

# Data Analytics

## Sentiment Analysis on Food Reviews

Relazione del progetto di:  
Alice Romagnoli 829833  
Davide Pietrasanta 844824

Anno Accademico 2020-2021

# Indice

<b>1</b>	<b>Introduzione</b>	<b>2</b>
<b>2</b>	<b>Dataset</b>	<b>2</b>
<b>3</b>	<b>Preprocessing</b>	<b>3</b>
<b>4</b>	<b>Distribuzione parole</b>	<b>4</b>
<b>5</b>	<b>Approccio basato sul Lessico</b>	<b>6</b>
5.1	Normalizzazione di Afinn . . . . .	7
5.2	Legame tra Afinn e Score . . . . .	8
5.3	Valutazione delle performance di Afinn . . . . .	9
5.4	Varianza del valore di Afinn . . . . .	10
<b>6</b>	<b>Approccio supervisionato</b>	<b>11</b>
6.1	Predizione di score . . . . .	11
6.2	Predizione di sentiment . . . . .	14
<b>7</b>	<b>Top Reviewers</b>	<b>16</b>
<b>8</b>	<b>Altro Processing</b>	<b>19</b>
8.1	Recensioni Duplicate . . . . .	19
8.2	Prodotti Duplicati . . . . .	20
<b>9</b>	<b>Nuove Predizioni</b>	<b>20</b>
<b>10</b>	<b>Nuovo Dataset</b>	<b>25</b>
<b>11</b>	<b>Dashboard</b>	<b>25</b>
<b>12</b>	<b>Conclusioni</b>	<b>26</b>

# 1 Introduzione

Per questo progetto è stato utilizzato il dataset *Amazon Fine Foods* per eseguire una Sentiment Analysis sulle recensioni di prodotti alimentari comprati su Amazon.

Il nostro obiettivo principale è stato quello di predire attraverso l'analisi del testo delle recensioni lo score associato dagli utenti ai prodotti acquistati e in generale se si trattasse di recensioni positive, negative o neutre. Per score è stata intesa una valutazione numerica da 1 a 5, dove 1 è il grado di soddisfazione minimo, mentre 5 è il grado massimo.

La predizione dello score attraverso l'analisi delle recensioni risulta essere un'attività molto importante per ricavare informazioni nei casi in cui lo score assegnato al prodotto non sia presente o venga considerato in una fase iniziale di ricerca del prodotto, ma poi per indirizzare una scelta d'acquisto venga dato un peso importante anche alle recensioni scritte.

Quest'ultime permettono infatti di articolare meglio l'opinione che si ha del prodotto e di evidenziarne eventuali aspetti positivi e negativi, a cui utenti diversi potrebbero dare diversa importanza.

Da ultimo, è bene tener conto del fatto che recensioni scritte e score assegnato possono differire significativamente, in quanto pur sempre frutto dell'opinione soggettiva dell'utente.

Per effettuare questo tipo di analisi sono stati principalmente utilizzati due approcci, uno basato sul lessico e uno di apprendimento supervisionato, applicati al testo delle recensioni dopo una fase di preprocessing.

In seguito, sono state valutate le performance di entrambi e sono state condotte altre analisi per meglio comprendere il dataset e i risultati ottenuti.

Infine, è stato eseguito un'ulteriore processing del testo delle recensioni e sono stati ricalcolati i risultati dei due approcci principali.

## 2 Dataset

Il dataset utilizzato, *Amazon Fine Foods*, è un dataset di 35172 records, rappresentanti singole recensioni e contenenti i valori di quattro colonne: `productid`, `userid`, `score`, `text`.

Le prime due colonne contengono quindi gli identificati per i prodotti recensiti e per gli utenti che effettuano recensioni, la colonna `score` contiene un valore da 1 a 5 assegnato ad ogni prodotto dall'utente che l'ha acquistato, infine la colonna `text` presenta il testo della recensione scritta.

Come prima cosa si è verificato che il dataset non contenesse missing values, quindi, per capirne meglio la composizione, sono state effettuate delle analisi sulla distribuzione del numero di recensioni.

Queste analisi sono state effettuate prima secondo `product_id` (quindi per prodotto) e poi secondo `user_id` (per utente).

Quello che si rileva nel raggruppamento per prodotto, visibile in Figura 1, è che la maggior parte dei prodotti hanno ricevuto un numero di recensioni minore di 10.

Infatti, ogni prodotto riceve un numero di recensioni molto variabile: si passa da un massimo di 632 a un minimo di 1 recensione.

Dal grafico, notiamo inoltre che la maggior parte dei prodotti hanno ricevuto una sola recensione e che all'aumentare del numero di reviews per prodotto cala drasticamente il numero di prodotti con tale quantità di recensioni.



Figura 1: Istogramma che rappresenta il numero di prodotti che hanno ricevuto un certo numero di recensioni

Andando invece a raggruppare per utente (Figura 2), notiamo subito che il numero massimo di reviews è 30 e il minimo 1; in generale ogni utente scrive mediamente una recensione ( $mean = 1.195$ ). Questo grafico permette di osservare come la maggior parte degli utenti abbia effettuato una sola recensione (25563) e come all'aumentare del numero di recensioni considerate ci siano sempre meno utenti che le abbiano effettuate.

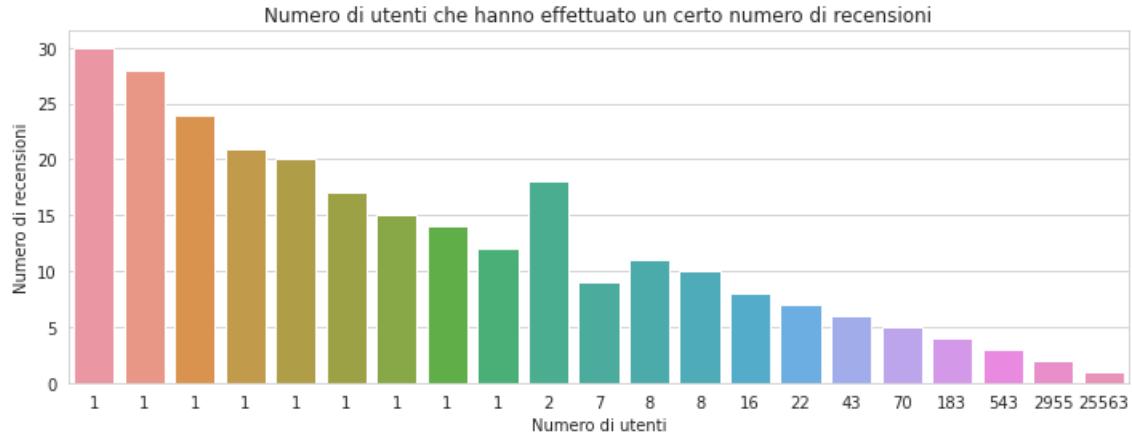


Figura 2: Istogramma che rappresenta il numero di utenti che hanno effettuato un certo numero di recensioni

### 3 Preprocessing

Come accennato nel Capitolo 1, prima di applicare i due approcci è stata effettuata una fase di preprocessing, volta a snellire le recensioni di tutte quelle parole che non risultano essere significative.

Questa operazione può infatti rivelarsi utile sia per migliorare le performance e le capacità predittive dei due approcci, sia per velocizzarne i tempi.

NLTK (Natural Language Toolkit) è una delle piattaforme più usate per lavorare con dati scritti in linguaggio naturale su Python.

Questa libreria è stata pertanto utilizzata per eseguire una *tokenizzazione* del testo delle recensioni e successivamente per identificare *stopwords* e segni di punteggiatura.

La *tokenization* risulta solitamente essere il primo passo quando si effettua NLP (Natural Language Processing) e consiste nel suddividere un testo o una porzione di testo in tanti tokens.

I tokens sono solitamente parole singole, ma in alcuni casi specifici e dove lo si ritenga opportuno potrebbero essere composti da più parole (almeno in lingue come l'inglese).

Nel progetto per esempio abbiamo mantenuto in un unico token parole collegate da un simbolo - o ', tag HTML e URLs.

I token, creati con questo processo di *tokenization*, vengono quindi utilizzati come input per altri tipi di analisi o attività.

Nella Tabella 1 sono riportate le 10 parole più frequenti nel dataset subito dopo il processo di *tokenization*, mentre nella Tabella 2 quelle più frequenti dopo la rimozione di *stopwords* e segni di punteggiatura.

Parola	Frequenza
.	171618
the	103443
,	89733
i	84916
and	71538
a	66568
to	55943
it	51869
of	43024
is	40793

Tabella 1: Frequenza iniziale delle parole dopo la tokenizzazione

Parola	Frequenza
 	28801
not	18592
like	14289
good	11429
coffee	9878
great	9849
taste	9612
one	9588
product	8344
flavor	8269

Tabella 2: Frequenza delle parole dopo aver rimosso stopwords e punteggiatura

Alla lista di *stopwords* e segni di punteggiatura fornita dalla libreria NLTK ne sono state rimosse altre tra cui i tag tipici di HTML (si pensa infatti che il dataset sia stato estratto usando una qualche tecnica di *scraping*), i numeri e i token costituiti da un solo carattere e i costrutti tipici della lingua inglese che vedono il pronome “I” seguito da un apostrofo e una forma contratta di un verbo.

Sono stati poi nuovamente osservati i token più frequenti al fine di rilevare ulteriori token poco significativi o anomali (Tabella 3).

Parola	Frequenza
not	18592
like	14289
good	11429
coffee	9878
great	9849
taste	9612
one	9588
product	8344
flavor	8269
love	7360
would	7162
tea	6478
food	6265
get	6057
really	5735
much	5190
time	4802
use	4711
amazon	4676
also	4595

Tabella 3: Frequenza delle parole dopo la rimozione di nuove stopwords

Una volta processato il testo delle recensioni del dataset è possibile iniziare a fare alcune analisi.

## 4 Distribuzione parole

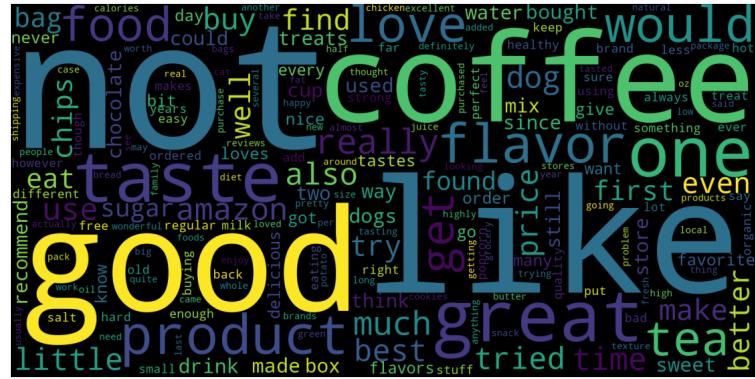
La distribuzione delle parole all’interno delle recensioni del dataset è stata valutata al termine della fase di preprocessing e quindi su parole in qualche modo significative.

In particolare, ci si è concentrati sulle parole con maggior frequenza.

Le 10 parole più frequenti all’interno del dataset risultano essere quelle riportate nella tabella in Figura 3.

Parola	Frequenza
not	18592
like	14289
good	11429
coffee	9878
great	9849
taste	9612
one	9588
product	8344
flavor	8269
love	7360

(a) 10 parole più frequenti con relativa frequenza



(b) Wordcloud delle parole più frequenti

Figura 3: Distribuzione delle parole all'interno del dataset

È stato anche riportato un insieme più ampio di parole molto frequenti, attraverso la rappresentazione visiva in Figura 3b.

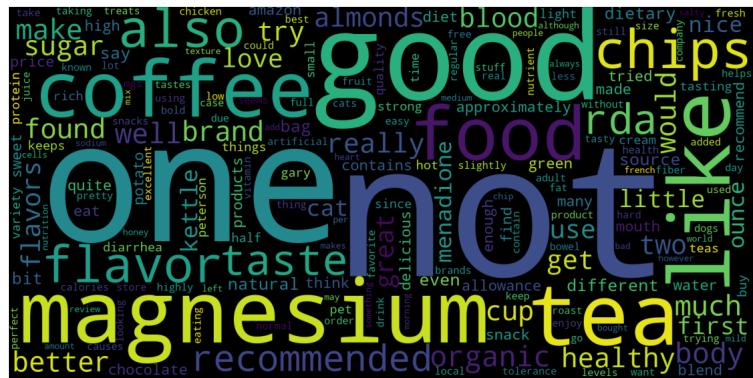
Come si può notare le parole più frequenti risultano essere aggettivi volti a descrivere i prodotti, nomi di prodotti alimentari e altre parole sempre relative all'area semantica degli alimenti e dello shopping online.

Non avendo trovato nulla che si allontanasse da ciò che ci si poteva aspettare, è stato tentato un nuovo approccio, andando a vedere se le parole utilizzate dalle persone che effettuano tante recensioni differiscono da quelle delle persone che hanno effettuato solo una recensione.

In Figura 4 sono quindi state riportate le stesse informazioni usate per valutare le parole più frequenti nel dataset ma considerando solo quelle utilizzate dallo 0.05% degli utenti che recensiscono di più, cioè dai 19 utenti che hanno effettuato più di 10 recensioni, chiamati top\_reviewers (I top\_reviewers verranno spiegati in modo più dettagliato nel capitolo 7).

Parola	Frequenza
not	255
one	204
good	192
magnesium	182
tea	175
like	171
coffee	145
food	130
chips	117
also	115

(a) 10 parole più frequenti con relativa frequenza



(b) Wordcloud delle parole più frequenti

Figura 4: Distribuzione delle parole utilizzate dai top reviewers

Mentre in Figura 5 sono riportate le informazioni relative ai moltissimi utenti che hanno effettuato una sola recensione, chiamati worst\_reviewers .

Parola	Frequenza
not	12290
like	9170
good	7780
great	7228
one	6446
product	6445
coffee	6275
taste	6237
love	5140
would	4889

(a) 10 parole più frequenti con relativa frequenza



(b) Wordcloud delle parole più frequenti

Figura 5: Distribuzione delle parole utilizzate dai worst reviewers

Com'è possibile notare, tutte le distribuzioni di parole considerate risultano essere molto simili e con la maggior parte delle parole in comune.

La somiglianza tra le parole utilizzate nell'intero dataset e quelle dei `worts_reviewers` non sorprende, in quanto quest'ultimi costituiscono il 73% degli utenti.

Quello che si può osservare è che benché costituiscano un gruppo molto ristretto e accomunato dall'attiva partecipazione al recensire, i `top_reviewers`, non presentino evidenti differenze nella scelta delle parole utilizzate, attenendosi ancora strettamente all'area semantica degli alimenti e dello shopping online.

## 5 Approccio basato sul Lessico

Il primo approccio utilizzato per predirre lo score è un approccio basato sul lessico.

In particolare, è stato utilizzato il lessico *Afinn*, che consiste in una lista di termini in inglese a cui è stato assegnato manualmente un punteggio da -5 (estremamente negativo) a +5 (estremamente positivo), volto a indicarne la polarizzazione. Un punteggio di 0 (nullo o neutro) rappresenta un vocabolo non trovato nella lista dei termini considerati da Afinn.

Questo punteggio è stato assegnato a tutte le parole delle recensioni, ed è così stato possibile calcolare un punteggio finale per ogni recensione.

I valori massimi e minimi riscontrati nelle recensioni sono stati rispettivamente 98 e -30. Questo indica che le recensioni positive tendono a raggiungere picchi più alti.

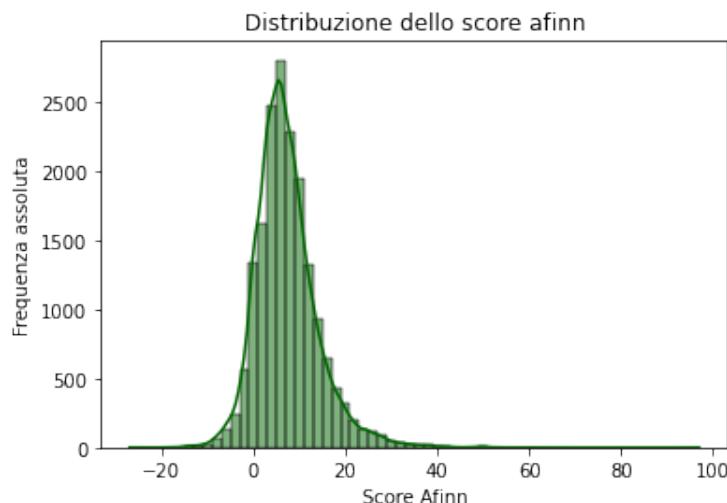


Figura 6: Distribuzione dell'afinn score all'interno del dataset

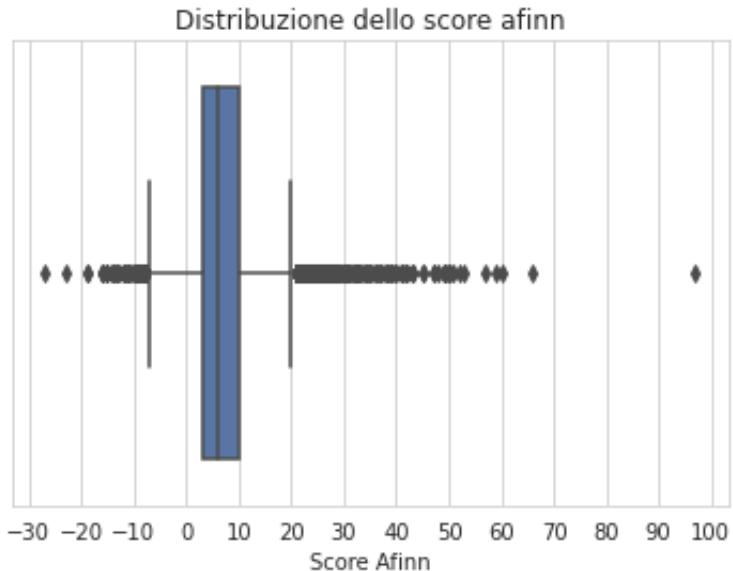


Figura 7: Distribuzione dell’afinn score all’interno del dataset

Come si può notare nelle Figure 6 e 7, il punteggio Afinn associato alle recensioni, si concentra maggiormente attorno a valori positivi ma non alti.

La distribuzione presenta un andamento normale, registrando il 50% delle recensioni nell’intervallo di valori [3, 10].

## 5.1 Normalizzazione di Afinn

Per poter confrontare al meglio il valore riportato da Afinn e lo score assegnato dagli utenti è stato necessario effettuare una normalizzazione del valore di Afinn.

Occorreva infatti passare da un valore potenzialmente compreso tra  $-\infty$  e  $+\infty$  a uno che fosse compreso nel range [1, 5], lo stesso range coperto dallo Score, senza alterarne però il significato.

Per operare questa normalizzazione sono stati presi in considerazione diversi approcci:

- *Opzione 1:* Normalizzare rispetto ai valori massimi e minimi riportati da Afinn.

Questo tipo di normalizzazione è una normalizzazione rispetto al corpus e presenta principalmente tre problemi: i valori di massimo e minimo di Afinn vengono trattati in maniera uguale (mentre il valore massimo 97 dovrebbe essere molto più positivo di quanto non sia negativo il valore minimo -30), ai valori rilevati da Afinn come neutri verrebbe assegnato uno score pari a 2, anziché il valore esatto 3 (i valori così trovati risultano non essere centrati) e in caso di presenza di *outliers* si tende a comprimere il range su valori centrali.

- *Opzione 2:* Normalizzare dividendo il valore di Afinn per il numero di parole nella frase.

Questa normalizzazione non presenta i problemi dell’Opzione 1, ma porta ad ottenere valori di score diversi in base al numero di parole presenti nelle recensioni.

In particolare, a parità di valore di Afinn, una frase con un maggior numero di parole avrà uno score più basso di un’altra con un minore numero di parole.

Questo succede perché molte parole non vengono riconosciute da Afinn o non presentano un’evidente polarizzazione: pertanto non dovrebbero partecipare all’assegnazione dello score.

Il valore di Afinn normalizzato dovrebbe essere infatti indipendente dal numero di parole totali presenti nelle recensioni.

- *Opzione 3:* Normalizzare effettuando una equal-depth (frequency) partitioning dei valori assumibili da Afinn.

Questa opzione presenta il problema di non essere centrata, quindi di non far corrispondere il valore 0 di Afinn con il valore 3 di score.

- *Opzione 4:* Normalizzare dividendo il valore di Afinn per il numero di parole riconosciute e valutate da Afinn nella frase.

Questa normalizzazione risulta essere simile all'Opzione 2, ma in questo caso si effettua una divisione per il numero di parole presenti nella frase riconosciute da Afinn.

Questa soluzione permette di risolvere appieno i problemi dell'Opzione 2, e per questo motivo è stata applicata al dataset, ottenendo un valore di Afinn normalizzato, chiamato `afinn_norm`.

I risultati ottenuti al termine della normalizzazione ricadevano nell'intervallo [-5, 5] e sono stati poi facilmente ricondotti all'intervallo desiderato ([1,5]).

In seguito, i valori di `afinn_norm` sono stati arrotondati a valori interi, in modo che fosse finalmente possibile confrontarli con lo score.

## 5.2 Legame tra Afinn e Score

È quindi stato valutato il legame che sussiste tra lo score normalizzato calcolato con Afinn, `afinn_norm`, e quello assegnato dagli utenti, `score`.

Per farlo, come prima cosa è stata calcolata la differenza tra questi due valori per ogni recensione.

Quello che si nota, grazie alla Figura 8, è che tendenzialmente essi differiscono di un punto (nel 61% dei casi) e il 25% delle recensioni differisce di due punti o più.

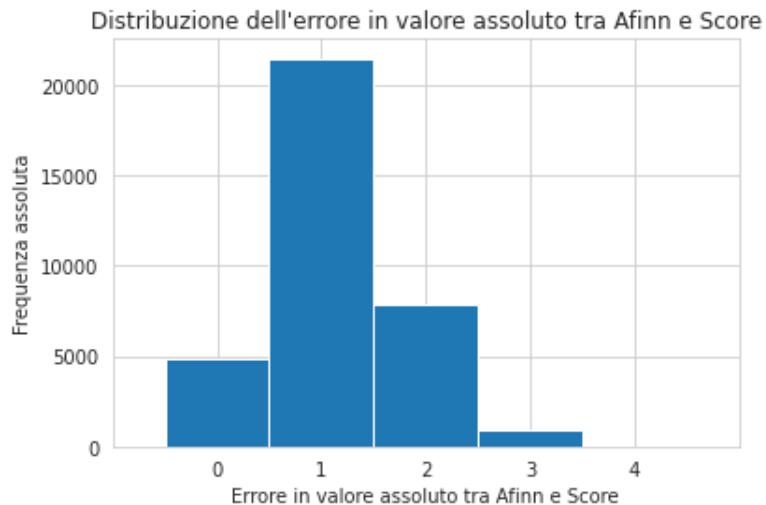


Figura 8: Distribuzione dell'errore in valore assoluto tra Afinn e Score

Successivamente siamo andati ad analizzare questa differenza più nello specifico (Figura Figura 9), osservando che generalmente è `score` che ha un valore più alto di un punto rispetto ad `afinn_norm`.

Il fatto che `score` assuma un valore più alto di 1 rispetto a `afinn_norm` accade nel 52% dei casi totali e nell'86% dei casi di differenza di un punto.

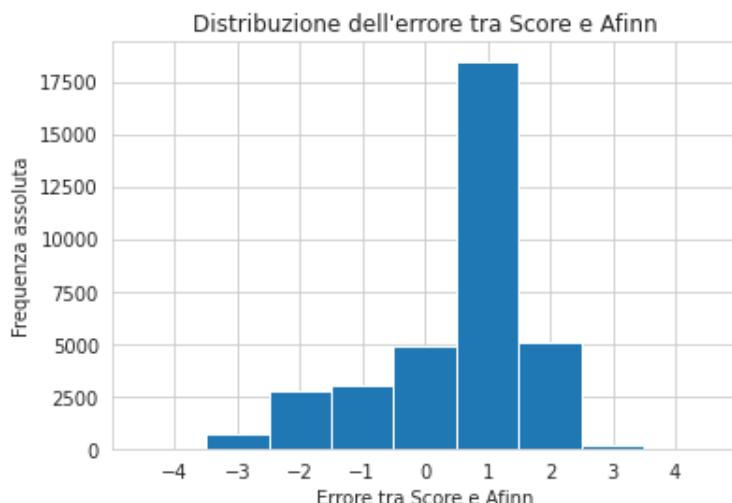


Figura 9: Distribuzione dell'errore tra Afinn e Score

Da ultimo, come si può osservare in Figura 10, nel 68% dei casi `score` risulta essere maggiore di `afinn_norm` e solo nel 14% dei casi coincidono.

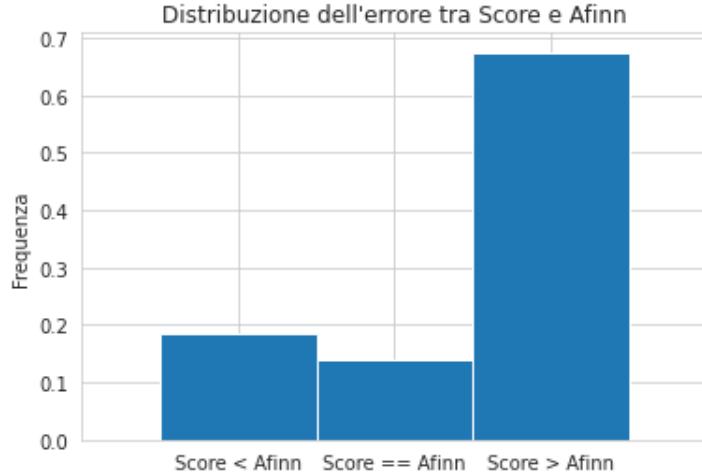


Figura 10: Relazione tra Afinn e Score

Da quest'analisi sul legame dei due valori, si può dedurre che, almeno per quanto riguarda il dataset in esame, gli utenti tendono ad essere più generosi con lo score assegnato al prodotto e più critici nelle corrispondenti recensioni scritte.

Data l'altissima presenza di recensioni valutate con punteggio 5 all'interno del dataset, si può pensare che le recensioni scritte siano state principalmente utilizzate per esplicitare i pochi e quasi trascurabili difetti dei prodotti acquistati.

### 5.3 Valutazione delle performance di Afinn

Dopo aver indagato il legame tra `afinn_norm` e `score`, sono state ricavate le matrici di confusione e le misure di performance ottenute utilizzando `afinn_norm` per predire il valore di `score` e di `sentiment`.

Con `sentiment` intendiamo la polarità di una recensione, ovvero il fatto di poter essere positiva, negativa o neutra.

I risultati ottenuti nella predizione di `score` sono osservabili in Figura 11 e quelli ottenuti nella predizione di `sentiment` in Figura 12.

		Afinn				
		1	2	3	4	5
Score	1	1	451	2058	701	7
	2	0	107	1177	719	3
	3	0	104	1358	1393	5
	4	1	63	1666	3332	9
	5	1	175	5038	16695	108

(a) Matrice di confusione

	precision	recall	f1-score	support
1.0	0.33	0.00	0.00	3218
2.0	0.12	0.05	0.07	2006
3.0	0.12	0.47	0.19	2860
4.0	0.15	0.66	0.24	5071
5.0	0.82	0.00	0.01	22017
accuracy			0.14	35172
macro avg	0.31	0.24	0.10	35172
avg	0.58	0.14	0.06	35172

(b) Misure di performance

Figura 11: Predizione di Score con Afinn

		Afinn		
		negativo	neutro	positivo
Score	negativo	559	3235	1430
	neutro	104	1358	1398
	positivo	240	6704	20144

(a) Matrice di confusione

	precision	recall	f1-score	support
negativo	0.62	0.11	0.18	5224
neutro	0.12	0.47	0.19	2860
positivo	0.88	0.74	0.80	27088
accuracy			0.63	35172
macro avg	0.54	0.44	0.39	35172
avg	0.78	0.63	0.66	35172

(b) Misure di performance

Figura 12: Predizione di Sentiment con Afinn

Osservando la Figura 11 si può subito notare come i valori di *recall* siano bassissimi per gli score 1, 2 e 5: questo significa che Afinn tende a non assegnare i valori di quella classe alle recensioni osservate e a prediligere i valori 3 e 4.

Ciò è avvalorato dalla precedente dimostrazione di come `afinn_norm` tendenzialmente assuma valori leggermente più bassi di `score` (bisogna infatti tener conto che `score` assume valore 5 nel 63% dei casi).

Un'altra misura che spicca è la *precision* della classe 5 di `score`, che risulta essere molto alta, rispetto a tutte le altre; considerando questo aspetto insieme al fatto che il relativo valore di *recall* è bassissimo, possiamo dedurre che Afinn non assegna quasi mai il valore 5 alle recensioni, ma le poche volte che lo fa risulta essere una predizione corretta.

L'accuratezza di questa previsione è comunque infima. Si è pensato di valutare la predizione della variabile `sentiment`.

Come si può notare in Figura 12, questa predizione presenta un'accuratezza generale nettamente migliore alla precedente (anche se ancora non soddisfacente).

Altre due differenze importanti rispetto a prima sono una miglior precisione nella predizione di uno score negativo (il valore negativo viene predetto meno volte delle effettive occorrenze nel dataset, ma quando viene predetto nel 62% dei casi viene predetto correttamente) e un notevole aumento della *recall* della classe positiva, che insieme al valore di *precision*, va garantire una buona predizione per questa classe.

In conclusione, Afinn non risulta essere adatto al compito di predire `score`, ma garantisce una buona previsione delle recensioni positive.

## 5.4 Varianza del valore di Afinn

Come ulteriore analisi è stata valutata la varianza del valore di Afinn associato alle singole parole. Questo perché, soprattutto in alcuni casi specifici, potrebbe essere utile capire la variabilità dei valori Afinn all'interno di una stessa frase.

In particolare, sono state analizzate le recensioni con valore di `afinn_norm` uguale a 3.

Queste vengono classificate da Afinn come recensioni neutre, ma questa classificazione potrebbe essere dovuto a due aspetti diversi: nessuna delle parole all'interno della recensione fa parte del lessico di Afinn, e quindi non viene riconosciuta (ciò determina un'assegnazione del valore neutro alla recensione) oppure alcune parole all'interno della frase vengono riconosciute da Afinn ma i loro valori sono tali da rendere neutro il valore complessivo della frase.

Ciò che si osserva andando a vedere il dataset è che la maggior parte delle recensioni classificate neutre da Afinn, in realtà risultano avere una polarizzazione, sia negativa che positiva, ma che è possibile cogliere solo leggendo l'intera recensione.

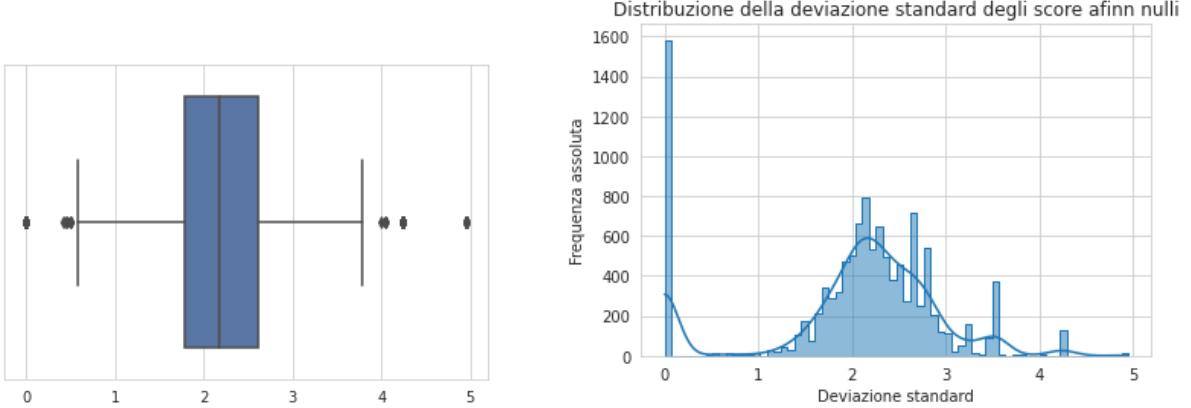
In particolare, il 4% delle recensioni del dataset contiene parole che non trovano alcun riscontro all'interno del lessico di Afinn, non permettendo così una corretta classificazione.

Andando ad approfondire l'analisi di questo sottoinsieme di recensioni, è stato notato che il motivo risiede nel fatto che in esse non sono presenti parole polarizzate, ma è il senso della frase a determinarne il `sentiment`. Questo ovviamente non può essere rilevato da Afinn, in quanto il calcolo del suo valore si basa sulle singole parole e non su gruppi di parole o sul contesto.

In Figura 13a è comunque possibile osservare che benché presenti come outliers, i casi neutri con varianza 0 (quelli in cui nessuna parola viene riconosciuta da Afinn) risultano essere una minoranza.

Tendenzialmente infatti, le recensioni con valore di Afinn uguale a 3 riportano una varianza poco superiore a 2, in quanto ragionevolmente descrivono sia aspetti positivi che negativi.

Come si può invece meglio osservare nella Figura 13b, i casi neutri a varianza zero non sono comunque da trascurare e si potrebbe pensare in futuro di trovare un modo per gestirli.



(a) Boxplot della distribuzione della varianza degli score Afinn neutri

(b) Istogramma della distribuzione della varianza degli score Afinn neutri

Figura 13: Distribuzione della varianza degli score Afinn neutri

## 6 Approccio supervisionato

Il secondo approccio utilizzato è un approccio supervisionato di Machine Learning volto a predire lo score. Viene infatti utilizzato un modello di regressione logistica, in quanto modello di regressione non lineare adatto per variabili dipendenti binarie.

Nel nostro caso infatti, il modello viene applicato a una *bag of words*, ovvero una matrice contenente per ogni recensione e per ogni parola l'informazione se quest'ultima sia presente o meno nella recensione considerata. La *bag of words* è stata creata in modo da contenere 5000 parole ed è stata applicata alle recensioni dopo il *preprocessing*, in modo da contenere il più possibile parole significative.

L'apprendimento è stato condotto effettuando una *cross-validation* con 10 folds applicata all'intero dataset, al fine di sfruttare appieno tutta l'informazione nota.

### 6.1 Predizione di score

Nel predire lo score, attraverso il modello di apprendimento supervisionato descritto, sono state computate le matrici di confusione e le misure di performance per ogni iterazione della *cross-validation*. I risultati ottenuti sono stati riportati nelle tabelle nelle Figure 14 e 15.

Si può così notare che la classe meglio predetta risulta sempre essere quella relativa allo score 5 (ciò è dovuto al fatto che il numero di recensioni con score 5 costituiscono il 63% del dataset), seguita dalla classe relativa allo score 1.

Come conseguenza alla scarsa capacità predittiva delle classi intermedie (2, 3, e 4) e ai risultati comunque non soddisfacenti per la classe 1, l'accuratezza in generale non supera il 70%.

		Predizione				
		1	2	3	4	5
Score	1	188	31	29	11	64
	2	38	54	33	20	51
	3	29	34	80	37	103
	4	8	21	53	124	288
	5	47	19	50	137	1969

(a) Iterazione 1

		Predizione				
		1	2	3	4	5
Score	1	161	44	23	14	65
	2	38	51	27	14	59
	3	38	34	77	53	89
	4	12	12	50	128	287
	5	34	25	50	131	2002

(b) Iterazione 2

		Predizione				
		1	2	3	4	5
Score	1	207	38	30	13	67
	2	34	51	26	19	46
	3	27	20	77	55	102
	4	11	18	45	129	289
	5	35	30	37	138	1973

(c) Iterazione 3

		Predizione				
		1	2	3	4	5
Score	1	186	35	23	17	74
	2	44	41	34	23	54
	3	30	31	79	55	111
	4	18	12	29	115	321
	5	39	23	45	135	1943

(e) Iterazione 5

		Predizione				
		1	2	3	4	5
Score	1	152	32	27	14	51
	2	61	48	31	18	58
	3	24	25	74	43	103
	4	20	11	40	132	334
	5	40	20	50	132	1977

(f) Iterazione 6

		Predizione				
		1	2	3	4	5
Score	1	171	36	27	17	74
	2	55	45	35	21	67
	3	28	52	69	38	102
	4	16	12	54	133	297
	5	42	16	41	137	1932

(g) Iterazione 7

		Predizione				
		1	2	3	4	5
Score	1	181	38	22	15	71
	2	39	56	45	19	39
	3	26	28	83	39	105
	4	19	17	53	131	298
	5	36	26	43	136	1952

(i) Iterazione 9

		Predizione				
		1	2	3	4	5
Score	1	184	28	28	13	73
	2	39	58	33	25	49
	3	25	27	81	47	85
	4	15	18	45	123	327
	5	41	21	47	133	1952

(j) Iterazione 10

Figura 14: Matrici di confusione generate ad ogni iterazione della cross-validation nella predizione di score

	precision	recall	f1-score	support
1.0	0.61	0.58	0.59	323
2.0	0.34	0.28	0.30	196
3.0	0.33	0.28	0.30	283
4.0	0.38	0.25	0.30	494
5.0	0.80	0.89	0.84	2222

accuracy			0.69	3518
macro avg	0.49	0.46	0.47	3518
avg	0.66	0.69	0.67	3518

(a) Iterazione 1

	precision	recall	f1-score	support
1.0	0.57	0.52	0.55	307
2.0	0.31	0.27	0.29	189
3.0	0.34	0.26	0.30	291
4.0	0.38	0.26	0.31	489
5.0	0.80	0.89	0.84	2242

accuracy			0.69	3518
macro avg	0.48	0.44	0.46	3518
avg	0.66	0.69	0.67	3518

(b) Iterazione 2

	precision	recall	f1-score	support
1.0	0.66	0.58	0.62	355
2.0	0.32	0.29	0.31	176
3.0	0.36	0.27	0.31	281
4.0	0.36	0.26	0.30	492
5.0	0.80	0.89	0.84	2213

accuracy			0.69	3517
macro avg	0.50	0.46	0.48	3517
avg	0.66	0.69	0.67	3517

(c) Iterazione 3

	precision	recall	f1-score	support
1.0	0.58	0.53	0.56	319
2.0	0.32	0.28	0.30	193
3.0	0.33	0.25	0.28	307
4.0	0.36	0.25	0.29	515
5.0	0.79	0.90	0.84	2183

accuracy			0.68	3517
macro avg	0.48	0.44	0.45	3517
avg	0.64	0.68	0.66	3517

(d) Iterazione 4

	precision	recall	f1-score	support
1.0	0.59	0.56	0.57	335
2.0	0.29	0.21	0.24	196
3.0	0.38	0.26	0.31	306
4.0	0.33	0.23	0.27	495
5.0	0.78	0.89	0.83	2185

accuracy			0.67	3517
macro avg	0.47	0.43	0.44	3517
avg	0.63	0.67	0.65	3517

(e) Iterazione 5

	precision	recall	f1-score	support
1.0	0.51	0.55	0.53	276
2.0	0.35	0.22	0.27	216
3.0	0.33	0.28	0.30	269
4.0	0.39	0.25	0.30	537
5.0	0.78	0.89	0.83	2219

accuracy			0.68	3517
macro avg	0.47	0.44	0.45	3517
avg	0.64	0.68	0.65	3517

(f) Iterazione 6

	precision	recall	f1-score	support
1.0	0.55	0.53	0.54	325
2.0	0.28	0.20	0.23	223
3.0	0.31	0.24	0.27	289
4.0	0.38	0.26	0.31	512
5.0	0.78	0.89	0.83	2168

accuracy			0.67	3517
macro avg	0.46	0.42	0.44	3517
avg	0.63	0.67	0.64	3517

(g) Iterazione 7

	precision	recall	f1-score	support
1.0	0.59	0.60	0.60	325
2.0	0.35	0.26	0.30	215
3.0	0.38	0.32	0.34	288
4.0	0.41	0.28	0.33	491
5.0	0.80	0.90	0.85	2198

accuracy			0.70	3517
macro avg	0.51	0.47	0.48	3517
avg	0.66	0.70	0.68	3517

(h) Iterazione 8

	precision	recall	f1-score	support
1.0	0.60	0.55	0.58	327
2.0	0.34	0.28	0.31	198
3.0	0.34	0.30	0.31	281
4.0	0.39	0.25	0.31	518
5.0	0.79	0.89	0.84	2193

accuracy			0.68	3517
macro avg	0.49	0.45	0.47	3517
avg	0.65	0.68	0.66	3517

(i) Iterazione 9

	precision	recall	f1-score	support
1.0	0.61	0.56	0.58	326
2.0	0.38	0.28	0.33	204
3.0	0.35	0.31	0.32	265
4.0	0.36	0.23	0.28	528
5.0	0.79	0.89	0.83	2194

accuracy			0.68	3517
macro avg	0.50	0.46	0.47	3517
avg	0.65	0.68	0.66	3517

(j) Iterazione 10

Figura 15: Misure di performance generate ad ogni iterazione della cross-validation nella predizione di score

Il risultato ottenuto non è pertanto ottimale: ritenendo però a nostro avviso più importante acquisire la capacità di predire il sentimento, ovvero la polarità (che può essere positiva, negativa o neutra), piuttosto che il valore preciso dello score, è stato effettuato un tipo di predizione meno specifica per vedere se, così facendo, si riuscivano ad ottenere risultati migliori.

## 6.2 Predizione di sentimento

Il valore di `sentiment` è stato così definito: per ogni recensione, dato lo score, la variabile `sentiment` assume valore positivo se `score` ha valore 4 o 5, negativo se `score` ha valore 1 o 2 e neutro se `score` è uguale a 3.

Nella predizione di `sentiment` è stato applicato lo stesso tipo di apprendimento supervisionato utilizzato per la predizione di `score`.

Le matrici di confusione e le misure di performance per ogni iterazione della cross-validation sono riportate rispettivamente alle Figure 16 e 17.

		Predizione		
		negativo	neutro	positivo
Sentiment	negativo	333	57	144
	neutro	66	88	140
	positivo	80	74	2536

(a) Iterazione 1

		Predizione		
		negativo	neutro	positivo
Sentiment	negativo	326	53	166
	neutro	66	59	158
	positivo	87	94	2509

(b) Iterazione 2

		Predizione		
		negativo	neutro	positivo
Sentiment	negativo	297	50	163
	neutro	83	69	159
	positivo	94	69	2533

(c) Iterazione 3

		Predizione		
		negativo	neutro	positivo
Sentiment	negativo	325	58	159
	neutro	68	58	163
	positivo	84	85	2517

(d) Iterazione 4

		Predizione		
		negativo	neutro	positivo
Sentiment	negativo	300	48	173
	neutro	58	77	141
	positivo	92	78	2550

(e) Iterazione 5

		Predizione		
		negativo	neutro	positivo
Sentiment	negativo	324	45	136
	neutro	62	77	135
	positivo	99	77	2562

(f) Iterazione 6

		Predizione		
		negativo	neutro	positivo
Sentiment	negativo	333	41	151
	neutro	51	73	174
	positivo	95	61	2538

(g) Iterazione 7

		Predizione		
		negativo	neutro	positivo
Sentiment	negativo	324	44	152
	neutro	64	58	141
	positivo	84	76	2574

(h) Iterazione 8

		Predizione		
		negativo	neutro	positivo
Sentiment	negativo	314	48	165
	neutro	70	79	159
	positivo	96	79	2507

(i) Iterazione 9

		Predizione		
		negativo	neutro	positivo
Sentiment	negativo	294	50	151
	neutro	60	59	145
	positivo	100	76	2582

(j) Iterazione 10

Figura 16: Matrici di confusione generate ad ogni iterazione della cross-validation nella predizione del sentimento

	precision	recall	f1-score	support
negativo	0.70	0.62	0.66	534
neutro	0.40	0.30	0.34	294
positivo	0.90	0.94	0.92	2690
accuracy			0.84	3518
macro avg	0.67	0.62	0.64	3518
avg	0.83	0.84	0.83	3518

(a) Iterazione 1

	precision	recall	f1-score	support
negativo	0.68	0.60	0.64	545
neutro	0.29	0.21	0.24	283
positivo	0.89	0.93	0.91	2690
accuracy			0.82	3518
macro avg	0.62	0.58	0.60	3518
avg	0.81	0.82	0.81	3518

(b) Iterazione 2

	precision	recall	f1-score	support
negativo	0.63	0.58	0.60	510
neutro	0.37	0.22	0.28	311
positivo	0.89	0.94	0.91	2696
accuracy			0.82	3517
macro avg	0.63	0.58	0.60	3517
avg	0.80	0.82	0.81	3517

(c) Iterazione 3

	precision	recall	f1-score	support
negativo	0.68	0.60	0.64	542
neutro	0.29	0.20	0.24	289
positivo	0.89	0.94	0.91	2686
accuracy			0.82	3517
macro avg	0.62	0.58	0.60	3517
avg	0.81	0.82	0.81	3517

(d) Iterazione 4

	precision	recall	f1-score	support
negativo	0.67	0.58	0.62	521
neutro	0.38	0.28	0.32	276
positivo	0.89	0.94	0.91	2720
accuracy			0.83	3517
macro avg	0.65	0.60	0.62	3517
avg	0.82	0.83	0.82	3517

(e) Iterazione 5

	precision	recall	f1-score	support
negativo	0.67	0.64	0.65	505
neutro	0.39	0.28	0.33	274
positivo	0.90	0.94	0.92	2738
accuracy			0.84	3517
macro avg	0.65	0.62	0.63	3517
avg	0.83	0.84	0.84	3517

(f) Iterazione 6

	precision	recall	f1-score	support
negativo	0.70	0.63	0.66	525
neutro	0.42	0.24	0.31	298
positivo	0.89	0.94	0.91	2694
accuracy			0.84	3517
macro avg	0.67	0.61	0.63	3517
avg	0.82	0.84	0.82	3517

(g) Iterazione 7

	precision	recall	f1-score	support
negativo	0.69	0.62	0.65	520
neutro	0.33	0.22	0.26	263
positivo	0.90	0.94	0.92	2734
accuracy			0.84	3517
macro avg	0.64	0.60	0.61	3517
avg	0.82	0.84	0.83	3517

(h) Iterazione 8

	precision	recall	f1-score	support
negativo	0.65	0.60	0.62	527
neutro	0.38	0.26	0.31	308
positivo	0.89	0.93	0.91	2682
accuracy			0.82	3517
macro avg	0.64	0.60	0.61	3517
avg	0.81	0.82	0.81	3517

(i) Iterazione 9

	precision	recall	f1-score	support
negativo	0.65	0.59	0.62	495
neutro	0.32	0.22	0.26	264
positivo	0.90	0.94	0.92	2758
accuracy			0.83	3517
macro avg	0.62	0.58	0.60	3517
avg	0.82	0.83	0.83	3517

(j) Iterazione 10

Figura 17: Misure di performance generate ad ogni iterazione della cross-validation nella predizione di sentiment

Nella predizione del sentimento, sono stati riscontrati dei miglioramenti a livello di performance. La classe positiva è predetta con buoni risultati, mentre la negativa con risultati sufficienti. Una scarsa predizione la si ha per la classe neutra, i cui valori risultano essere notevolmente più bassi rispetto

alle altre due.

I motivi dietro questo risultato potrebbero essere: un minor numero di campioni disponibili per l'addestramento (la classe neutra è l'unica costituita da una sola classe di score e non due) e una probabile maggior variabilità delle parole usate nelle recensioni di questa classe.

Le recensioni neutre, infatti, non essendo né esplicitamente negative né positive, possono presentare sia termini positivi che negativi per descrivere i punti di forza e le critiche sul prodotto; sta poi al singolo utente scegliere quali aspetti enfatizzare (e quindi quale polarizzazione far emergere) nella sua recensione scritta.

Possiamo comunque notare che l'accuratezza generale si attesta tra l'81% e l'84% apportando un miglioramento del 15%.

## 7 Top Reviewers

Come già visto nel Capitolo 2, la maggior parte degli users considerati ha effettuato una sola recensione (ciò corrisponde anche al numero minimo di recensioni necessarie per far parte del dataset).

La media di recensioni per utente è 1.195, con una varianza di 0.527.

Questo valore alto per la varianza è spiegato dal fatto che poche persone effettuano più di una recensione e chi lo fa, ne scrive tante.

Come introdotto nel Capitolo 4, sono stati chiamati `top_reviewers` tutti coloro che hanno effettuato più di 10 recensioni e rientrano pertanto nell'0.05% degli utenti che recensiscono di più.

È stato quindi deciso di andare a studiare meglio questa ristretta classe di individui per cercare di comprenderne meglio il comportamento e capire come mai si discostano così tanto dal resto degli utenti.

Gli appartenenti ai `top_reviewers` risultano essere 19 individui:

- 'AY12DBB0U420B': 30 review
- 'A1Z54EM24Y40LL': 28 review
- 'A3OXHLG6DIBRW8': 24 review
- 'A2SZLNSI5KOQJT': 21 review
- 'A1YUL9PCJR3JTY': 20 review
- 'A31N6KB160O508': 18 review
- 'A3PJZ8TU8FDQ1K': 18 review
- 'A1TMAVN4CEM8U8': 17 review
- 'A281NPSIMI1C2R': 15 review
- 'A3HPCRD9RX351S': 14 review
- 'A2PNOU7NXB1JE4': 12 review
- 'A1ZH9LWMX5UCFJ': 11 review
- 'AQQLWCMRNDFGI': 11 review
- 'A250AXLRBVYKB4': 11 review
- 'AF3BYMPWKWO8F': 11 review
- 'A35R32TA60XD57': 11 review
- 'A2TN9C5E4A0I3F': 11 review
- 'AY1EF0GOH80EK': 11 review
- 'A1XGFW5016CGQI': 11 review

La domanda che ci si pone è: "I top reviewers presentano questo comportamento anomalo perché sono in qualche modo fraudolenti o scrivono cattive recensioni?".

Si vuole capire se i top reviewers sono tali per spam o perché effettivamente recensiscono onestamente.

Per effettuare questa verifica, è stata calcolata la media degli score di ogni prodotto per ricavare l'errore relativo percentuale tra quest'ultimo e lo score assegnato dall'utente.

In questo modo, più l'errore è basso meno ragioni si hanno di credere che un utente sia fraudolento.

La differenza di un punto tra score medio del prodotto e score assegnato da un top reviewers è stata considerata accettabile.

Per capire meglio, il come è stato gestito il calcolo dell'errore percentuale, basti pensare che un utente con soli voti da 3 stelle ad una lista di prodotti da 4, determina un errore percentuale del 25 %. Un utente che invece assegna sempre 2 stelle alla stessa lista di prodotti da 4, determinerà un errore del 50 %.

Per questo motivo, è stato deciso che un errore maggiore del 25 % possa indicare utenti con intenzioni fraudolente o poco oggettivi nel recensire.

Sono stati dunque considerati tutti gli utenti tra i top reviewers con più del 25 % di errore percentuale.

- 'A1YUL9PCJR3JTY' : 9.518785% — 20 review
- 'AY12DBB0U420B' : 13.043038% — 30 review
- 'AQQLWCMRNDFGI' : 13.798255% — 11 review
- 'A3HPCRD9RX351S' : 15.836790% — 14 review
- 'A3OXHLG6DIBRW8' : 17.178787% — 24 review
- 'A1XGFW5016CGQI' : 17.872723% — 11 review
- 'A281NPSIMI1C2R' : 18.374898% — 15 review
- 'A1Z54EM24Y40LL' : 19.220213% — 28 review
- 'A1ZH9LWMX5UCFJ' : 19.259065% — 11 review
- 'AY1EF0GOH80EK' : 19.354133% — 11 review
- 'A31N6KB160O508' : 20.465523% — 18 review
- 'A2SZLNSI5KOQJT' : 22.460137% — 21 review
- 'A1TMAVN4CEM8U8' : 24.404645% — 17 review
- 'A2PNOU7NXB1JE4' : 24.499311% — 12 review
- 'A3PJZ8TU8FDQ1K' : 25.367257% — 18 review
- 'A250AXLRBVYKB4' : 29.809660% — 11 review
- 'A35R32TA60XD57' : 42.540865% — 11 review
- 'A2TN9C5E4A0I3F' : 50.551699% — 11 review
- 'AF3BYMPWKWO8F' : 55.740516% — 11 review

Si può notare che la maggior parte dei top reviewers ha errori relativamente bassi (sotto al 25 %).

Emergono però users con errori più alti: in generale comunque su 19 users solo 5 hanno errori più alti della norma.

- 'A3PJZ8TU8FDQ1K' : 25.367257% — 18 review
- 'A250AXLRBVYKB4' : 29.809660% — 11 review
- 'A35R32TA60XD57' : 42.540865% — 11 review
- 'A2TN9C5E4A0I3F' : 50.551699% — 11 review

- 'AF3BYMPWKWO8F' : 55.740516% — 11 review

Non essendoci molte recensioni abbiamo deciso che il modo più semplice ed accurato per validare la nostra tesi fosse quello di leggerle (non processate) e determinare se effettivamente fossero casi di votazioni sbagliate o comportamenti fraudolenti.

A questo scopo, sono stati analizzati gli users 'AF3BYMPWKWO8F', 'A2TN9C5E4A0I3F', 'A35R32TA60XD57', 'A250AXLRBVYKB4' e 'A3PJZ8TU8FDQ1K'.

- AF3BYMPWKWO8F : È subito possibile notare come 9 delle 11 recensioni effettuate da questo utente riguardino lo stesso prodotto e come 8 di queste siano addirittura la stessa recensione copia-incollata, con voto 1 e relativa al prodotto 'B001BDDTB2' (Cibo per gatti), che ha una media di 4.15 stelle.

Leggendo il testo di queste recensioni, si evince che l'utente sta cercando di comunicare il fatto che questo prodotto sia in realtà tossico.

Non essendo in possesso di ulteriori informazioni non possiamo determinare con certezza se si tratti di fake news o meno, ma lo presumiamo, vedendo le altre recensioni.

*"According to the manufacturer's website, this (and many of their other products) contains menadione sodium bisulfite complex. Menadione is a synthetic precursor for Vitamin K and has been reported to have toxic effects. The use of menadione over the natural alternatives (such as leafy greens and kelp) is only for cost-saving reasons, hence usually only found in cheaper and lower quality animal food. It's surprising that pet food at this high of a price actually contains this substance. Menadione is banned by the Food & Drug Administration in over-the-counter supplements due to its potential for organ toxicity. It's also banned in Europe for human consumption. A study in rats showed that exposure to menadione produced lesions in the kidney, heart, liver and lung. Chiou et al. Toxicology 1997. Menadione's toxicity seems to be due to its ability to induce oxidative stress in cells. Do an online search for menadione and you'll find more info from a pet food watchdog trying to stop the inclusion of this substance. In light of recent episodes of toxic pet food and the questionable behaviors of manufacturers and our regulatory agencies, I would advise you to avoid feeding this to your cats."*

Notiamo inoltre come un altro utente (A3FKGKUCI3DG9U) ha recensito in modo simile lo stesso prodotto, copia-incollando 3 volte lo stesso testo:

*"My two traditional striped cats eat mostly dry Science Diet cat food with a few NuCat vitamins added. They also get Iams canned chicken, turkey, or beef once a day. Brownie prefers the dry food but will eat some canned food only when the can is freshly opened. Pi likes both. We tried the Petite Cuisine Variety Pack (Yellowfin, Snapper, Tuna & Sole, Tuna & Shrimp) and the Petite Cuisine Variety Pack (Chinese Chicken & Chicken Pot Pie) over the course of sixteen days. Each cat got about one-quarter can each day. I alternated cans of chicken and fish varieties. At first, the felines were enthusiastic. After a while they tired of the chicken. Pi would only lick up the gravy, while Brownie refused to even look at it. The fish was better received. They like the yellowfin and the tuna & shrimp. The cat vote is for two of the six varieties There is, however, a question of mercury in ocean fish. A 1995 study in Japan found that cats that ate tuna had a higher concentration of mercury in their fur than cats that ate dry cat food. The study and its results are discussed in an article that you can find on the web by using the search term all about tuna fish. I posted a question on the Petite Cuisine web site as to whether they had tested their products for mercury. I have not received an answer. My conclusion is that it may be ok to give the cats fish as an occasional treat but not as a steady diet."*

Questo commento risulta però essere più pacato del precedente e con un voto di 3, discostandosi di 1,15 da quello reale (lo consideriamo pertanto concorde).

- A2TN9C5E4A0I3F: Notiamo come invece questo utente scriva la stessa recensione sia per il prodotto 'B0058AMY74' (Kettle chips), che per il prodotto 'B000G6RYNE' (Kettle chips).

Notiamo inoltre che questi due prodotti ricevono le stesse recensioni dagli stessi utenti e che quindi probabilmente sono esattamente lo stesso prodotto.

Pensiamo dunque che non sia un comportamento fraudolento, ma che l'utente abbia semplicemente recensito in maniera discordante rispetto agli altri utenti, assegnando il voto 1 a un prodotto con media 4,013.

*"I've bought these at the local supermarket and enjoyed them although they are so salty that a few leave my tongue and roof of my mouth burning. Keeps you from eating too many! Occasionally I get really stale items*

*from Amazon.com and this was one. Unedible. Beware of the quality of food items on this website that are on special as they can be very close to due dates or in this case, not expired but stale and unedible just the same.”*

- A35R32TA60XD57 : Notiamo come quest’utente ripeta la stessa recensione per due prodotti diversi.  
In particolare, vengono ripetute Frase 1 e Frase 2 per i prodotti ’B003VXFK44’ (Caffe) e ’B006N3IG4K’ (Caffe), mentre Frase 3 per i prodotti ’B000EH2QPQ’ (Cibo per cani) e ’B000EH2QP6’ (Cibo per cani).
  1. Frase 1: *“I wasn’t expecting this one to be so rich or full! Yet it has a smoky finish that is typical of a French roast but doesn’t taste burnt. I think a lot of times a French roast tastes thin and yet bitter instead of full and smoky, but not this one! Very pleased with it and will definitely order again. If you like a stronger cup in the morning, without being bitter, give this one a try.”*
  2. Frase 2: *“This one will not disappoint! I got this yesterday and my husband and I tried it this morning with breakfast and we both loved it. I have gotten many different kinds of the k-cups to try but after trying this one, I think I could be happy with just this one and black tiger for those mornings when I want a really strong cup of coffee. The rodeo drive blend is delicious with a nice round feel on the palate and an interesting flavor on the finish with just a hint of smokiness. Perfectly balanced and absolutely wonderful! Plus, I have to admit...I love the label. The different color looks very appealing on my carousel with the other k-cups.”*
  3. Frase 3: *“I really wanted my little shihtzu, Sheldon, to like this food but it was a no go here. He barely sniffed it and walked away. I loved the size of the cans as well as the impressive ingredient list. I do have to say that it didn’t smell a lot like chicken to me though so I understand him not even tasting it. Too bad though. I would have loved to add this to the very small list of foods he will eat.”*

Ai prodotti B003VXFK44 e B006N3IG4K(entrambi di media 3.934066) viene assegnato il voto 5; ai prodotti B000EH2QPQ e B000EH2QP6 (entrambi di media 3.72) viene assegnato il voto 2.

Osservando la presenza delle stesse recensioni per entrambi i prodotti, supponiamo che B003VXFK44 e B006N3IG4K siano in realtà lo stesso prodotto. Lo stesso discorso vale anche per la coppia di prodotti B000EH2QP6 e B000EH2QPQ.

Non riteniamo perciò che il comportamento dell’utente sia anomalo, ma solo che abbia recensito male.

- ’A250AXLRBVYKB4’ e ’A3PJZ8TU8FDQ1K’: Non sembrano avere comportamenti fraudolenti e pensiamo abbiano recensito male.

Andando dunque a leggere le varie recensioni ci siamo resi conto che considerare l’errore relativo percentuale di un utente può essere una soluzione ragionevole per valutare i casi fraudolenti o di recensione discorde con quella media del prodotto.

## 8 Altro Processing

Come già menzionato nel Capitolo 7, andando ad analizzare i `top_reviewers`, ci siamo accorti che il dataset può essere ulteriormente processato.

Sono stati infatti notati due aspetti:

- Esistono utenti che scrivono più volte la stessa recensione per lo stesso prodotto.
- Esistono prodotti diversi che hanno le stesse recensioni.

### 8.1 Recensioni Duplicate

Dal momento che esistono users, come ad esempio ’AF3BYMPWKWO8F’, che eseguono recensioni duplicate, si è pensato di accorparli e di aggiungere un contatore `n_repeated_review` che possa identificare il numero di volte che la recensione è stata ripetuta.

Si è deciso di tener traccia di queste recensioni duplicate in modo tale da permettere, in futuro, altri studi o altri approcci al dataset.

Quest’informazione infatti potrebbe influire sul calcolo dello score medio di ogni prodotto.  
Infatti lo score medio di un prodotto potrebbe essere calcolato come:

- una media di tutti i voti, anche quelli ripetuti;
- una media di tutti i voti, ad eccezione di quelli assegnati da persone che hanno effettuato recensioni ripetute;
- una media di tutti i voti, considerando solo una volta le recensioni ripetute.

## 8.2 Prodotti Duplicati

Successivamente è stato considerato un altro aspetto, ovvero il fatto che esistono prodotti con `productid` diversi, ma con le stesse identiche recensioni.

Abbiamo ragione di credere che questi prodotti siano in realtà lo stesso prodotto con una qualche caratteristica diversa, ad esempio il peso o la quantità.

Dopo un lungo confronto sono stati dunque estratti i `sinonimi`, ovvero i prodotti con identificativo diverso ma che ragionevolmente riguardano lo stesso prodotto:

- “B0001VWGWS”, “B0001VWGWI” (Dark Chili Powder)
- “B0019IPKFC”, “B002KE33QC” (Cibo per cani)
- “B0007ZPY2C”, “B0007ZNW1W” (Caffe) - Jablum Jamaican
- “B00375LB6M”, “B00375LB6W” (Jimmies Sprinkle)
- “B001T4WKJ0”, “B001T4ZOK2” (AmeriColor colorante edibile)
- “B000EH2QP6”, “B000EH2QPQ” (Cibo per cani)
- “B006N3IG4K”, “B003VXFK44” (Caffe) - Wolfgang Puck K-Cups
- “B0058AMY74”, “B000G6RYNE” (Kettle chips)

Di questi prodotti, non essendo molti, sono infatti state lette le recensioni per capire effettivamente di cosa parlassero, identificando l'oggetto recensito.

Osservando gli identificativi di questi prodotti, è stato notato che molto spesso la radice del `productid` risulta essere uguale e variano solo gli ultimi caratteri.

Questo aspetto avvalorà l'ipotesi che si tratti effettivamente dello stesso prodotto.

Infatti, questa pratica viene adottata da molti negozi e si nota spesso che, in base per esempio al colore di un prodotto, l'ultima lettera del codice può cambiare.

## 9 Nuove Predizioni

Di seguito in Figura 18 e 19 sono stati riportati i nuovi risultati dell'approccio supervisionato descritto nel Capitolo 6 applicato al dataset in seguito alla seconda fase di *processing* (descritta nel Capitolo 8).

		Predizione				
		1	2	3	4	5
Score	1	174	44	27	10	66
	2	49	49	34	20	64
	3	26	31	79	35	117
	4	15	17	49	129	321
	5	36	23	45	140	1908

(a) Iterazione 1

		Predizione				
		1	2	3	4	5
Score	1	177	36	32	12	70
	2	45	46	23	20	42
	3	29	33	85	48	91
	4	17	18	59	121	316
	5	36	30	47	103	1972

(c) Iterazione 3

		Predizione				
		1	2	3	4	5
Score	1	182	38	27	13	93
	2	42	43	35	16	56
	3	33	42	77	40	98
	4	18	14	43	122	299
	5	42	24	41	135	1935

(e) Iterazione 5

		Predizione				
		1	2	3	4	5
Score	1	177	40	27	12	70
	2	39	42	37	16	47
	3	29	29	81	50	87
	4	17	16	31	140	314
	5	41	19	53	127	1967

(g) Iterazione 7

		Predizione				
		1	2	3	4	5
Score	1	187	38	28	14	70
	2	49	54	26	22	65
	3	24	32	70	54	105
	4	10	18	39	126	299
	5	27	10	48	148	1944

(i) Iterazione 9

		Predizione				
		1	2	3	4	5
Score	1	167	34	24	12	69
	2	41	45	39	16	51
	3	32	21	81	51	82
	4	17	21	41	122	299
	5	38	24	43	138	2000

(b) Iterazione 2

		Predizione				
		1	2	3	4	5
Score	1	162	42	17	15	53
	2	45	68	40	26	51
	3	24	35	69	53	90
	4	10	19	34	148	294
	5	38	18	48	152	1957

(d) Iterazione 4

		Predizione				
		1	2	3	4	5
Score	1	188	29	28	18	66
	2	43	48	32	22	55
	3	29	36	75	48	97
	4	15	12	46	127	312
	5	25	18	41	141	1957

(f) Iterazione 6

		Predizione				
		1	2	3	4	5
Score	1	179	36	23	13	65
	2	41	54	35	21	63
	3	24	29	99	47	101
	4	19	15	40	126	307
	5	35	14	43	140	1939

(h) Iterazione 8

		Predizione				
		1	2	3	4	5
Score	1	166	43	13	10	64
	2	45	32	32	24	54
	3	32	39	71	45	116
	4	15	11	40	126	278
	5	33	17	46	150	2005

(j) Iterazione 10

Figura 18: Matrici di confusione generate ad ogni iterazione della cross-validation nella predizione di score

	precision	recall	f1-score	support
1.0	0.58	0.54	0.56	321
2.0	0.30	0.23	0.26	216
3.0	0.34	0.27	0.30	288
4.0	0.39	0.24	0.30	531
5.0	0.77	0.89	0.82	2152

accuracy			0.67	3508
macro avg	0.47	0.43	0.45	3508
avg	0.63	0.67	0.64	3508

(a) Iterazione 1

	precision	recall	f1-score	support
1.0	0.57	0.55	0.56	306
2.0	0.31	0.23	0.27	192
3.0	0.36	0.30	0.33	267
4.0	0.36	0.24	0.29	500
5.0	0.80	0.89	0.84	2243

accuracy			0.69	3508
macro avg	0.48	0.44	0.46	3508
avg	0.66	0.69	0.67	3508

(b) Iterazione 2

	precision	recall	f1-score	support
1.0	0.58	0.54	0.56	327
2.0	0.28	0.26	0.27	176
3.0	0.35	0.30	0.32	286
4.0	0.40	0.23	0.29	531
5.0	0.79	0.90	0.84	2188

accuracy			0.68	3508
macro avg	0.48	0.45	0.46	3508
avg	0.65	0.68	0.66	3508

(c) Iterazione 3

	precision	recall	f1-score	support
1.0	0.58	0.56	0.57	289
2.0	0.37	0.30	0.33	230
3.0	0.33	0.25	0.29	271
4.0	0.38	0.29	0.33	505
5.0	0.80	0.88	0.84	2213

accuracy			0.69	3508
macro avg	0.49	0.46	0.47	3508
avg	0.66	0.69	0.67	3508

(d) Iterazione 4

	precision	recall	f1-score	support
1.0	0.57	0.52	0.54	353
2.0	0.27	0.22	0.24	192
3.0	0.35	0.27	0.30	290
4.0	0.37	0.25	0.30	496
5.0	0.78	0.89	0.83	2177

accuracy			0.67	3508
macro avg	0.47	0.43	0.44	3508
avg	0.64	0.67	0.65	3508

(e) Iterazione 5

	precision	recall	f1-score	support
1.0	0.63	0.57	0.60	329
2.0	0.34	0.24	0.28	200
3.0	0.34	0.26	0.30	285
4.0	0.36	0.25	0.29	512
5.0	0.79	0.90	0.84	2182

accuracy			0.68	3508
macro avg	0.49	0.44	0.46	3508
avg	0.65	0.68	0.66	3508

(f) Iterazione 6

	precision	recall	f1-score	support
1.0	0.58	0.54	0.56	326
2.0	0.29	0.23	0.26	181
3.0	0.35	0.29	0.32	276
4.0	0.41	0.27	0.32	518
5.0	0.79	0.89	0.84	2207

accuracy			0.69	3508
macro avg	0.48	0.45	0.46	3508
avg	0.65	0.69	0.67	3508

(g) Iterazione 7

	precision	recall	f1-score	support
1.0	0.60	0.57	0.58	316
2.0	0.36	0.25	0.30	214
3.0	0.41	0.33	0.37	300
4.0	0.36	0.25	0.30	507
5.0	0.78	0.89	0.83	2171

accuracy			0.68	3508
macro avg	0.50	0.46	0.48	3508
avg	0.65	0.68	0.66	3508

(h) Iterazione 8

	precision	recall	f1-score	support
1.0	0.63	0.55	0.59	337
2.0	0.36	0.25	0.29	216
3.0	0.33	0.25	0.28	285
4.0	0.35	0.26	0.29	492
5.0	0.78	0.89	0.83	2177

accuracy			0.68	3507
macro avg	0.49	0.44	0.46	3507
avg	0.64	0.68	0.66	3507

(i) Iterazione 9

	precision	recall	f1-score	support
1.0	0.57	0.56	0.57	296
2.0	0.23	0.17	0.19	187
3.0	0.35	0.23	0.28	303
4.0	0.35	0.27	0.31	470
5.0	0.80	0.89	0.84	2251

accuracy			0.68	3507
macro avg	0.46	0.43	0.44	3507
avg	0.65	0.68	0.66	3507

(j) Iterazione 10

Figura 19: Misure di performance generate ad ogni iterazione della cross-validation nella predizione di score

		Predizione								
		negativo	neutro	positivo			Predizione			
Score	negativo	306	47	143	Score	negativo	312	34	143	
	neutro	51	82	142		neutro	63	77	170	
	positivo	96	96	2545		positivo	84	72	2553	
(a) Iterazione 1										
		Predizione						Predizione		
		negativo	neutro	positivo			negativo	neutro	positivo	
Score	negativo	354	45	162	Score	negativo	346	48	161	
	neutro	64	70	146		neutro	62	78	148	
	positivo	106	75	2486		positivo	79	81	2505	
(b) Iterazione 2										
		Predizione						Predizione		
		negativo	neutro	positivo			negativo	neutro	positivo	
Score	negativo	343	45	153	Score	negativo	278	50	153	
	neutro	82	73	142		neutro	62	64	150	
	positivo	104	85	2481		positivo	94	76	2581	
(c) Iterazione 3										
		Predizione						Predizione		
		negativo	neutro	positivo			negativo	neutro	positivo	
Score	negativo	326	52	186	Score	negativo	298	45	163	
	neutro	73	67	164		neutro	59	67	143	
	positivo	96	80	2464		positivo	75	77	2581	
(d) Iterazione 4										
		Predizione						Predizione		
		negativo	neutro	positivo			negativo	neutro	positivo	
Score	negativo	278	61	154	Score	negativo	319	51	148	
	neutro	58	64	178		neutro	51	76	125	
	positivo	99	67	2548		positivo	87	70	2580	
(e) Iterazione 5										
		Predizione						Predizione		
		negativo	neutro	positivo			negativo	neutro	positivo	
Score	negativo	326	52	186	Score	negativo	298	45	163	
	neutro	73	67	164		neutro	59	67	143	
	positivo	96	80	2464		positivo	75	77	2581	
(f) Iterazione 6										
		Predizione						Predizione		
		negativo	neutro	positivo			negativo	neutro	positivo	
Score	negativo	278	61	154	Score	negativo	319	51	148	
	neutro	58	64	178		neutro	51	76	125	
	positivo	99	67	2548		positivo	87	70	2580	
(g) Iterazione 7										
		Predizione						Predizione		
		negativo	neutro	positivo			negativo	neutro	positivo	
Score	negativo	278	61	154	Score	negativo	319	51	148	
	neutro	58	64	178		neutro	51	76	125	
	positivo	99	67	2548		positivo	87	70	2580	
(h) Iterazione 8										
		Predizione						Predizione		
		negativo	neutro	positivo			negativo	neutro	positivo	
Score	negativo	278	61	154	Score	negativo	319	51	148	
	neutro	58	64	178		neutro	51	76	125	
	positivo	99	67	2548		positivo	87	70	2580	
(i) Iterazione 9										
		Predizione						Predizione		
		negativo	neutro	positivo			negativo	neutro	positivo	
Score	negativo	278	61	154	Score	negativo	319	51	148	
	neutro	58	64	178		neutro	51	76	125	
	positivo	99	67	2548		positivo	87	70	2580	
(j) Iterazione 10										

Figura 20: Matrici di confusione generate ad ogni iterazione della cross-validation nella predizione del sentimento

	precision	recall	f1-score	support
negativo	0.68	0.62	0.64	496
neutro	0.36	0.30	0.33	275
positivo	0.90	0.93	0.91	2737

accuracy		0.84	3508
macro avg	0.65	0.61	0.63
avg	0.83	0.84	0.83

(a) Iterazione 1

	precision	recall	f1-score	support
negativo	0.68	0.64	0.66	489
neutro	0.42	0.25	0.31	310
positivo	0.89	0.94	0.92	2709

accuracy		0.84	3508
macro avg	0.66	0.61	0.63
avg	0.82	0.84	0.83

(b) Iterazione 2

	precision	recall	f1-score	support
negativo	0.68	0.63	0.65	561
neutro	0.37	0.25	0.30	280
positivo	0.89	0.93	0.91	2667

accuracy		0.83	3508
macro avg	0.64	0.60	0.62
avg	0.81	0.83	0.82

(c) Iterazione 3

	precision	recall	f1-score	support
negativo	0.71	0.62	0.66	555
neutro	0.38	0.27	0.32	288
positivo	0.89	0.94	0.91	2665

accuracy		0.83	3508
macro avg	0.66	0.61	0.63
avg	0.82	0.83	0.83

(d) Iterazione 4

	precision	recall	f1-score	support
negativo	0.65	0.63	0.64	541
neutro	0.36	0.25	0.29	297
positivo	0.89	0.93	0.91	2670

accuracy		0.83	3508
macro avg	0.63	0.60	0.61
avg	0.81	0.83	0.82

(e) Iterazione 5

	precision	recall	f1-score	support
negativo	0.64	0.58	0.61	481
neutro	0.34	0.23	0.27	276
positivo	0.89	0.94	0.92	2751

accuracy		0.83	3508
macro avg	0.62	0.58	0.60
avg	0.82	0.83	0.82

(f) Iterazione 6

	precision	recall	f1-score	support
negativo	0.66	0.58	0.62	564
neutro	0.34	0.22	0.27	304
positivo	0.88	0.93	0.90	2640

accuracy		0.81	3508
macro avg	0.62	0.58	0.60
avg	0.79	0.81	0.80

(g) Iterazione 7

	precision	recall	f1-score	support
negativo	0.69	0.59	0.64	506
neutro	0.35	0.25	0.29	269
positivo	0.89	0.94	0.92	2733

accuracy		0.84	3508
macro avg	0.65	0.59	0.62
avg	0.82	0.84	0.83

(h) Iterazione 8

	precision	recall	f1-score	support
negativo	0.64	0.56	0.60	493
neutro	0.33	0.21	0.26	300
positivo	0.88	0.94	0.91	2714

accuracy		0.82	3507
macro avg	0.62	0.57	0.59
avg	0.80	0.82	0.81

(i) Iterazione 9

	precision	recall	f1-score	support
negativo	0.70	0.62	0.65	518
neutro	0.39	0.30	0.34	252
positivo	0.90	0.94	0.92	2737

accuracy		0.85	3507
macro avg	0.66	0.62	0.64
avg	0.84	0.85	0.84

(j) Iterazione 10

Figura 21: Misure di performance generate ad ogni iterazione della cross-validation nella predizione di sentiment

Come è possibile notare, la seconda fase di *processing* non ha influenzato visibilmente i risultati ottenuti con la tecnica supervisionata.

Questo risultato è dovuto al fatto che il *processing* effettuato riguarda casistiche molto rare, influenzando solo una piccolissima parte di dati.

Il lavoro risulta comunque essere interessante, soprattutto nell'ottica di una possibile futura estensione dell'analisi in seguito all'acquisizione di nuovi dati.

## 10 Nuovo Dataset

Per rendere più fruibili i nostri risultati ad altre persone abbiamo deciso di salvare il nuovo Dataset processato in un csv.

I campi sono i seguenti:

- `productid` : ID del prodotto.
- `userid` : ID dell'utente/user.
- `score` : Rate (da 1 a 5) dato dall'utente al prodotto.
- `sentiment` : Valore di sentiment (positivo, negativo o neutro) del prodotto, ricavato da `score`.
- `text` : Testo processato.
- `afinn` : Score di afinn calcolato sul testo.
- `afinn_norm` : Score di afinn normalizzato da 1 a 5 calcolato sul testo.
- `afinn_sentiment` : Valore di sentiment (positivo, negativo o neutro) del prodotto, ricavato da `afinn_norm`.
- `var_afinn` : Varianza di afin calcolata sul testo, utile per casi neutri.
- `initial_pred_ML_score` : Predizione iniziale di `score` ottenuta con approccio supervisionato.
- `initial_pred_ML_sentiment` : Predizione iniziale di `sentiment` ottenuta con approccio supervisionato.
- `n_repeated_review` : Numero di volte che una recensione identica a questa è stata scritta.
- `alternative_prod_id` : Lista di `productid` alternativi usati per identificare lo stesso prodotto.
- `final_pred_ML_score` : Predizione finale di `score` ottenuta con approccio supervisionato.
- `final_pred_ML_sentiment` : Predizione finale di `sentiment` ottenuta con approccio supervisionato.

## 11 Dashboard

Al fine di rendere più fruibili le informazioni e i risultati riscontrati con le analisi effettuate, abbiamo creato una semplice *Dashboard* che permetta di interagire con il dataset e potrebbe tornare utile per eventuali analisi future.

La *Dashboard* è composta da due Tab con due scopi differenti:

- *Search Tab* : Dà la possibilità all'utente di selezionare una parola e di cercare le recensioni in cui compare. Oltre ad un boxplot, che mostra la distribuzione dello score legato alle recensioni in cui la parola è presente, viene mostrato anche l'insieme delle recensioni in cui la parola occorre.
- *Top Reviewers Tab* : In fase di esplorazione ed analisi del dataset abbiamo fissato i “Top Reviewers” come i top 0.05% dei recensori, per numero di recensioni.

In realtà questa scelta potrebbe esser cambiata. Abbiamo dunque costruito questa Tab per poter dar modo di esplorare, anche se brevemente, cosa potrebbe succedere se cambiamo il numero di recensioni minime per essere considerati “Top Reviewers”.

Questa Tab mostra dunque la percentuale di “Top Reviewers” rispetto al totale, i prodotti più recensiti da questa categoria e, per ogni utente, il numero di recensioni e l'errore relativo percentuale.

## Food Reviews



Figura 22: Search Tab

## Food Reviews

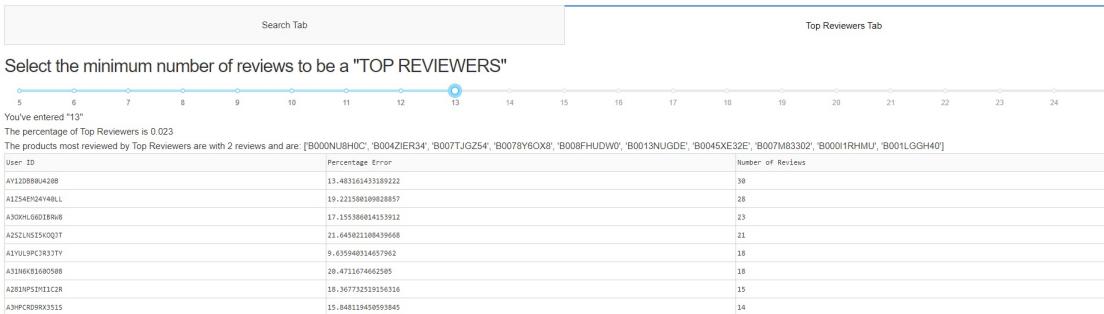


Figura 23: Top Reviewers Tab

## 12 Conclusioni

La nostra analisi è stata guidata da alcune domande.

- È possibile predire lo Score sfruttando un approccio basato sul lessico?

Non è possibile predire in maniera sufficiente utilizzando un approccio basato sul lessico, in quanto per definizione vengono considerate le singole parole, senza relazionarle fra loro e senza valutare il contesto.

I risultati ottenuti con questa tecnica risultano essere pessimi.

Un sostanziale miglioramento lo si ottiene nel limitarsi a predire il sentiment, che con questo approccio garantisce per lo meno una buona previsione della classe positiva.

Si veda il Capitolo 5.

- È possibile predire lo Score sfruttando un approccio di apprendimento supervisionato?

È possibile predire con risultati abbastanza buoni sia la classe di score corrispondente al valore 5, con risultati scarsi la classe 1 e con risultati pessimi le 3 classi centrali.

Limitandosi alla predizione del sentiment si può invece notare come si riescano ad ottenere risultati soddisfacenti sia per classe positiva che per quella negativa. La classe neutra invece, in quanto via di mezzo tra le altre due, risulta sempre essere quella più difficile da identificare.

Si veda il Capitolo 6.

- Quali sono i “Top Reviewers” che si possono considerare come fraudolenti o cattivi recensori?

Per vedere se è possibile considerare un “Top Reviewers” fraudolento o cattivo recensore abbiamo guardato l’errore relativo tra i voti dati ai prodotti e lo score medio del prodotto. Abbiamo notato che utenti con errori relativi percentuali sopra al 25% si possono considerare tali.

Si veda il Capitolo 7.

- La distribuzione delle parole tra i “Top Reviewers” e gli altri utenti è significativamente diversa?

La distribuzione di parole tra i “Top Reviewers” e gli altri utenti non varia in modo considerevole.  
Si veda il Capitolo 4.

- Esistono fenomeni anomali all’interno del dataset?

Si, esistono prodotti duplicati, cioè con le stesse recensioni ma con `productid` diverso. Inoltre esistono utenti che scrivono le stesse recensioni, più volte per uno stesso prodotto.

Si veda il Capitolo 8.