TEXT MINING AND SEARCH
AMAZON'S REVIEWS TOPIC CLASSIFICATION

A PREPRINT

Federico Maria Viotti
785867
Università degli studi di Milano Biocca
f.viotti@campus.unimib.it

Andrea Armando Tinella
771399
Università degli studi di Milano Biocca
a.tinella@campus.unimib.it

June 12, 2019

Contents

ABSTRACT

Questo progetto si propone di risolvere un task di topic classification su un insieme di recensioni di nove categorie di prodotti venduti sul sito e-commerce Amazon.

Dopo una prima parte dedicata all'esplorazione dei dati e alla sentiment analysis delle recensioni, viene approfondito il tema del preprocessing, dell'implementazione dei modelli e dei relativi risultati della topic classification.

1 Introduzione

Il sito di e-commerce Amazon è uno dei leader di mercato dei nostri giorni. Nata nel 1994 come e-commerce di libri cartacei, è riuscita, sfruttando le potenzialità delle nuove tecnologie digitali a portare i suoi servizi sul mercato globale, partendo dal retailing fino ad ampliare la sua gamma di prodotti e servizi a settori completamente diversi e variegati (cloud computing, streaming, videogaming, etc.).

Ponendo l'attenzione sul sito di e-commerce di Jeff Bezos, uno degli aspetti caratterizzanti l'esperienza di acquisto è la possibilità da parte dei clienti di effettuare recensioni sui prodotti, sia per mezzo di una valutazione numerica (le classiche 5 stelline) che per argomentazioni in linguaggio naturale, utili ai potenziali nuovi acquirenti che possono così leggere cosa ne pensa chi ha già acquistato il prodotto.

Oltre ai loro clienti, le recensioni vengono utilizzate dall'azienda stessa per effettuare analisi che permettono di migliorare e ottimizzare il processo aziendale attraverso un riscontro diretto; data l'enorme mole di questa tipologia di informazioni, risulta impossibile anche per un'azienda grande come Amazon, dare l'onere dell'analisi delle recensioni a del personale.

Proprio per questo motivo, si rendono necessari la creazione e l'utilizzo di strumenti di analisi di dati semi-strutturati che possano, in tempi brevi, analizzare le informazioni con risultati soddisfacenti.

2 Analisi esplorativa

Il primo passo ha riguardato un'analisi di tipo esplorativo/descrittiva al fine di ottenere una visione globale delle informazioni, step fondamentale di qualunque tipologia di analisi.

I dataset utilizzati per il progetto, ottenuti al seguente link, riguardavano le recensioni di 9 categorie di prodotti venduti su Amazon, raccolte tra Maggio 1996 e Luglio 2014. La scelta delle categorie è stata guidata sia dai gusti personali che dalla grandezza del dataset finale che, per motivi computazionali, non doveva essere impeditiva; dunque delle 24 categorie il subset selezionato, di grandezza pari a 900Mb, comprendeva le seguenti:

- Automotive
- Beauty
- Digital Music
- Health & Care
- Musical Instruments
- Office Products
- Patio & Garden
- Toys & Games
- Video Games

Le colonne del dataset fornivano informazioni riguardanti il recensore, il prodotto e le caratteristiche delle recensioni stesse. In particolare:

- *reviewerID* - l'ID del cliente che ha effettuato la recensione
- *asin* - l'ID del prodotto
- *reviewerName* - il nome del cliente che ha fatto la recensione

- *helpful* - rating di quanto la recensione è stata utile
- *reviewText* - il testo della recensione
- *overall* - il rating del prodotto
- *summary* - il sommario della review
- *unixReviewTime* - data della recensione in UNIXTIME
- *reviewTime* - data della recensione

Dopo aver convertito il dataset in csv e aver eliminato le colonne non necessarie all'analisi esplorativa sono state prodotte 3 dashboard con il software Tableau, visualizzabili al seguente link e delle quali è mostrato qui un estratto:

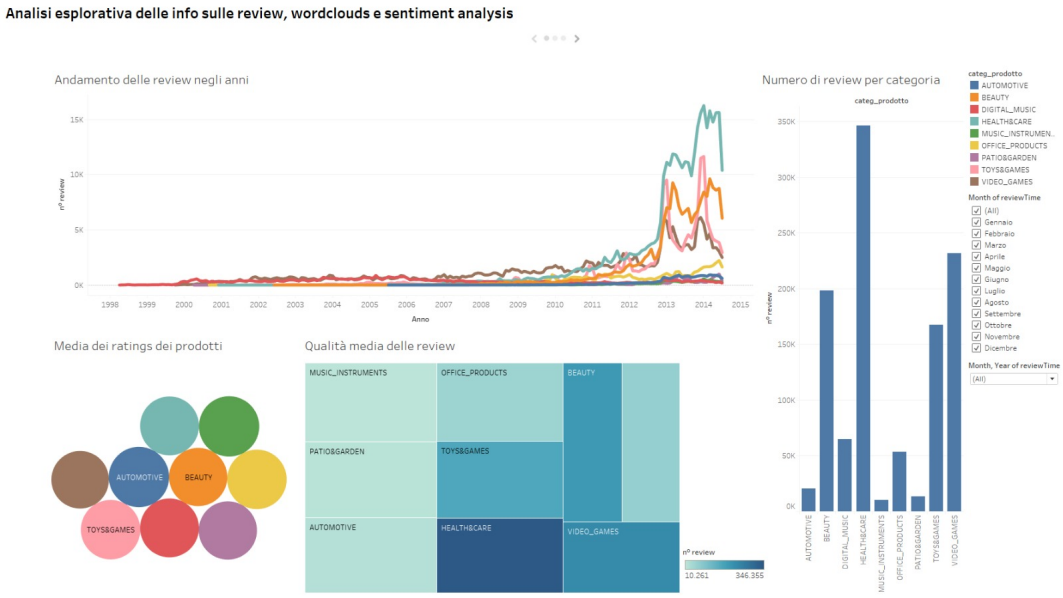

Figure 1: Analisi esplorativa

Alcuni insight ricavabili dalla dashboard sono, ad esempio, la numerosità delle recensioni che aumenta in modo considerevole a partire dal 2012, una tendenziale uniformità nel giudizio dato ai prodotti delle varie categorie, l'esiguità di recensioni per le categorie degli strumenti musicali e dei prodotti da giardino, contrapposta all'abbondanza di recensioni relativamente ai prodotti per la salute e di bellezza e dei videogiochi.

3 Sentiment analysis

La sentiment analysis è una tecnica di elaborazione del linguaggio naturale utilizzata con lo scopo di identificare ed estrarre le opinioni da un testo.
In riferimento alle recensioni sui prodotti venduti da Amazon, l'analisi del sentimento risulta uno strumento molto utile per capire cosa pensano i clienti dei prodotti acquistati e per poter effettuare analisi comparative come ad esempio il confronto tra il rating dato ad un prodotto e la relativa recensione scritta dall'acquirente. Per questo si è deciso di implementare un'analisi del sentiment su R.

3.1 Sentiment analysis delle recensioni

Dopo aver effettuato le necessarie operazioni di pulizia del testo delle recensioni, contenute nella colonna reviewText del dataset, la sentiment analysis è stata declinata in tre parti e svolta per ognuna delle nove categorie di prodotto.

- Il confronto tra sentiment positivo e negativo rispetto alla totalità delle recensioni di una particolare categoria di prodotto;

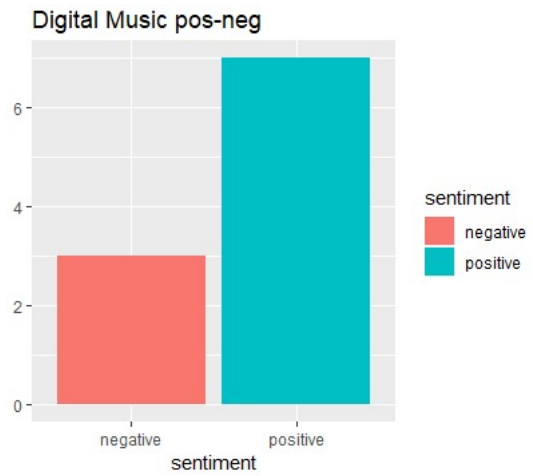

Figure 2: Sentiment analysis: recensioni con sentiment positivo e negativo

- Il "ventaglio" di emozioni date da una particolare categoria di prodotti, secondo il lessico NRC, basato sulle otto emozioni base, tarato sulla lingua inglese;

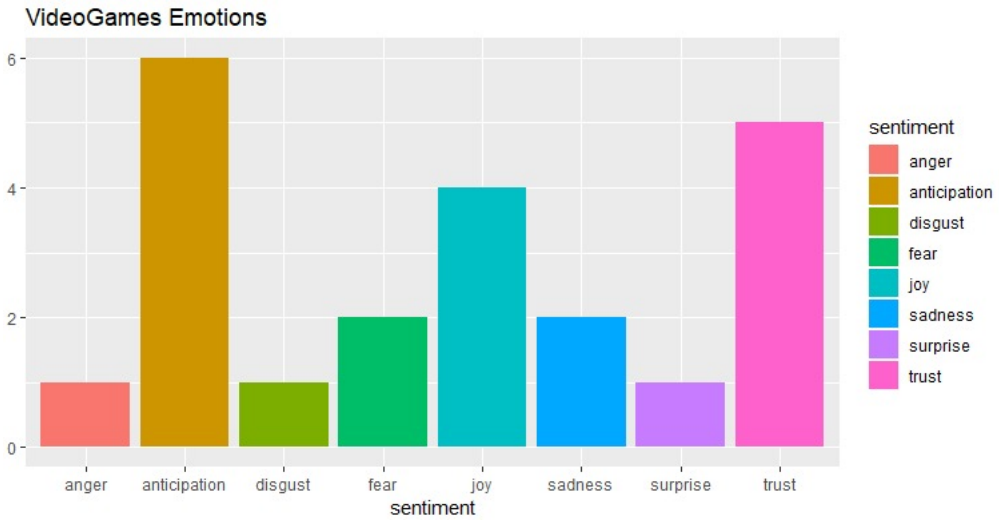

Figure 3: Sentiment analysis: lessico nrc

- La wordcloud delle parole più ricorrenti all'interno di ogni categoria di prodotti;

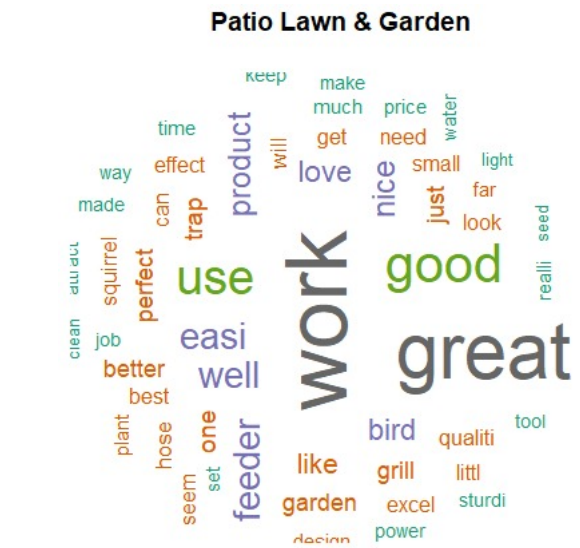

Figure 4: Sentiment analysis: wordcloud

- Infine, il confronto tra le parole più utilizzate di tre categorie di prodotti contemporaneamente, per valutare eventuali parole in comune a più categorie.

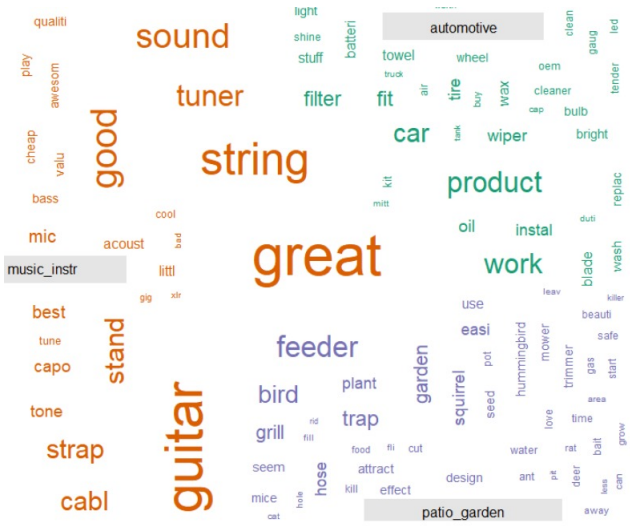

Figure 5: Sentiment analysis: wordcloud su tre categorie

4 La topic classification

L'obiettivo del progetto era l'applicazione di uno o più algoritmi di classificazione alle recensioni al fine di riuscire a prevederne la categoria di prodotto associata.

4.1 Metodologia

Dopo un'iniziale definizione del subset oggetto di analisi, il progetto è stato svolto seguendo le fasi tipiche del text mining:

- preprocessing dei dati
- definizione e modellazione degli algoritmi
- valutazione e comparazione dei risultati

Si è infine scelto di usare il linguaggio python con le necessarie librerie (pandas, numpy e in particolare per i task di text mining *NLTK* e *keras*); nonostante la sentiment analysis sia stata effettuata su R, python è stato preferito per l'implementazione degli algoritmi di classificazione poiché più adatto a gestire moli di dati maggiori e quindi più efficiente.

4.2 Preprocessing

La fase di preprocessing ha seguito nell'ordine le seguenti fasi atte a pulire e normalizzare il testo delle recensioni così da poter successivamente applicare al meglio la classificazione.

- il testo è stato trasformato tutto in minuscolo;
- è stata rimossa la punteggiatura;
- sono state rimosse eventuali emoji ed eventuali url presenti nel testo;
- sono state rimosse le stopwords
- è stata effettuata la lemmatizzazione (preferita per il costo computazionale più basso)
- si sono eliminati tutti gli spazi vuoti aggiuntivi
- si è generata la document-term matrix riempita con i pesi tf-idf, limitata alle prime 10000 parole

Successivamente si sono definiti il train e il test set (rispettivamente 85% e 15%) e la variabile outcome è stata trasformata in una dummy (binarizzata) per poter essere utilizzata con i modelli della libreria keras.

4.3 I modelli

Considerando la mole dei dati a disposizione e la loro complessità sia da un punto di vista computazionale che qualitativo, si è scelto di utilizzare due modelli di deep learning, relativi alla famiglia dei modelli MLP (multilayer perceptron).
L'idea è stata quella di applicare due diversi ottimizzatori (l'*SGD* e l'*adam*) a due reti neurali con identica architettura per valutarne le differenze in termini di efficacia ed efficienza. Per semplicità gli ottimizzatori usati non sono stati tunati ma si sono mantenute le caratteristiche di default.

Le caratteristiche della rete utilizzata sono riassunte nella seguente tabella:

Caratteristica	Primo modello	Secondo modello
Livelli nascosti	3	3
Dropout (in %)	30	30
Neuroni per livello	500-500-250	500-500-250
Ottimizzatore	SDG	adam
Numero di epoche	10	10
Dimensione del batch	128	128

5 Risultati

I due modelli hanno portato, come previsto, ottimi risultati in valore assoluto fin dalle epoche iniziali. Il confronto dei risultati dei due modelli può essere sintetizzato come segue:

Metrica	Modello SGD	Modello adam
Accuracy su training set	93.8	98
Accuracy su validation set	92	93
Accuracy su test set	92.5	93
Precision media	0.92	0.93
Recall media	0.92	0.93
Tempo di esecuzione per epoca	6 minuti	12 minuti
F-1 score medio	0.92	0.93

Come si può osservare dalla tabella dei risultati, i due ottimizzatori portano a risultati molto simili sia nell'accuracy che nelle altre metriche di valutazione. Le due differenze principali riguardano il tempo di training dei due algoritmi

e il loro overfitting: osservando queste due caratteristiche si noti come l'ottimizzatore *adam* ottenga un overfitting maggiore (da 0.98 a 0.93) contro l'*SGD*, che passa da 0.938 di accuracy sul training a 0.925 sul test set e inoltre vede aumentare quest'ultima sull'accuracy calcolata sul validation set; infine il tempo passando da un modello all'altro raddoppia, portando alla conclusione che sia preferibile scegliere come ottimizzatore il gradiente stocastico a discapito di 1 punto percentuale sulle metriche di performance.

6 Conclusioni

I due modelli presi in considerazione non sono certamente esaustivi, ma portano a ottimi risultati già con le caratteristiche attuali.
Per ottenere risultati ancora migliori, si potrebbe attuare un fine tuning dei parametri, "smorzando" e ottimizzando i valori in gioco per ottenere risultati ancora migliori e ridurre l'overfitting.
Come ben risaputo, inoltre, le reti neurali mostrano i loro limiti più importanti per quanto riguarda l'interpretabilità del modello: bisogna dunque considerare se è più importante ottenere ottime performance del modello in valore assoluto, oppure ripiegare su altri modelli più facilmente spiegabili e mirati a una maggiore interpretazione da parte dell'uomo.

References

[1] Fonte Dataset di partenza

[2] Python

[3] Rstudio

[4] keras

[5] Tableau software

[6] Link all'analisi Tableau