DATA ANALYTICS

Amazon

Ceccon Martina 817613
Costantino Chantal 860621
Maggio Vittorio 817034

Indice

Elenco delle tabelle

Elenco delle figure

Capitolo 1

Richiami teorici

L'analisi del *sentiment* o *sentiment analysis* (nota anche come *opinion mining*) è un campo dell'elaborazione del linguaggio naturale che si occupa di costruire sistemi per l'identificazione ed estrazione di opinioni dal testo, basandosi sui principali metodi di linguistica computazionale e di analisi testuale. Per identificare le informazioni soggettive che denotano opinioni , è necessario determinare la polarità insita in esse. Questa può essere di tre tipi: positiva, negativa o neutra.

polarità positiva: l'informazione che ricaviamo è positiva.

polarità negativa: l'informazione che ricaviamo è negativa.

polarità neutra: non ricaviamo alcuna informazione sulla polarità. Questo può avvenire per due diversi motivi:

- le opinioni positive sono bilanciate da quelle negative;
- non si manifesta alcun sentimento.

Per il calcolo della polarità esistono diversi tipi di approccio: basato sul lessico, supervisionato, semi-supervisionato e non supervisionato.

Approccio basato sul lessico: la polarità viene calcolata analizzando la frase.

Approccio supervisionato: la polarità viene calcolata utilizzando le tecniche del *Machine Learning*. Inserisco prima le parole in cui la polarità è esplicita, poi aggiungo le altre per fare inferenza.

Approccio semi-supervisionato: ????????

Approccio non supervisionato: ????????

Vedremo che per il nostro non è stato necessario andare a calcolare il *sentiment*; in quanto un attributo si prestava bene per questo tipo di analisi.

Capitolo 2

Sentiment Analysis

Il nostro primo obiettivo è quello di studiare l'intrattenimento, in particolare quali siano i prodotti più recensiti all'interno delle nostre quattro categorie. Come primo passo abbiamo quindi scelto di effettuare lo studio su un numero limitato di prodotti, riducendoli a duecento, e di analizzare la loro distribuzione. I risultati ottenuti sono presentati nella Figura 2.1

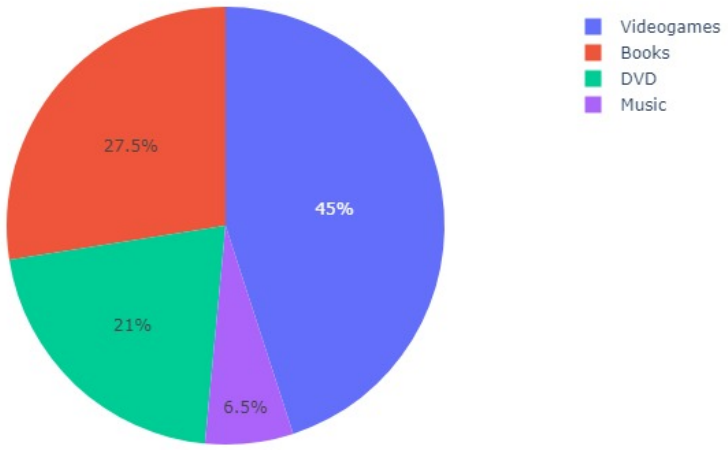


Figura 2.1: Distribuzione delle recensioni per ogni categoria.

Dalla figura è possibile notare che, in percentuale, i prodotti più recensiti appartengono alla categoria dei videogiochi (`45%`), seguiti dai libri (`27.5%`), dai dvd (`21%`) e dalla musica (`6.5%`). Quest'ultima categoria è davvero esigua

perché delle analisi diano risultati consistenti e si possano ricavare informazioni utili; inoltre poiché lo studio su cui abbiamo posto l'attenzione riguardava i più popolari, non aveva senso esaminare dei prodotti quasi privi di recensioni. Da questa considerazione è derivata la scelta di escludere `music` e procedere ad analizzare l'intrattenimento solo sulle tre migliori categorie, quindi `videogames`, `books` e `dvd`. Arrivati a questo punto il nostro studio è stato diviso in due diversi passaggi, durante i quali abbiamo cercato di rispondere a due domande ovvero quali siano i prodotti più recensiti e perché proprio loro.

2.1 Primo passaggio: distribuzione del sentiment

Come già accennato, per poter valutare la polarità di una recensione è necessario calcolare il *sentiment*. Abbiamo visto che questo calcolo può essere eseguito secondo diverse metodologie, tuttavia il nostro *dataset* relativo ai prodotti, presentava al suo interno un campo che ben si prestava a questo tipo di analisi. L'attributo in questione è quello delle stelle. *Amazon* infatti per ogni recensione riferita a un determinato prodotto associa una valutazione in stelle da `0` a `5`. Questo ha permesso di effettuare un conteggio per ogni prodotto di tutte le sue recensioni, dividendole in positive negative e neutre, secondo lo schema di seguito.

- **positive:** la recensione aveva un punteggio di maggiore di tre stelle.
- **neuter:** la recensione aveva un punteggio uguale a tre stelle.
- **negative:** la recensione aveva un punteggio minore di tre stelle.

Procedendo secondo queste modalità, preso per esempio il prodotto "Zelda", con `200` recensioni, si calcolano quante di queste sono risultate positive negative oppure neutre e un possibile risultato potrebbe essere costituito da `100` recensioni positive, `70` negative e `30` neutre.
Ottenuto questo elenco abbiamo calcolato la distribuzione probabilistica associata alle tre diverse polarità. In particolare per ognuna associata a uno specifico prodotto, abbiamo effettuato un rapporto tra il conteggio delle recensioni positive e negative, escludendo le neutre (dato non rilevante per la nostra analisi) e il numero di recensioni totali.

La distribuzione probabilistica della polarità ottenuta è stata quindi combinata al sottoinsieme dei prodotti più recensiti/più popolari e il risultato di questo *match* è stata l'identificazione per ogni categoria dei primi cinque prodotti più popolari suddivisi in due elenchi. Il primo contenente i prodotti valutati più negativamente e nel secondo quelli valutati più positivamente.
I risultati ottenuti sono visibili nelle Figure 2.2 2.3, 2.4, 2.5, 2.6, 2.7.

ID	Numero di recensioni	Titolo del libro	Recensioni positive	Recensioni negative	Recensioni neutrali
8817055778	262	Per questo mi chiamo Giovanni. Da un padre a un figlio il racconto della vita di Giovanni Falcone	99.24%	0.38%	0.38%
8867143336	298	Il Gruffalò. Ediz. illustrata	97.99%	1.34%	0.67%
8858012534	174	I colori delle emozioni. Ediz. illustrata (pop-up)	97.70%	0.00%	2.30%
8868364026	150	25 grammi di felicità . Come un piccolo riccio può cambiarti la vita	97.33%	1.33%	1.33%
8817109444	122	Metodo universitario. Come studiare meglio in meno tempo e superare gli esami senza ansia	96.72%	2.46%	0.82%

Figura 2.2: Primi cinque libri con più recensioni positive

ID	Numero di recensioni	Titolo del libro	Recensioni positive	Recensioni negative	Recensioni neutrali
8820067706	130	After	46.15%	40.77%	13.08%
8869876616	131	Il magico potere del riordino. Il metodo giapponese che trasforma i vostri spazi e la vostra vita	63.36%	29.01%	7.63%
8830100463	322	La verità sul caso Harry Quebert	58.70%	28.57%	12.73%
8891817597	235	Divertiti con Luì e Sofì. Il Fantalibro dei Me contro Te	62.55%	28.09%	9.36%
8890955503	556	Vivere 120 anni. Le verità che nessuno vuole raccontarti	71.40%	20.68%	7.91%

Figura 2.3: Primi cinque libri con più recensioni negative

2.2 Secondo passaggio: analisi delle parole più usate

Avendo risposto al primo quesito ci siamo potuti soffermare sulla seconda interrogazione, ovvero perché fossero risultati proprio questi prodotti. Siamo così andati alla ricerca delle parole più utilizzate all'interno delle recensioni (negative e positive), cercando una corrispondenza tra queste e i prodotti migliori. Tuttavia per poter eseguire questo confronto sono state necessarie delle operazioni atte a uniformare il *dataset*; i dati infatti non sempre sono puliti, in essi sono presenti errori di battitura, intere frasi scritte in maiuscolo, eccessiva punteggiatura, etc. Ecco dunque il motivo del termine "uniformare", intendiamo con esso il processo di riscrittura delle frasi seguendo degli specifici passaggi, che saranno descritti nel seguito.

tokenizzazione: suddivide un testo in singole parole, i*token*), che saranno utilizzati per altri tipi di analisi o attività.

standardizzazione: riscrittura delle parole da *upper case* a *lower case*.

rimozione delle *stopwords*, parole comuni prive di significato, ma che ricorrono spesso all'interno della frasi.

rimozione delle cifre numeriche

ID	Numero di recensioni	Titolo del dvd	Recensioni positive	Recensioni negative	Recensioni neutrali
B004XKRO6Q	1166	The Lord of the Rings - The Motion Picture Trilogy, Extended Edition	98.00%	1.00%	1.00%
B0041KWD6Y	122	Pomi d'ottone e manici di scopa	97.00%	2.00%	2.00%
B078H2XDQ2	264	Coco	96.00%	2.00%	2.00%
B00AGD6MUS	196	Quasi amici	95.00%	2.00%	3.00%
B0041KY2Q8	146	Gli Aristogatti	95.00%	3.00%	2.00%

Figura 2.4: Primi cinque dvd con più recensioni positive

ID	Numero di recensioni	Titolo del dvd	Recensioni positive	Recensioni negative	Recensioni neutrali
B0781YXHRH	152	Star Wars: Gli Ultimi Jedi	64.00%	25.00%	11.00%
B01BMCRNFY	135	Deadpool	74.00%	18.00%	8.00%
B019C00JYA	246	Star Wars: Il Risveglio della Forza	74.00%	14.00%	12.00%
B07C8FCK6G	174	Harry Potter Collection (Standard Edition) (8 Dvd)	86.00%	12.00%	2.00%
B0071AO594	157	Rocky - La Collezione Completa	80.00%	11.00%	9.00%

Figura 2.5: Primi cinque dvd con più recensioni negative

rimozione della punteggiatura, tuttavia nella nostra implementazione questa fase è subentrata all'interno della `tokenizzazione`.

stemming: processo di riduzione delle parole flesse (o talvolta derivate) alla loro forma di origine, base o radice.

Al termine del processo, l'*output* risultante era composto da un elenco di parole scritte Italiano corretto, privo di segni di punteggiatura o numeri, scritto in forma minuscola, ridotto alla radice. A questo punto è stato quindi possibile effettuare un'operazione di visualizzazioni delle parole più usate all'interno delle recensioni, per comprendere il motivo per cui proprio quei prodotti presenti nella lista rientrino nei migliori, per entrambe le polarità. La tecnica visiva utilizzata è quella dei `wordclouds`, che consiste nella raccolta delle parole più usate nelle recensioni (`word`), rappresentate per mezzo di `cloud`, ha permesso di confrontare le recensioni positive e negative per ogni categoria e di comprendere che cosa determinava l'appartenenza dei prodotti all'insieme dei "migliori" o dei peggiori.

2.2.1 Wordclouds libri

Nella Figura 2.8 sono mostrati i *wordclouds* ottenuti per la categoria libri. In particolare l'immagine di sinistra rappresenta le parole più usate

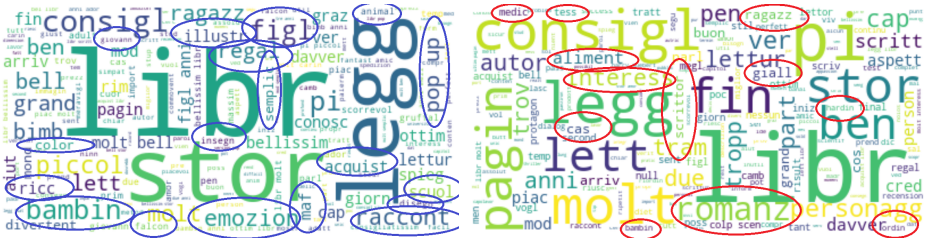
ID	Numero di recensioni	Titolo del videogioco-accessorio	Recensioni positive	Recensioni negative	Recensioni neutrali
B072N597H5	453	Super Mario Odyssey - Nintendo Switch	98.68%	0.88%	0.44%
B01N7QUSQ3	439	The Legend of Zelda: Breath of the Wild - Nintendo Switch	98.41%	0.23%	1.37%
B01MS894C1	406	Mario Kart 8 Deluxe - Nintendo Switch	98.28%	1.23%	0.49%
B01M8MO8XS	125	Red Dead Redemption 2 - Xbox One	97.60%	0.80%	1.60%
B07BZV8R5R	151	Super Smash Bros Ultimate - Nintendo Switch	97.35%	1.99%	0.66%

Figura 2.6: Primi cinque videogiochi con più recensioni positive

ID	Numero di recensioni	Titolo del videogioco-accessorio	Recensioni positive	Recensioni negative	Recensioni neutrali
B0746PN4PZ	198	Nacon Revolution Pro Controller 2, Nero - PlayStation 4	59.60%	30.81%	9.60%
B06XWYC4QP	140	Destiny 2 + DLC Esclusivo Amazon - PlayStation 4	55.71%	27.86%	16.43%
B0746R8TKN	130	Nacon Compact Controller, Blu - PlayStation 4	62.31%	24.62%	13.08%
B01BVEEJG4	580	Gran Turismo Sport - PlayStation 4	68.45%	23.10%	8.45%
B00499DB7W	436	Two Dots Universal Pro Driving Simulator Supporto universale ripiegabile	57.34%	21.33%	21.33%

Figura 2.7: Primi cinque videogiochi con più recensioni negative

nelle recensioni positive, mentre a destra sono rappresentate quelle negative.



(a) Recensioni positive (b) Recensioni negative

Figura 2.8: Wordclouds libri

Recensioni positive: nello schema sono rappresentate con dimensione maggiore le parole `libr`, `stori`, `legg`, a dimostrazione del fatto che queste sono in assoluto le parole più frequenti all'interno delle recensioni degli utenti. Non sorprende che i tre termini si riferiscano alla categoria selezionata, nel caso attuale quella dei libri; vedremo che questo comportamento ricorrerà sempre durante l'analisi tramite textitwordclouds. Concentriamoci invece sulle parole che racchiudono informa-

zioni, ricordando che l'obiettivo primario è comprendere perché i libri di Figura 2.2 siano risultati i migliori.

- `ricc`, `figl`, `falcon`, `giovann`, `"emozion"`, `pop up`:
  sono tutte parole presenti in alcuni dei titoli della lista dei migliori, in ordine *"Per questo mi chiamo Giovanni. da un padre a un figlio il racconto della vita di Giovanni Falcone"*, *"I colori delle emozioni (pop-up)"*, "*25 grammi di felicità - Come un piccolo riccio può cambiarti la vita*".
- `maf`:
  termine che se esteso indica la mafia, riferito quindi al primo libro dell'elenco.
- `illustr`, `bambin`, `color`, `semplic`, `disegn`, `raccont`, `animal` :
  sono tutti termini riguardanti l'infanzia. Da questa informazioni comprendiamo che la maggior parte delle recensioni commenta positivamente i libri per bambini, che sono semplici, colorati, ricchi di disegni e di racconti basati sugli animali. Quindi questi vocaboli sono riferiti per lo più al secondo e al terzo libro dell'elenco dei migliori; tuttavia l'ultimo termine può essere esteso anche quarto prodotto.
- `insegn`, `scuol`:
  sono vocaboli riferiti all'ambiente scolastico; quindi probabilmente riferito all'ultimo libro, criticato positivamente poiché insegna tematiche relative alla scuola.
- `acquist`, `regal`:
  molto probabilmente alcuni libri, acquistati per un regalo, hanno suscitato delle impressioni positive nell'utente, che ha quindi recensito positivamente.

Recensioni negative: nello schema sono rappresentate con dimensione maggiore le parole `libr`, `pagin`, `legg`, a dimostrazione del fatto che queste sono in assoluto le parole più frequenti all'interno delle recensioni degli utenti. Notiamo che due su tre sono parole già trovate all'interno dell'elenco delle parole positive, poiché anche in questo caso sono termini utili a determinare la categoria in questione. Studiamo ora le altre parole in relazione all'elenco già stilato (Figura 2.3).

- `hardin`, `tess`, `romanz`:
  sono i termini più espliciti dell'insieme, indicano infatti i nomi dei protagonisti del romanzo "*After*"
- `ordin`, `cas`:
  sono parole usate nel secondo prodotto dell'elenco

- `scrittor`, `giall`:
  si riferiscono probabilmente a "*La verità del caso di Harry Quebert*", un giallo che vede protagonista uno scrittore. `bambin`:
  fra tutti i prodotti dell'elenco quello che certamente si adatta alle linee giovanili è sicuramente il quarto libro dell'elenco, pieno di attività creative e non proposte ai più piccini. `aliment`:
  vocabolo riferito a "*Vivere 120 anni. la verità che nessuno vuole raccontarti*". `medic`, `ragazz`, `interessant`:
  tra i proposti sono quelli che più di tutti racchiudono le possibili cause di una critica. Il libro, presupponibilmente l'ultimo dell'elenco, probabilmente aveva un approccio troppo rivolto alla medicina o forse troppo poco. Il secondo è forse più rivolto ad "*After*", criticato spesso per rivolgersi esclusivamente a un pubblico femminile. Più generico, invece, è il vocabolo `interessante`, che ben si adatta a ognuno dei prodotti proposti.

2.2.2 Wordclouds dvd

Nella Figura 2.9 sono mostrati i *wordclouds* ottenuti per la categoria dvd. In particolare l'immagine di sinistra rappresenta le parole più usate nelle recensioni positive, mentre a destra sono rappresentate quelle negative.

(a) Recensioni positive (b) Recensioni negative

Figura 2.9: Wordclouds dvd

Recensioni positive: nello schema sono rappresentate con dimensione maggiore le parole `dvd`, `film`, `blu ray`, riferite genericamente alla categoria presa in esame.

- `signore anell`, `tolkien`, `tre film`:
  I termini selezionati sono accomunati dalla tematica "*Signore degli Anelli*", trilogia ispirato all'omonimo romanzo scritto da *Tolkien*, che ha raccolto molti consensi, dato che più parole lo caratterizzano.
- `bambin`:
  I dvd criticati positivamente sono stati quelli adatti a un gio-

vane pubblico. Da notare che il termine risponde ancora alle caratteristiche del "*Signore degli Anelli*".

- `version estes`:
  Tra i tanti comprati o visti, i dvd in versione estesa hanno riscontrato un consenso positivo.

Si noti che le parole trovate ben si accordano con i risultati trovati nei passaggi precedenti (Figura 2.4), dove il miglior prodotto era la trilogia del *"Signore degli anelli*", seguito poi da altri film adatti ai bambini ("*Coco*", terzo posto, e "*Gli Aristogatti*", ultimo posto, sono addirittura dei cartoni animati).

Recensioni negative: nello schema i termini generici riscontrabili con la categoria in questione sono `film`, `dvd` , `blu ray`. Poiché nella Figura 2.5, abbiamo trovato cinque prodotti di diversa natura come maggiormente criticati, ci aspettiamo di trovare termini che li identifichino e che argomentino le recensioni negative.

- `war`, `marvel`, `rocky`:
  sono tutti termini utili a identificare il dvd in questione. Come ci aspettavamo, il primo di questi si riferisce al film di "*Star Wars*", il primo peggio criticato, il secondo invece riguarda "*Deadpool*", un film della marvel e il terzo del film "*Rocky*".
- `ser`, `episod`, `vecc`:
  In queste parole sono racchiuse le cause delle critiche. I primi due probabilmente rivolti a "*Star Wars*", una serie di nove film, ognuno dei quali denominato con il termine di "episodio". Il terzo vocabolo, che molto probabilmente si rifesce a "*Rocky*", criticato invece per l'età.

2.2.3 Wordclouds videogiochi

Nella Figura 2.10 sono mostrati i *wordclouds* ottenuti per la categoria videogiochi. In particolare l'immagine di sinistra rappresenta le parole più usate nelle recensioni positive, mentre a destra sono rappresentate quelle negative.

Recensioni positive: nello schema sono rappresentate con dimensione maggiore le parole `gioc`, `grafic`, riferite genericamente alla categoria presa in esame, di cui si apprezza la natura del gioco e la grafica. Per quanto riguarda invece il contenuto informativo, soffermiamoci sulla Figura 2.6; dalla quale apprendiamo quali sono i cinque prodotti migliori e cerchiamo di ricostruire cosa li caratterizza.

- `super mar`, `mar kart`, `zeld`:
  I termini selezionati rappresentano infatti il podio dei migliori,

(a) Recensioni positive (b) Recensioni negative

Figura 2.10: Wordclouds videogiochi

rispettivamente "*Super Mario Odyssey*", "*The Legend of Zelda*" e *Mario Kart 8 Deluxe*".

- `nint switch`:
  i videogiochi più apprezzati sono quelli per *ninntendo switch*; non a caso infatti i tre migliori sono proprio accomunati da questa caratteristica.
- `avventur`, `online`, `modalit`, `open world`, `livell`, `gameplay`:
  tutti questi vocaboli sono caratteristiche positive dei giochi; probabilmente apprezzati quindi dalle avventure proposte dalla storia, dalla possibilità di giocare online in più modalità, compresa anche quella esplorativa, organizzate in livelli o dal *gameplay* intuitivo, etc.
- `prezz`: l'appartenenza dei giochi alla lista dei migliori può anche essere stata determinata dal prezzo, trovato forse dagli utenti accessibile o anche basso.

Recensioni negative: nello schema i termini generici più rilevanti per la categoria sono `gioc`, `prodott`. Concentriamoci ora sulla *top five* dei prodotti negativi (Figura 2.7), contrariamente ai casi precedenti abbiamo rappresentati tra i pprimi cinque prodotti ci sono sia videogiochi, sia accessori appartenenti alla categoria in questione.

- `gran turism`, `destiny`, `controller`:
  sono tutti termini appartenenti all'elenco dei primi cinque prodotti con più recensioni negative.
- `gt`, `pedalier`, `macchin`, `volant`:
  Sono parole riferite al gioco di corse automobilistiche "*Grand Turism*", apprezzato soprattutto quando si gioca con un volante o una pedaliera.
- `grafic`, `stabil`: i giochi sono probabilmente apprezzati per la grafica efficace e con il termine `stabile` ci si vuole forse riferire alla qualità da parte del gioco di mantenere la fluidità dell'immagine priva di scatti, o anche la capacità del gioco di non "*crashare*".

- `prezz`: anche in questo caso il prodotto è stato recensito rispetto al suo prezzo, contrariamente a prima però questa volta trovato troppo alto.

2.3 Correlazione tra prezzo e sentiment

Avendo trovato un riscontro positivo alle domande poste, siamo passati a una seconda analisi del *sentiment* sul *dataset* prodotti. In particolare ci siamo chiesti se il prezzo associato a ogni prodotto, fosse determinante per la polarità delle recensioni. Esiste infatti un'idea duale legato a un prezzo ridotto o eccessivo. Nel primo caso un prezzo ridotto può essere fonte di:

- **reazioni positive**, legate all'esigua quantità di denaro spesa per un prodotto funzionante.
- **reazioni negative**, legate alla qualità infima dei materiali.

Medesimo comportamento è legato ad alti costi. In questo caso abbiamo:

- **reazioni positive**, perché il prodotto oltre che a essere funzionante è di ottima qualità.
- **reazioni negative**, dovute al costo eccessivo di alcuni prodotti, dove si paga non tanto la qualità, quanto piuttosto il marchio.

Bibliografia