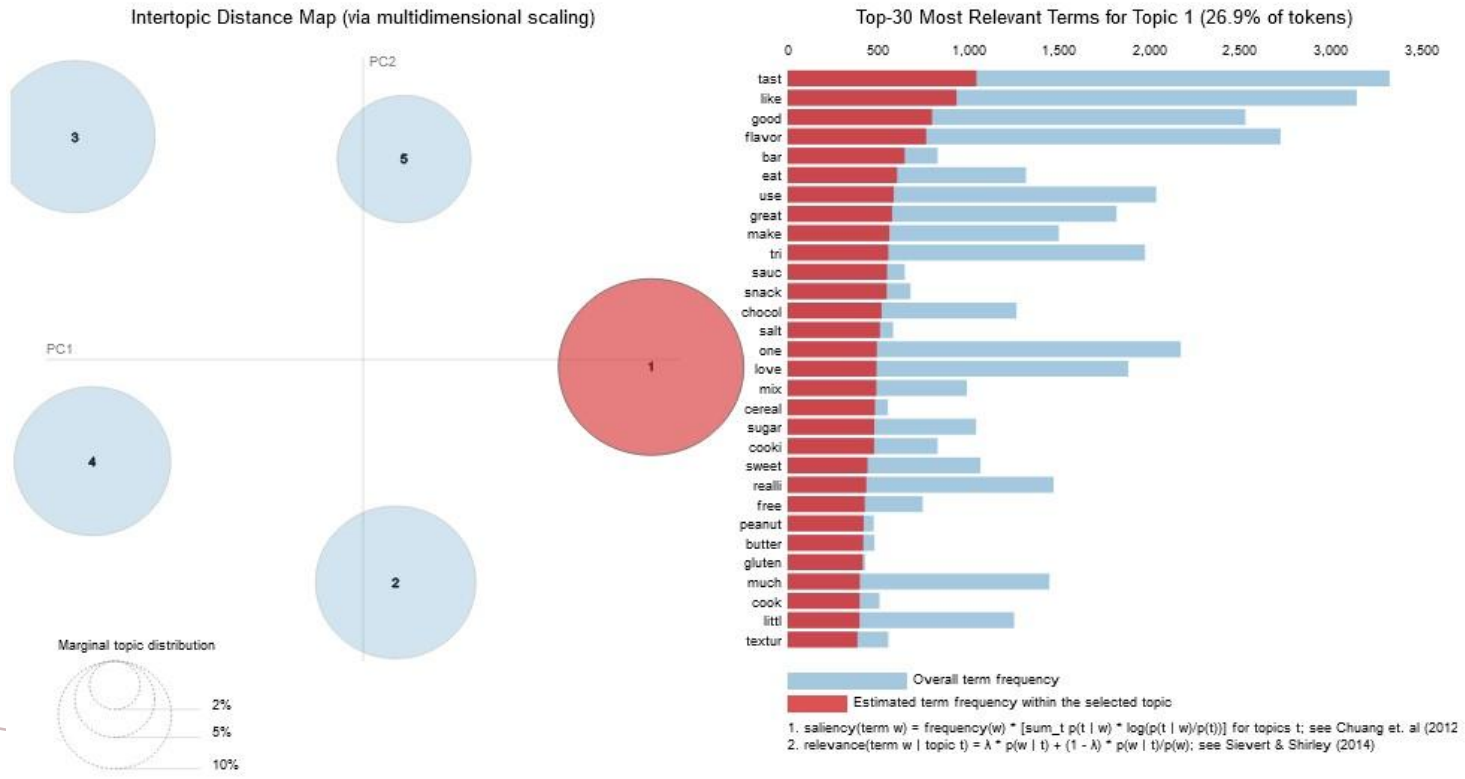
6. TOPIC MODELING

Visti i risultati, parzialmente incoraggianti, ottenuti a livello semantico con k-means si è provato ad approcciare il problema dell'estrazione dei contenuti (topic) presenti nelle diverse recensioni. L'analisi precedente ha evidenziato la presenza di almeno *5 topic* (animali, caffè, the, ordini, cioccolato/snacks).
Si è dunque cercato un numero di topic ≥ 5 che minimizzino la metrica della *perplexity*, la scelta ottimale per l'estrazione degli argomenti è risultata essere proprio quella di estrarre 5 topic differenti.

Per l'estrazione dei topic si è utilizzata la tecnica LDA.

6.1 RISULTATI

I risultati ottenuti con questa tecnica, seppur basici, sono incoraggianti e sembra che effettivamente sia possibile estrarre i topic presenti nelle recensioni.



In particolare:

- topic1 = generico/dolci
- topic2 = ordini, spedizioni
- topic3 = animali
- topic4 = the
- topic5 = caffè

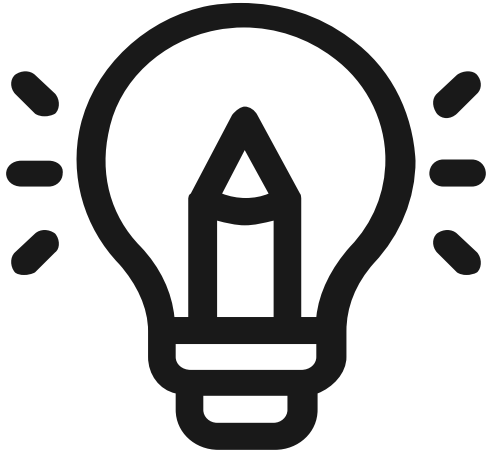
Estrazione topic LDA

7. CONCLUSIONI

In conclusione si può affermare che:

- La rappresentazione TF-IDF risulta essere più performante di BOW.
- La classificazione da dei buoni risultati ma ci sono problemi di recall con la classe negative.
- La classificazione multiclasse non dà risultati ideali, evidenziando i limiti del modello. Forse anche dovuto al fatto che non vi è una chiara distinzione testuale per i diversi voti.
- Il clustering non ha dato i risultati sperati e si è rivelato più complesso rispetto al task di classification. Anche per via della difficile valutazione.
- Il topic modeling, anche se approcciato velocemente, evidenzia come esistano diversi argomenti estraibili dalle recensioni con risultati soddisfacenti.

Complessivamente è possibile dire che i modelli di classificazione sono abili nel predire la classe binaria con una certa accuratezza ma è difficile creare un modello completo che, data una nuova recensione, restituisca automaticamente lo score della stessa.



DOMANDE?

GRAZIE PER L’ATTENZIONE

7. RIFERIMENTI BIBLIOGRAFICI

- G. Pasi and M. Viviani, “Dispense e slide del corso di Text Mining & Search” 2021.
- J. McAuley and J. Leskovec, “From amateurs to connoisseurs: Modeling the evolution of user expertise through online reviews”
- S. N. A. Project, “Amazon fine food reviews” 2017. https://www.kaggle.com/snap/amazon-fine-food-reviews