Text Mining Coursework

Analysis on Amazon Alexa reviews

Matteo Fasulo

04/7/2021

Contents

Introduction

L'analisi riportata ha come oggetto le recensioni di Amazon Alexa Echo Dot (4th Gen). Nonostante il prodotto abbia oltre 275,141 global ratings, Amazon permette di visualizzare solo le prime 500 pagine di recensioni ognuna contenente 10 recensioni. A partire da questo dataset ho proposto diverse analisi tra cui:

Analisi in frequenza:

- Parole con maggior frequenza
- Rappresentazione grafica del rating in stelle
- Analisi dei voti utili delle recensioni
- Confronto tra acquisti verificati e non verificati
- Wordcloud delle parole più frequenti

Analisi del Sentiment:

- Empirical Distribution Function del Sentiment
- Rappresentazione della valenza emotiva in funzione del tempo
- Rappresentazione della media del Sentiment per ogni categoria di rating in stelle
- Analisi dei "falsi negativi" e "falsi positivi"
- Analisi della percentuale di polarità delle recensioni
- Andamento del Sentiment in funzione del tempo

Topic Modelling:

- Modello BTM (Biterm Topic Model)

Web Scraping

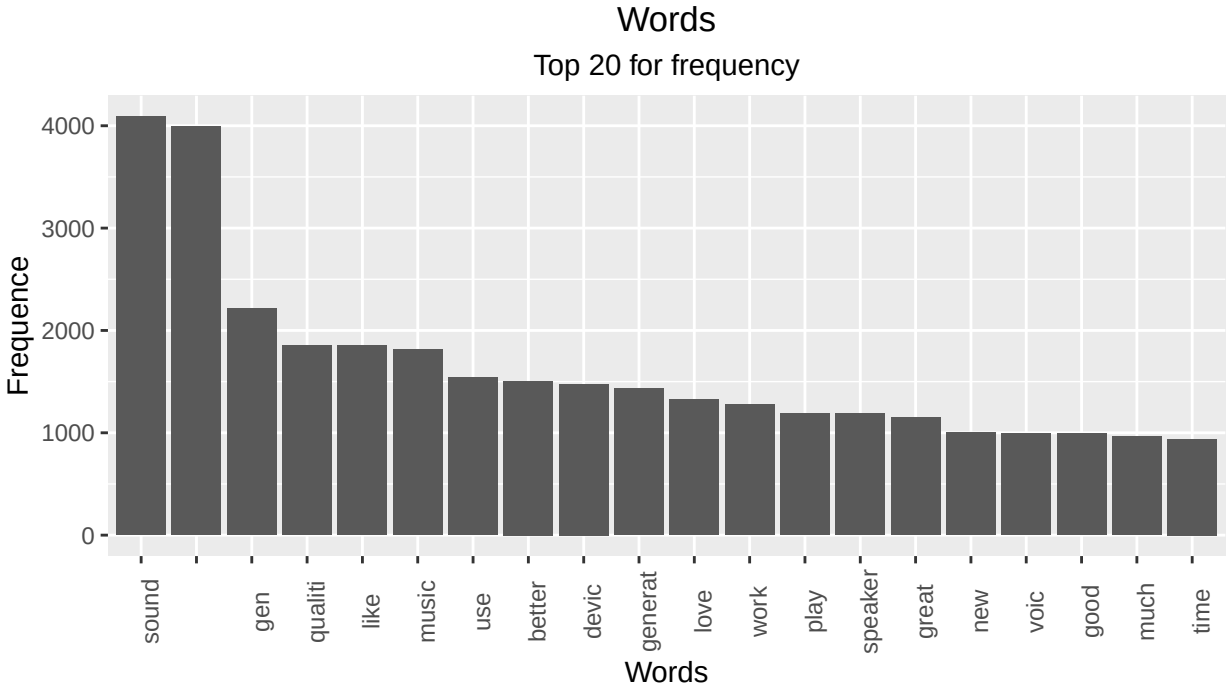
Per ottenere le recensioni di Alexa ho sviluppato uno Scraper Web basato sulla libreria *rvest*. In questo modo sono riuscito ad ottenere le informazioni già in formato *data.frame* analizzando il testo presente nei singoli *div* delle pagine *html*. Lo scraping è stato limitato alle prime 500 pagine, ognuna contenente 10 recensioni. Per evitare qualunque incoveniente, il file "alexa_echo_dot.csv" contiene il dataset in formato *.csv* aggiornato all'ultima data di scraping. Purtroppo per problemi di conversione le emoji presenti nel testo originario di Amazon, una volta esportate in csv, non vengono visualizzate correttamente. Per evitare una possibile limitazione alle richieste http, ho inserito anche un *throttle* tra una richiesta e la successiva.

Preprocessing

Una volta ottenuto il dataset "review" sono passato alla fase di "pulizia" dei dati. Poiché non tutte le analisi richiedono lo stesso tipo di purità, ho sviluppato diverse funzioni in base alle necessità. La fase di preprocessing prevede prima la conversione del testo a Vettore Corpus e successivamente la rimozione di numeri, punteggiatura e stopwords. Per questo motivo ho creato due dataset, il primo con solo le operazioni di rimozione di numeri, punteggiatura e stopwords mentre il secondo con stemming. A tale proposito ho utilizzato le librerie *tm* (VCorpus(), tm_map()), *stopwords* e *syuzhet* (get_sentences()).

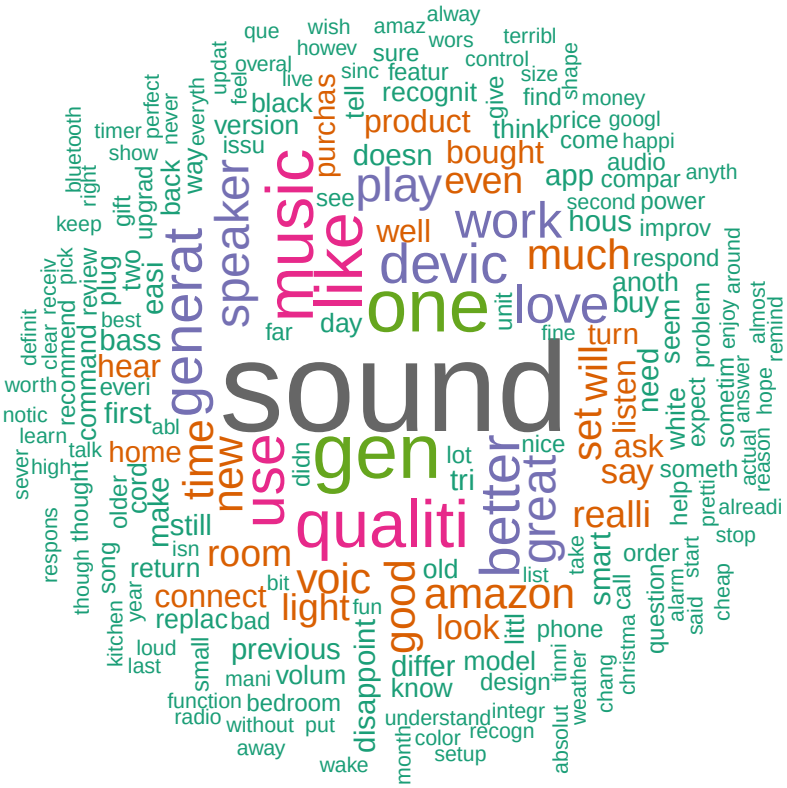
Analisi in frequenza

La prima analisi che ho effettuato di questa sezione sono le prime 20 parole per numero di occorrenze nel *dataset*. A partire dal dataset con le recensioni, ho selezionato le singole parole e, una volta considerate come *table*, ordinate per frequenza decrescente. Si noti come la parola "sound" sia la più frequente e rappresenti la caratteristica principale tra tutte le recensioni. Tutte le parole sono state prima "filtrate" da una fase di stemming per portarle ad una forma comune laddove vi erano più "declinazioni" dello stesso lemma.

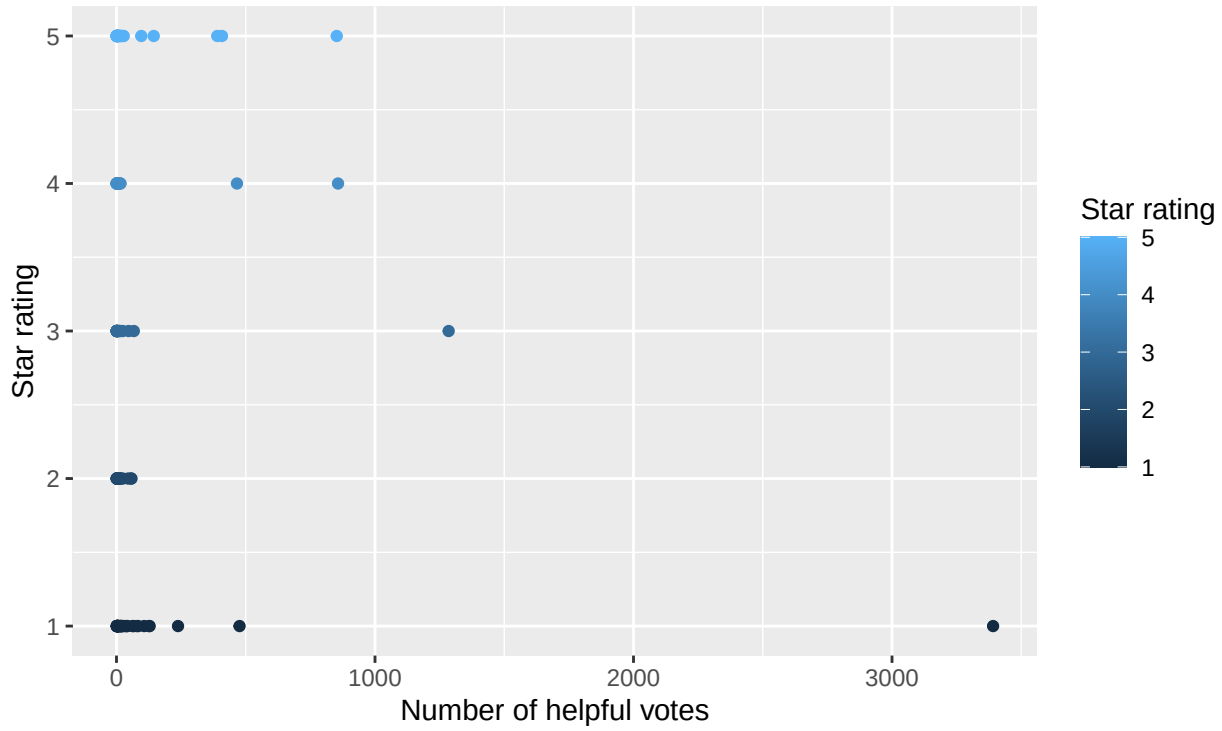


(based on previous stemming phase)

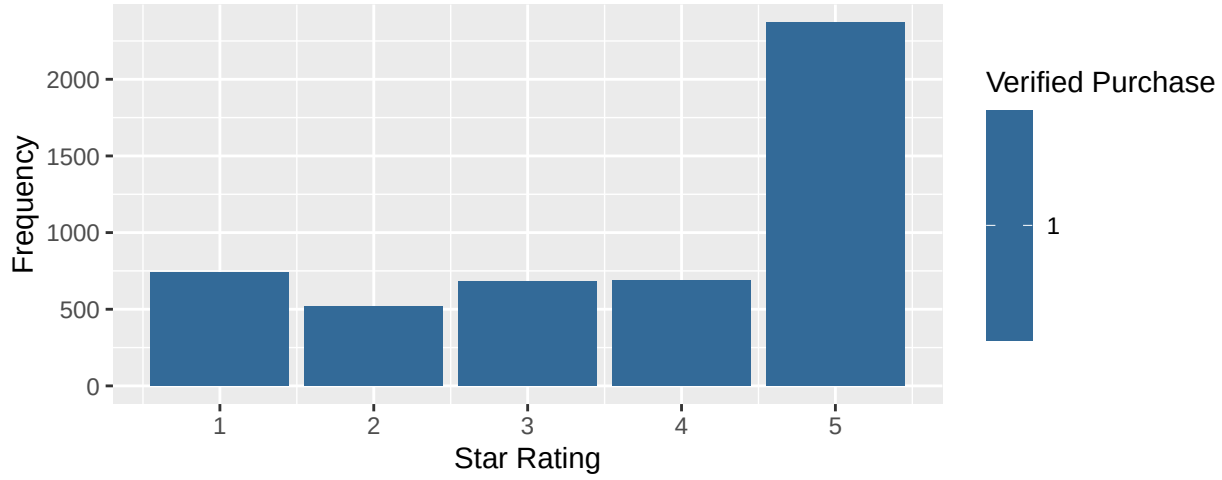
Lo stesso risultato ottenuto nella precedente analisi può essere visualizzato anche tramite il WordCloud. E' una rappresentazione più compatta dove è possibile raffigurare la frequenza delle top 200 parole. In questo tipo di rappresentazione grafica, la grandezza del carattere è proporzionale al numero di occorrenze di tale parola nel testo di appartenenza. Nonostante sia una rappresentazione statisticamente poco informativa, ha un forte impatto visivo.



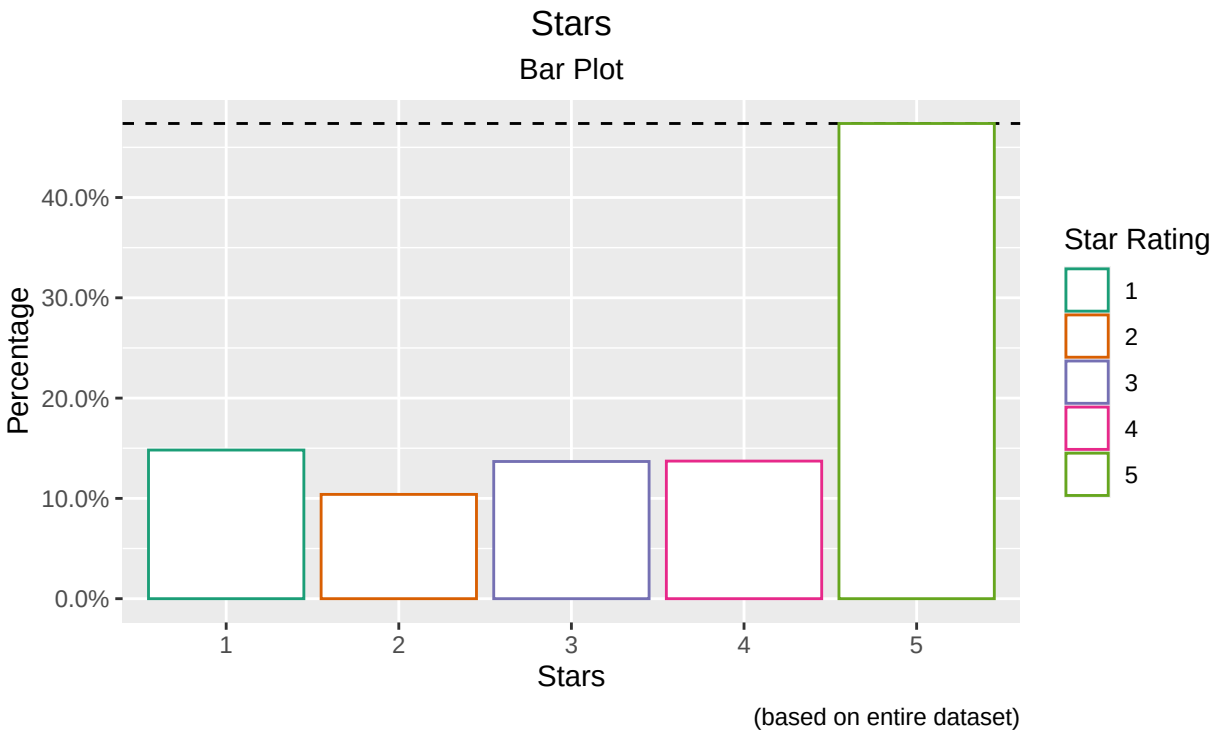
Durante la fase di Web Scraping sono inoltre riuscito ad ottenere anche il numero di voti utili in ogni recensione presente su Amazon. Ho analizzato il numero di voti utili espresso nelle recensioni confrontandolo con il rispettivo rating di appartenenza in stelle. Sfortunatamente solo poche recensioni all'interno del dataset erano considerate utili e votate dagli altri recensori, tuttavia il commento con più voti utili (oltre 3000) ha una sola stella di rating.

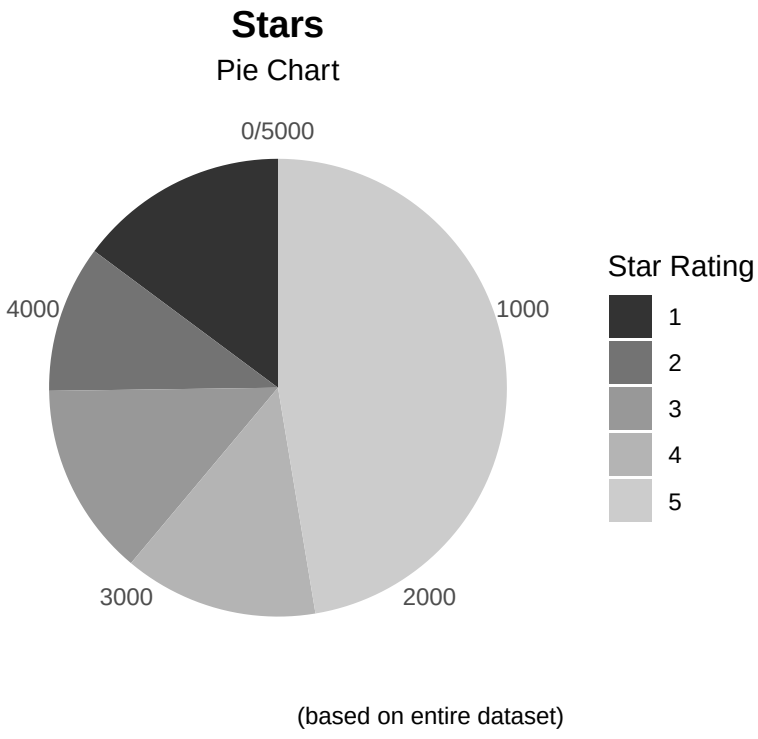


Con l'intenzione di voler effettuare un test sulle medie tra prodotti con acquisto verificato e non, ho analizzato il numero di acquisti verificati notando che esso comprende tutte le 5000 recensioni (essendo Alexa un prodotto progettato da Amazon ci sarà una maggiore attenzione nella scelta delle review da esporre).



Sono passato poi alla fase di analisi del rating in stelle. A tal caso ho deciso di rappresentarlo sia in forma di grafico a barre (barplot) sia in forma di diagramma a torta (piechart). Dalla rappresentazione è chiaro che la maggior parte dei commenti riporta una valutazione di 5 stelle su 5.

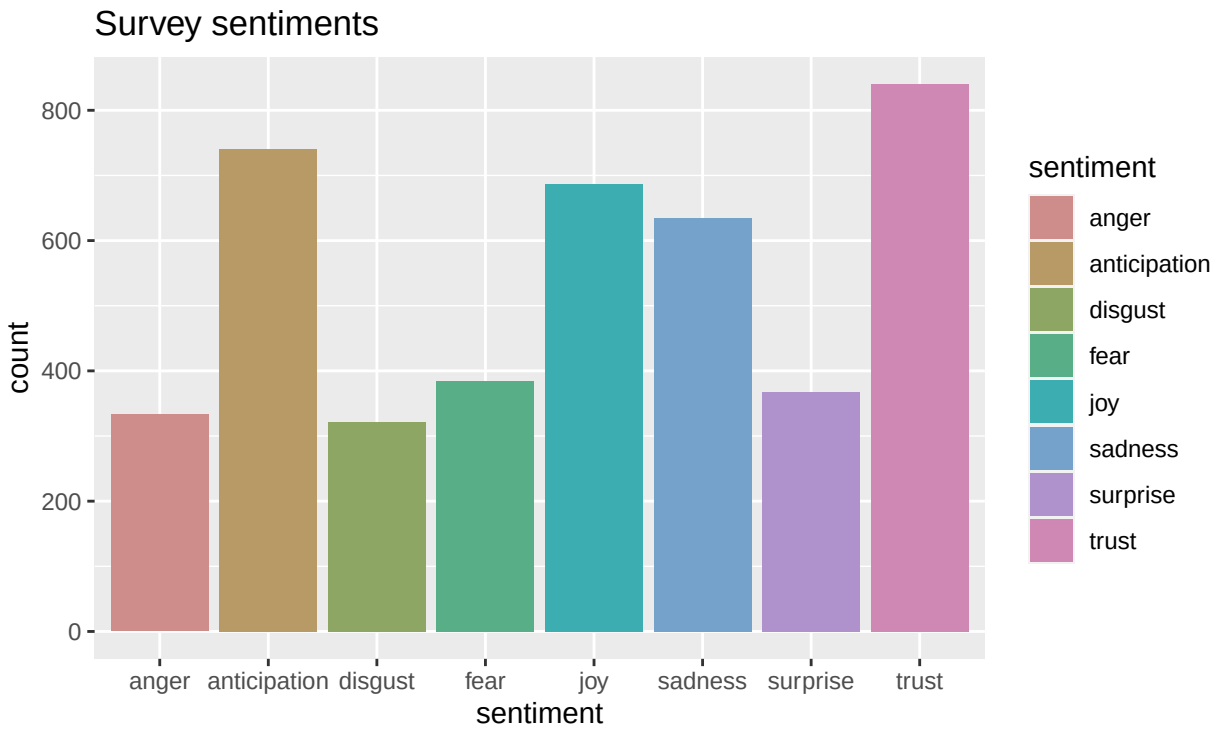




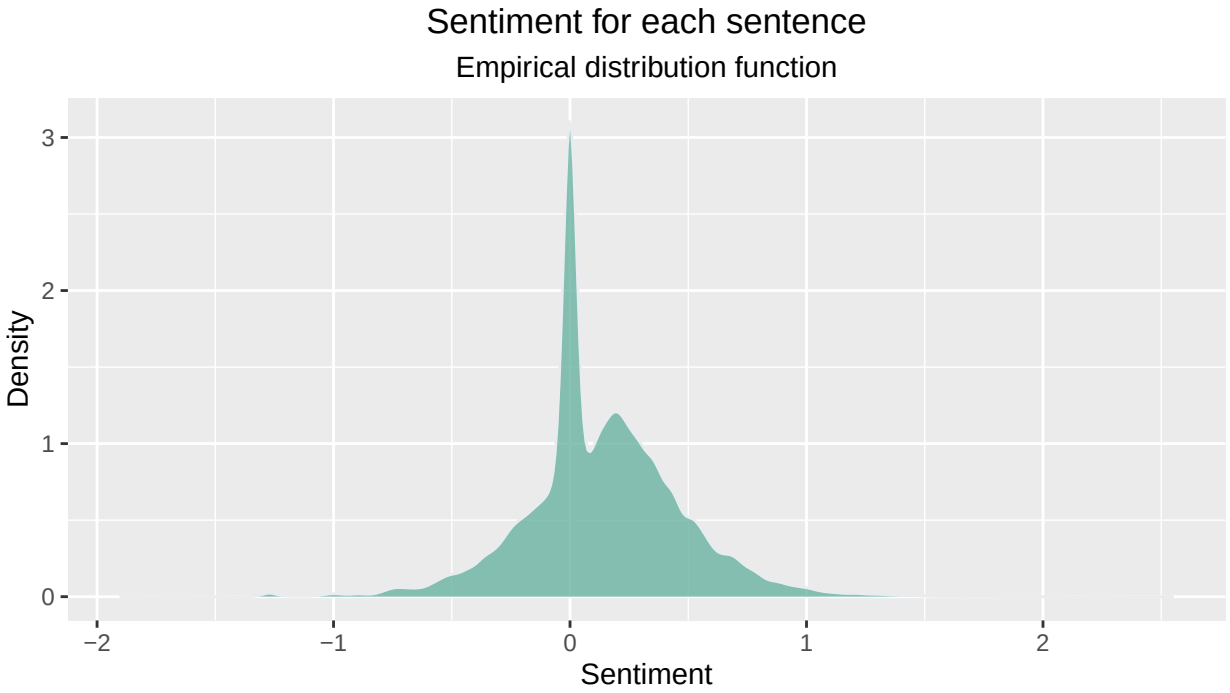
Analisi del Sentiment

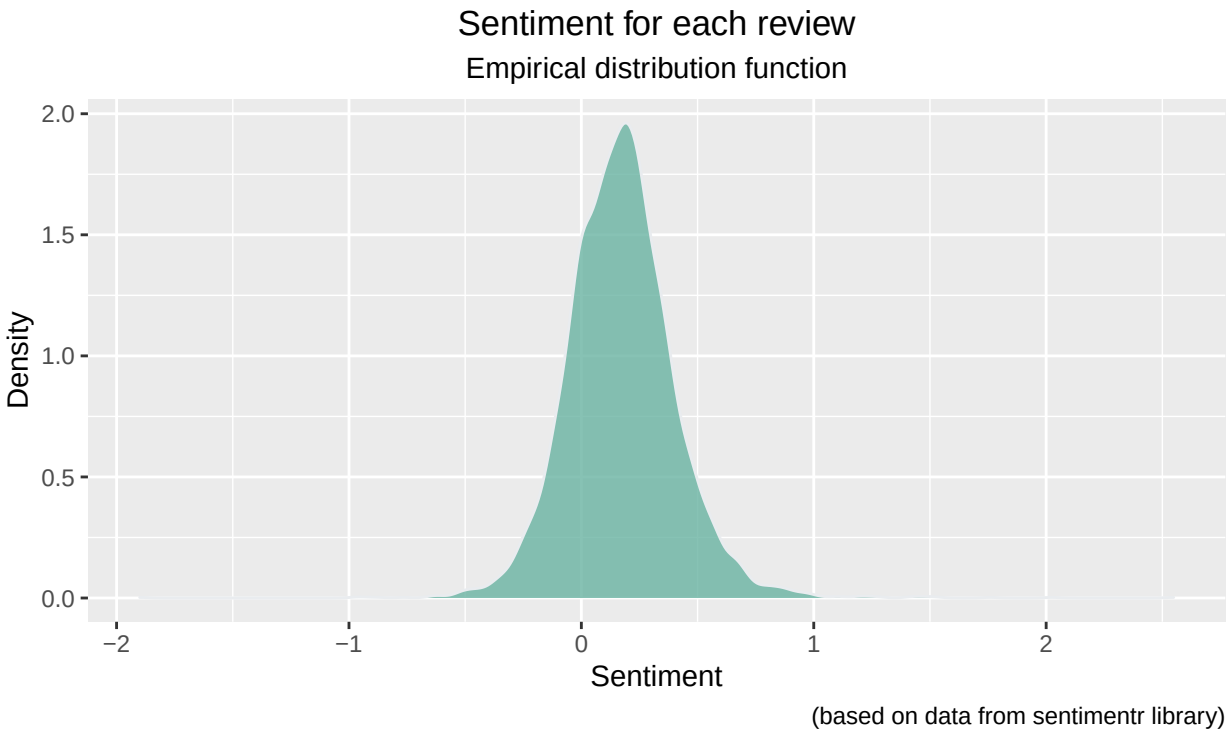
Per Sentiment Analysis si intende quel processo tramite cui è possibile comprendere e misurare il grado di soddisfazione di un giudizio espresso a parole. E' una tecnica molto utilizzata a livello decisionale per

interpretare se il prodotto è di gradimento ai clienti. Dovendo quindi analizzare il giudizio a parole, è fondamentale sapere quanto ogni parola influenzi il valore del sentiment. Le librerie citate in questo contesto usano tutte un approccio *bag-of-words* dove il sentiment è determinato sulla base delle singole parole nel testo. Le parole vengono poi confrontate con dei lessici dove i termini positivi e negativi sono associati ad un grado di intensità. Librerie come *syuzhet* si limitano ad analizzare le singole parole e non sempre questo approccio fornisce il miglior risultato, infatti per l'analisi del sentiment ho scelto di utilizzare la libreria *sentimentr* che tra le librerie di Sentiment Analysis ha la maggiore accuratezza. SentimentR adotta i *valence shifter* (negators e amplifiers/deamplifiers) che invertono, aumentano o diminuiscono la polarità delle parole. Tuttavia per alcune analisi sulle emozioni percepite nelle recensioni, ho usato il *lexicon nrc* reperibile tramite *syuzhet*. Ho voluto prima classificare le emozioni presenti nelle recensioni attraverso tale lessico.



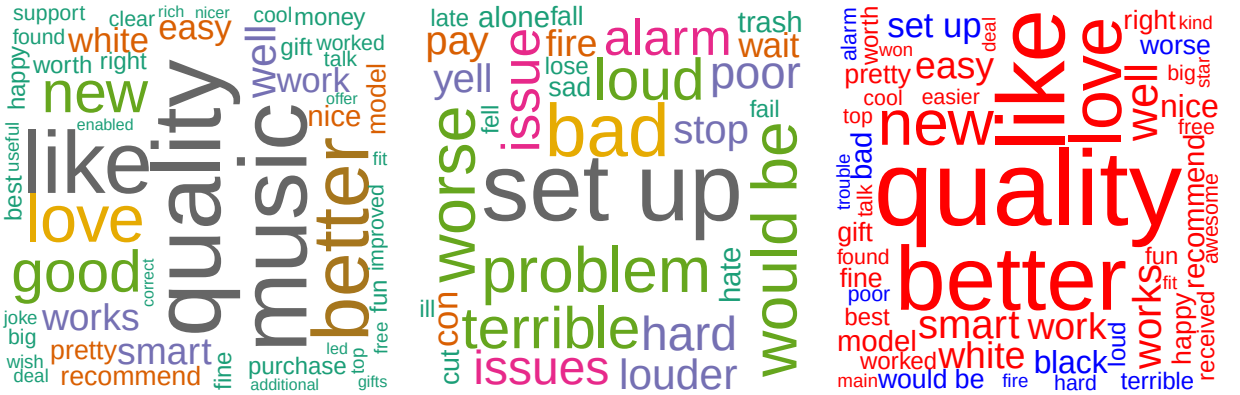
Successivamente ho sviluppato due diverse Empirical Distribution Function del Sentiment sia per singole frasi che per recensione. Queste due curve sottostanti approssimano la distribuzione del sentiment che nel caso relativo alle singole frasi presenta un'area minore nella regione di piano con sentiment negativo coerentemente con quanto riportato nell'analisi grafica di tutte le recensioni in cui l'area del semiasse negativo di sentiment risulta visibilmente minore rispetto al semiasse positivo.



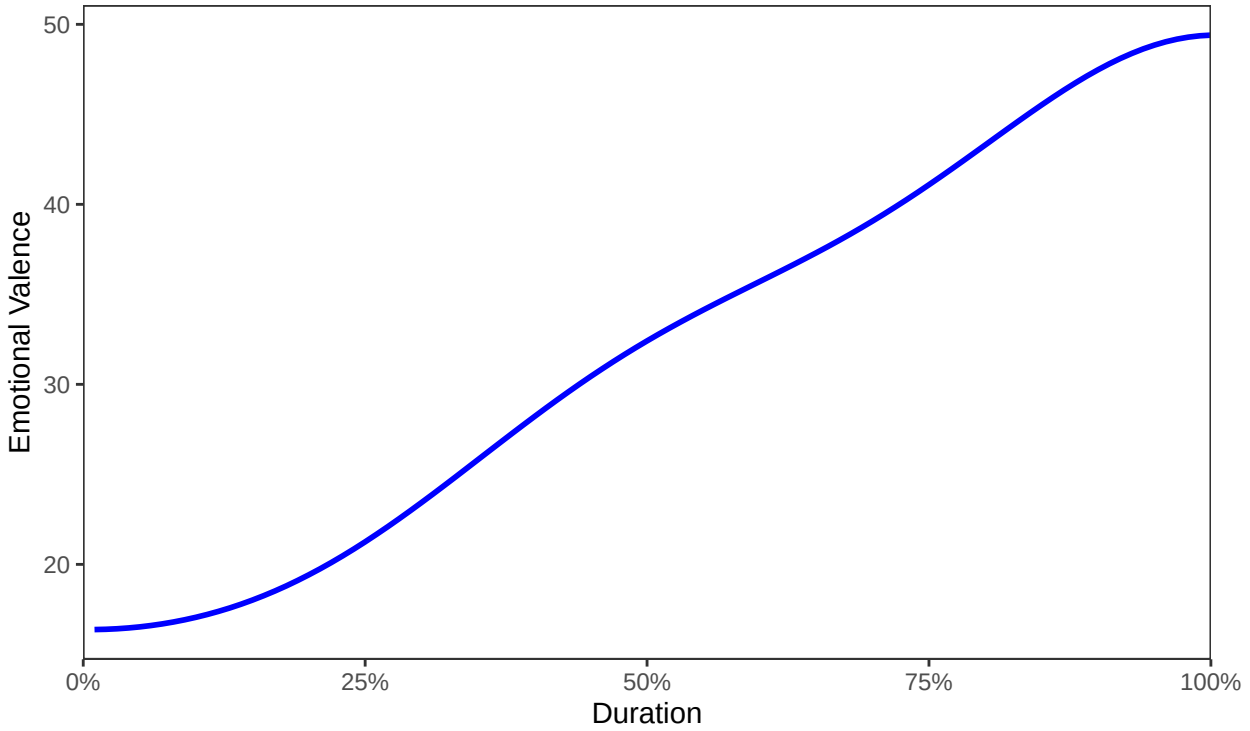


Attraverso la funzione *extract_sentiment_terms()* ho generato tre diversi WordCloud con Positive, Negative e Positive/Negative words. Coerentemente con gli altri dati, dal 3° modello si evince che le parole con maggiore occorrenza sono per lo più di grado positivo.

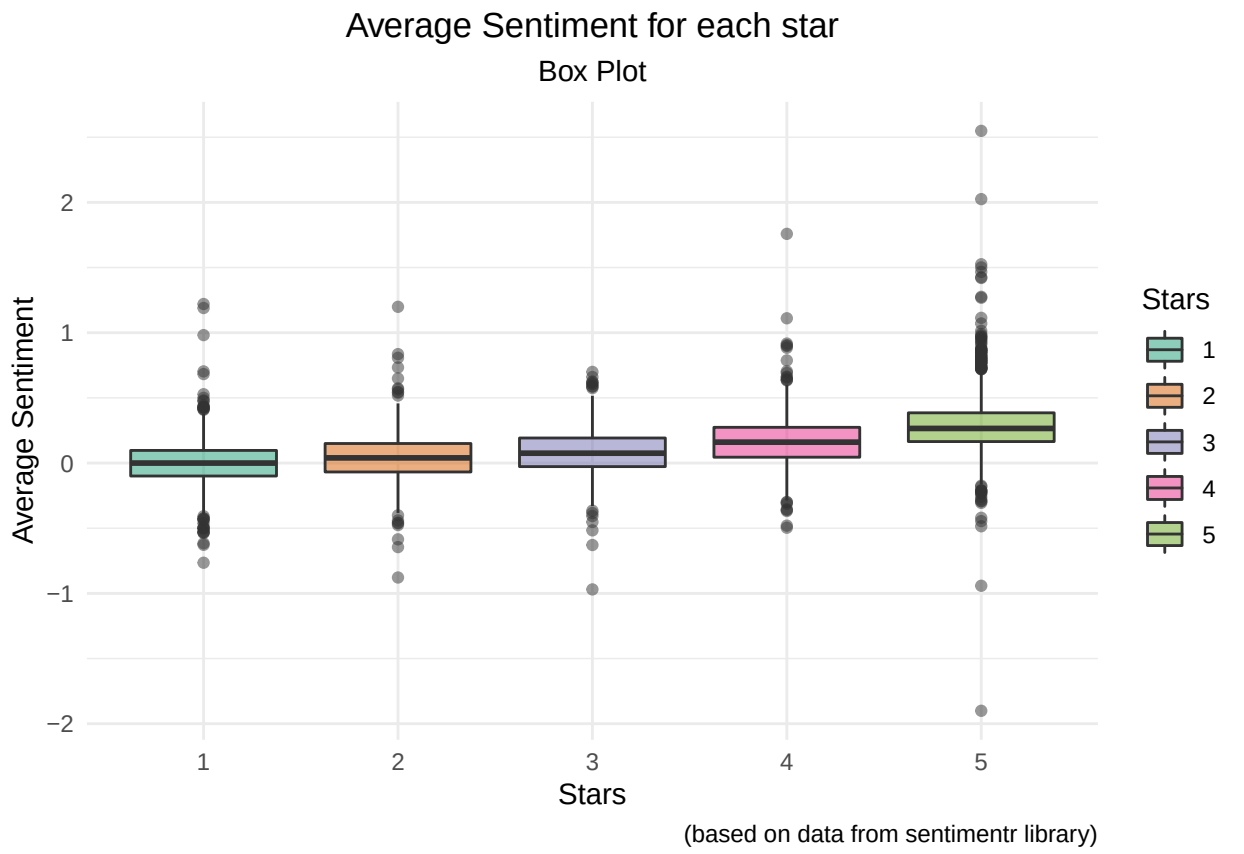
Positive Words Negative Words Positive(red) & Negative(blue) Wo



La rappresentazione sottostante è ottenuta tramite la funzione *get_transformed_values()* di *syuzhet* che attraverso la trasformata di Fourier genera una curva smussata e normalizzata della valenza emotiva lungo tutto l'arco del dataset. Nel paper della liberia syuzhet viene definito come: *shape smoothing and normalization using a Fourier based transformation and low pass filtering.*



Usando la rappresentazione a Boxplot ho poi notato come le recensioni con 1 stella apportino un contributo neutro al sentiment mentre a partire già da recensioni con 2 stelle abbiamo un aumento verso il grado positivo fino ai commenti con 5 stelle. Notiamo inoltre come vi siano molti *outliers* sia in corrispondenza di 1 stella che di 5 stelle.

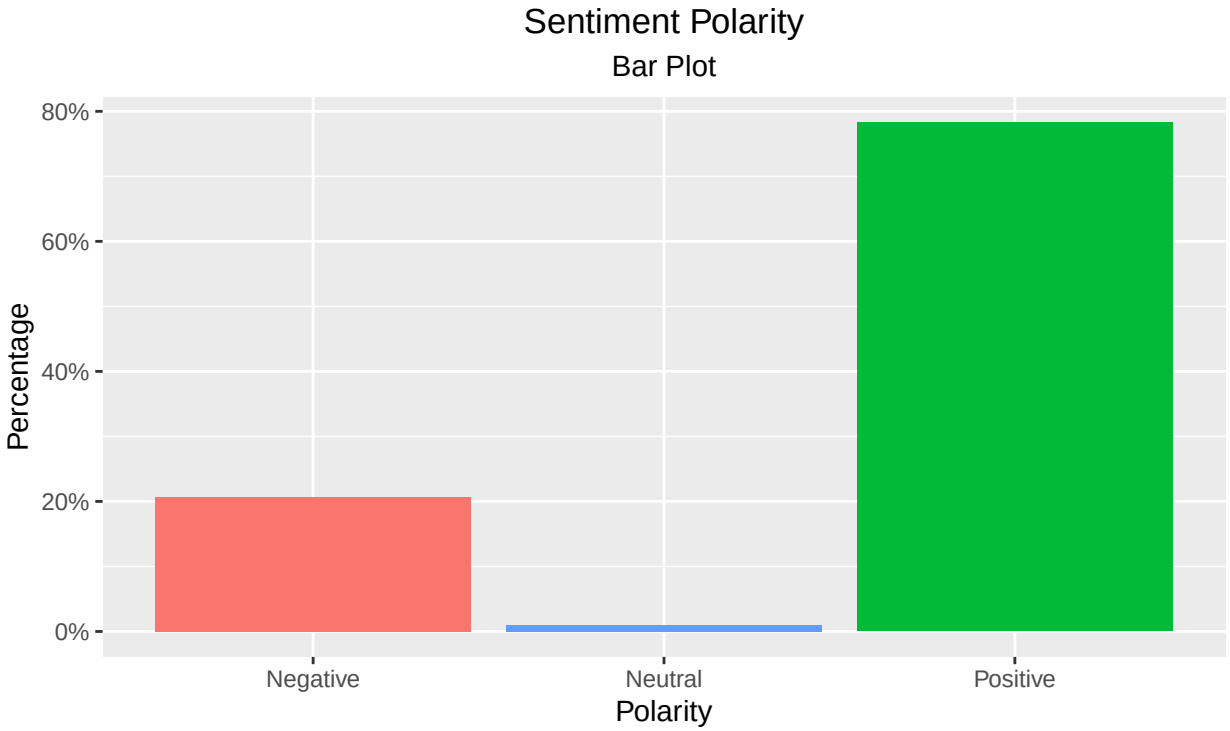


Le discrepanze sopra citate solitamente sono dovute a valutazioni errate da parte del recensore o da recensioni di stampo sarcastico (spesso difficili da individuare). L'analisi relativa agli outliers è stata svolta nel successivo grafico in cui ho confrontato tutte le possibilità:

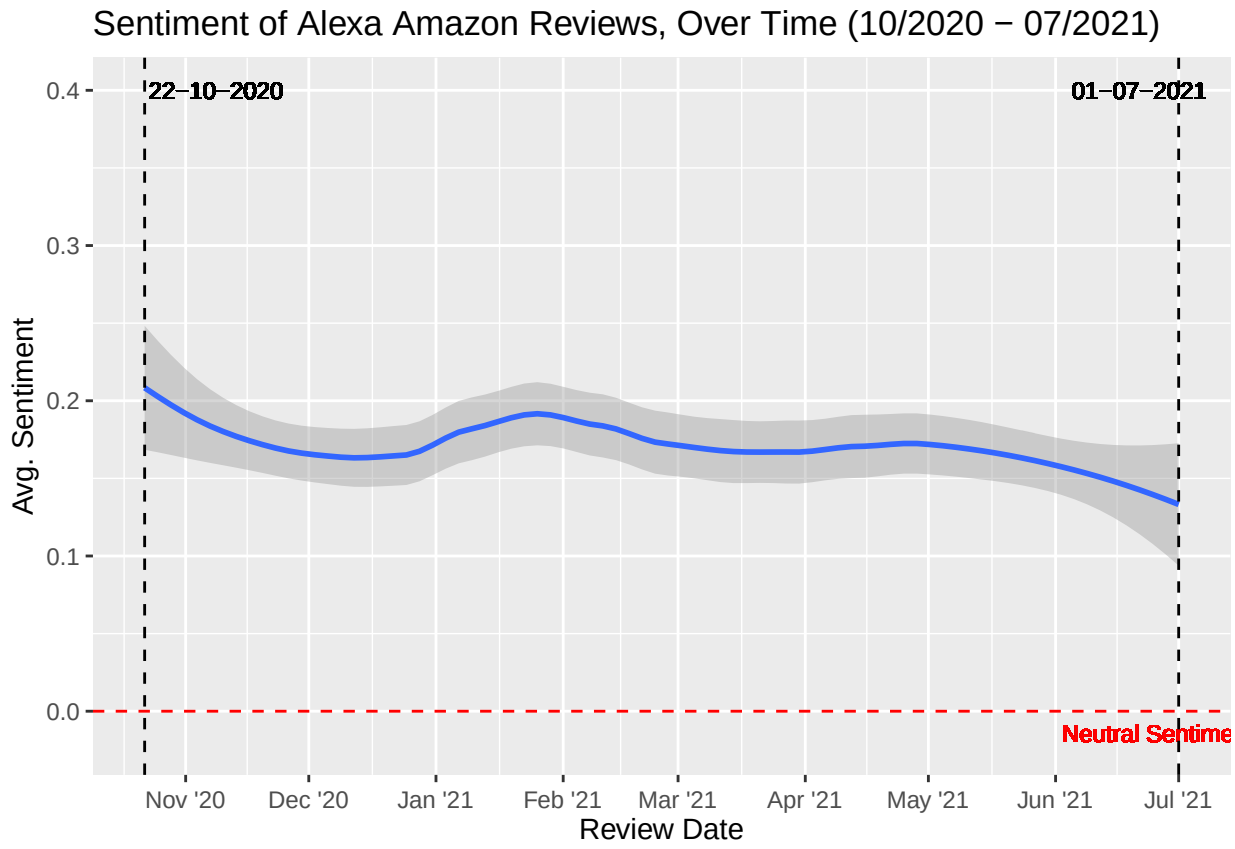
- Positive Review - Positive Sentiment
- Positive Review - Negative Sentiment
- Negative Review - Positive Sentiment
- Negative Review - Negative Sentiment

Ho scelto come soglia di recensioni positive tutte quelle superiori a 3 selle dividendo lo spazio in 4 piani.

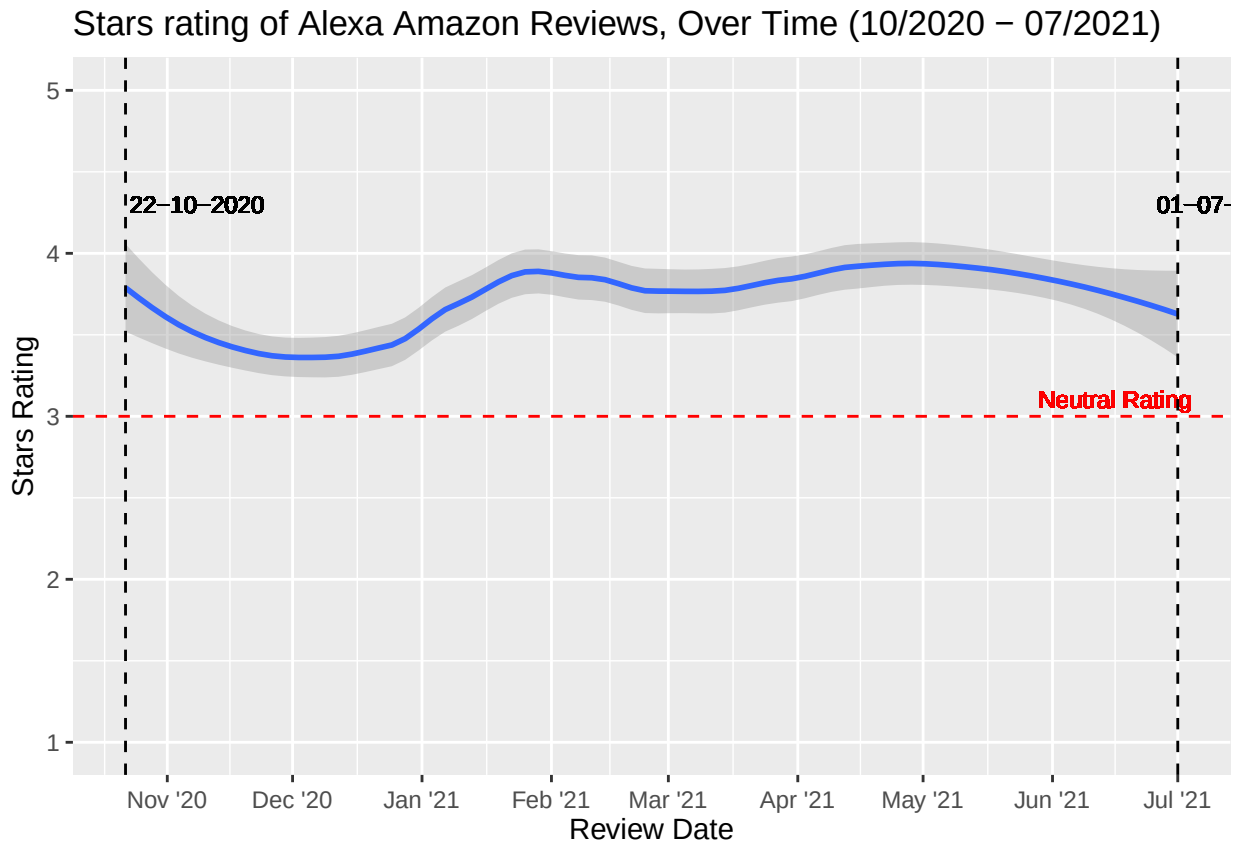
E' stato poi condotto uno studio sulla polarità delle recensioni classificandole in negative, neutre e positive sulla base del sentiment per recensione. Il risultato conferma tutte le supposizioni fin ora descritte inclusa quella che il maggior numero di recensioni è polarizzato positivamente.



Analizzando poi la media del sentiment nell'intervallo compreso tra Ottobre 2020 e Luglio 2021, si nota che la variazione totale è compresa tra 0.1 e 0.25 con un calo nel trend di positività nell'ultimo periodo. Sfortunatamente Amazon non mette a disposizione recensioni antecedenti a tale periodo, perciò l'analisi è relativa solo agli ultimi 10 mesi.



Discorso analogo per quanto rigurada la distribuzione dei rating in stelle al variare del tempo nello stesso intervallo in cui tendenzialmente si è sempre rimasti sulla media di 4 stelle su 5 tranne nel periodo a cavallo fra Novembre e Febbraio. Verificando in un secondo momento, ho notato che Amazon ha prodotto un modello di Echo Dot di colore *Twilight Blue*. Potrebbe trattarsi sempre del discorso relativo al colore del cavo di alimentazione di Alexa.



Topic Modelling

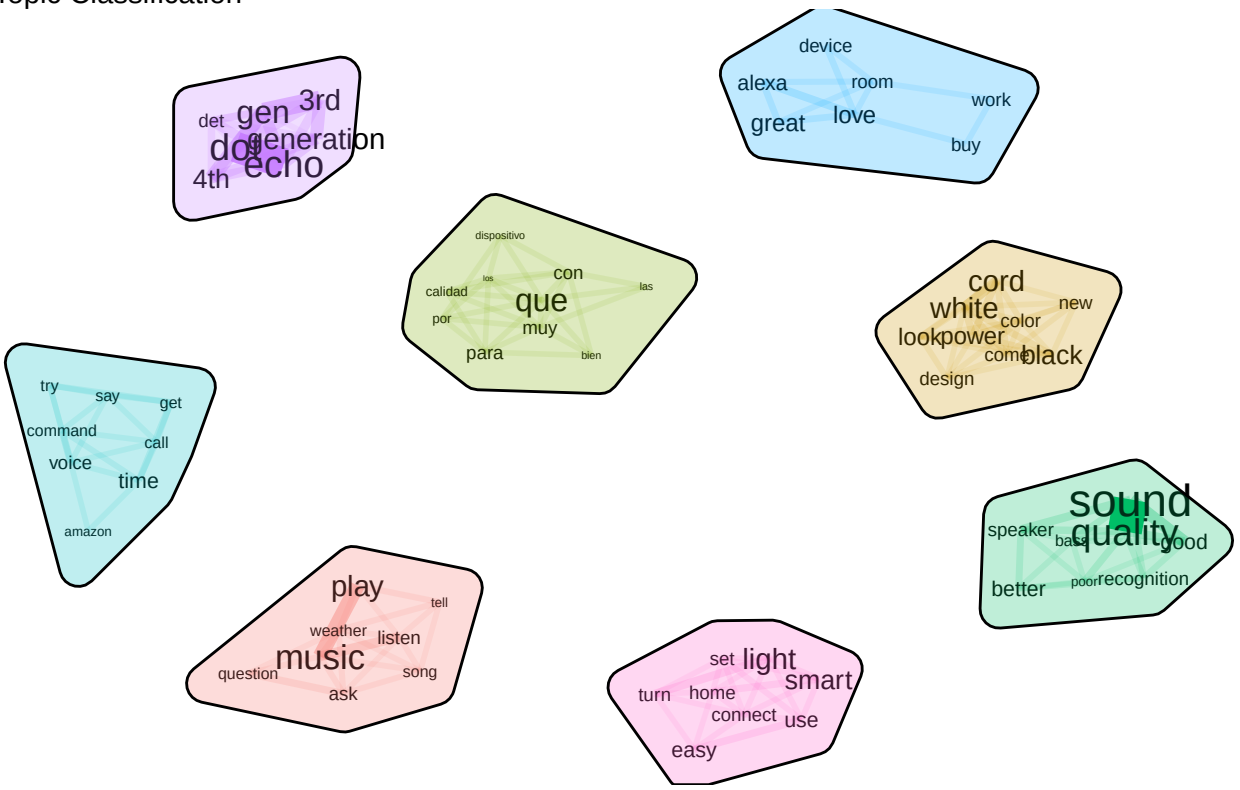
Il Topic Modelling è la tecnica che permette di individuare gli argomenti o *topic* di cui tratta un certo testo. Rientra nelle tecniche di Unsupervised Learning ovvero la branca di modelli di Machine Learning che non richiede addestramento o *training*. Tramite questo metodo possiamo associare ogni documento ad un determinato numero di topic permettendo di aggregare documenti simili che trattano gli stessi topic. Il più noto tra gli algoritmi di Topic Modelling è *LDA* (Latent Dirichlet Allocation). Per la nostra analisi però il modello LDA presenta delle criticità tra cui il fatto che per addestrare il modello occorre lavorare sulle occorrenze nel singolo documento e in documenti brevi (come le recensioni) ci sono poche occorrenze per addestrare il modello (vi è infatti molta sparsità nel dataset). Per questo motivo ho scelto di utilizzare il modello *BTM* (Biterm Topic Modelling) che considera le occorrenze di bigrammi come caratteristici del topic e spesso riporta risultati migliori per documenti brevi.

```
## 2021-07-05 16:08:41 Start Gibbs sampling iteration 1/2000
## 2021-07-05 16:08:47 Start Gibbs sampling iteration 101/2000
## 2021-07-05 16:08:54 Start Gibbs sampling iteration 201/2000
## 2021-07-05 16:09:00 Start Gibbs sampling iteration 301/2000
## 2021-07-05 16:09:06 Start Gibbs sampling iteration 401/2000
## 2021-07-05 16:09:13 Start Gibbs sampling iteration 501/2000
## 2021-07-05 16:09:19 Start Gibbs sampling iteration 601/2000
## 2021-07-05 16:09:26 Start Gibbs sampling iteration 701/2000
## 2021-07-05 16:09:34 Start Gibbs sampling iteration 801/2000
## 2021-07-05 16:09:41 Start Gibbs sampling iteration 901/2000
## 2021-07-05 16:09:48 Start Gibbs sampling iteration 1001/2000
```

```
## 2021-07-05 16:09:55 Start Gibbs sampling iteration 1101/2000
## 2021-07-05 16:10:02 Start Gibbs sampling iteration 1201/2000
## 2021-07-05 16:10:09 Start Gibbs sampling iteration 1301/2000
## 2021-07-05 16:10:16 Start Gibbs sampling iteration 1401/2000
## 2021-07-05 16:10:23 Start Gibbs sampling iteration 1501/2000
## 2021-07-05 16:10:30 Start Gibbs sampling iteration 1601/2000
## 2021-07-05 16:10:37 Start Gibbs sampling iteration 1701/2000
## 2021-07-05 16:10:45 Start Gibbs sampling iteration 1801/2000
## 2021-07-05 16:10:55 Start Gibbs sampling iteration 1901/2000
```

BTM Model

Topic Classification



Dalla rappresentazione notiamo diversi cluster di token divisibili in categorie di aspetti di Alexa:

- Sound quality
- Power cord
- Play music
- Smart light
- Voice command

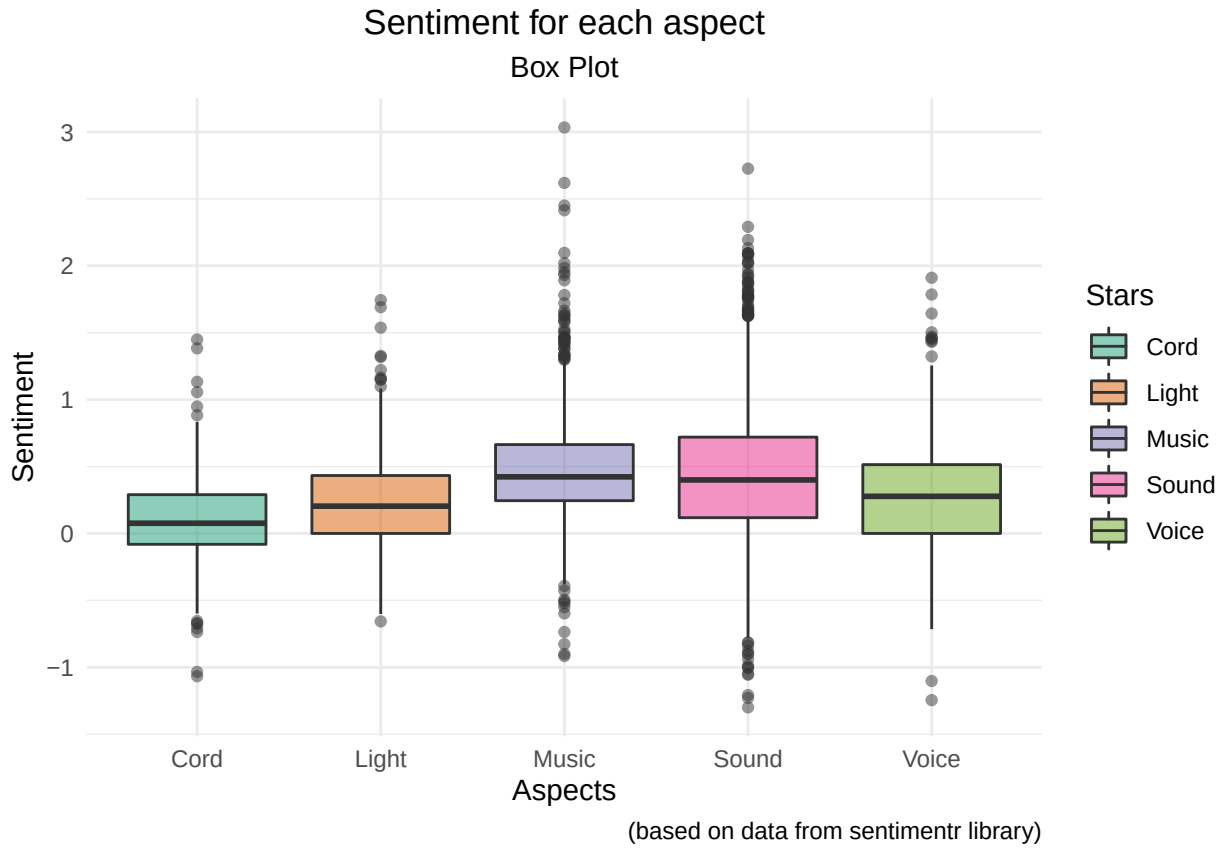
I restanti cluster non sono del tutto precisi come per esempio quello riguardante la generazione di Alexa (3rd vs 4th) oppure quello relativo alle recensioni scritte in lingua Spagnola e pubblicate comunque dagli Stati Uniti. Sulla base di questi aspetti è possibile andare a verificare effettivamente le parole "vicine" (per numero di occorrenze concomitanti) tramite il metodo *findAssocs()* di *tm*.

```
## $sound
## qualiti  better     gen    good   tinni    bass    much generat speaker   cheap
```

```
##    0.52    0.27    0.16    0.16    0.15    0.14    0.13    0.12    0.12    0.11
##  compar   great  hollow horribl    wors
##    0.11    0.11    0.10    0.10    0.10
```

Procedo dunque a verificare il sentiment delle frasi che contengono i 5 aspetti individuati per verificare l'opinione dei recensori. A tale scopo ho scritto una funzione che preso un dataframe e un vettore di parole da analizzare, ritorna un dataframe contenente il sentiment di ogni frase diviso in base agli elementi del vettore. Ho potuto così rappresentare parallelamente tutti i Boxplot di questi 5 aspetti confrontando le distribuzioni del sentiment.

```
aspect_analysis <- function(df, word=c("Sound","Cord","Music","Light","Voice")){
  word <- tolower(word)
  final_df <- NULL
  for (i in 1:length(word)){
    sound_vec <- df$text[as.vector(which(str_detect(df$text, word[i]) == TRUE))]
    sound_sentences <- sentimentr::get_sentences(sound_vec)
    sound_sentiment <- sentimentr::sentiment(sound_sentences)
    sound_sentiment <- as.data.frame(sound_sentiment)
    sound_sentiment["word"] <- word[i]
    final_df <- rbind(final_df, sound_sentiment)
  }
  return(final_df)
}
```



Da quest'ultima analisi è chiaro come l'aspetto più gradito di Alexa sia la possibilità di riprodurre contenuti musicali con un'ottima qualità del suono. Aspetto molto meno gradito è il *Power Cord* ovvero il cavo per

alimentare Alexa poiché molti recensori si sono lamentati del diverso colore di tale cavo rispetto all'Echo Dot. In molte recensioni infatti è presente questa forte critica alla diversità di colore tra i due oggetti che ha influenzato il sentiment generale delle precedenti analisi.