Università degli Studi di Milano Bicocca
Scuola di Scienze
Dipartimento di Informatica, Sistemistica e Comunicazione
Corso di laurea magistrale in Informatica

Progetto Data Analytics

Corsair Reviews Analysis on Amazon

829612 Giuseppe Magazzù
829685 Gaetano Magazzù
829889 Mirco Malanchini

A.A. 2022 - 2023

`https://github.com/saiteki-kai/amazon-reviews`

Indice

1 Introduzione

In questo progetto è stato analizzato uno dei dataset disponibili riguardante le recensioni di prodotti Amazon [4].

Al progetto è stato dato un taglio ben preciso, ovvero ci siamo messi nella condizione in cui un brand avesse contattato la nostra azienda per eseguire delle indagini in grado di rappresentare nel modo più fedele possibile l'esperienza del pubblico.

Per soddisfare la richiesta abbiamo cercato di rispondere alle seguenti domande in modo da delineare una linea guida per le analisi:

- Qual'è il gradimento del pubblico in merito al brand ?
- Sono presenti prodotti non apprezzati dal pubblico ?
- Quali sono gli aspetti che caratterizzano il brand e come sono percepiti dal pubblico ?
- Quali sono i principali competitors ?
- Come il brand è percepito rispetto ai competitors ?
- Un confronto degli aspetti che caratterizzano il brand rispetto ai competitors.

Il brand scelto è **Corsair** leader nella produzione di componenti per computer come schede RAM, memorie, alimentatori e sistemi di raffreddamento.

Nei successivi capitoli verranno descritte le scelte progettuali che hanno permesso l'analisi aspect-based sentiment dei dati in merito alla linea guida scelta.

Inoltre verranno mostrate le tecniche utilizzate per gestire le problematiche introdotte dal linguaggio naturale e le decisioni prese per poter caratterizzare in modo corretto le recensioni evitando di introdurre artefatti non riconducibili al brand, ma al servizio offerto da Amazon, i quali possono distorcere le conclusioni tratte dalle analisi.

2 Dataset

Il dataset [4] utilizzato in questo lavoro contiene i prodotti e le recensioni di Amazon dal 1999 fino al 2018 raggruppate in macro-categorie. Per esigenze computazionali il dataset è stato ridotto selezionando casualmente 30,000 recensioni dalla sotto-categoria *Internal Components*.

Nelle tabelle 1 e 2 vengono rispettivamente descritti tutti gli attributi dei prodotti e delle recensioni forniti. In questo lavoro sono stati analizzati e utilizzati solo i seguenti attributi: *asin*, *title*, *description*, *brand*, *price*, *categories*, *imageURLHighRes*, *reviewText*, *vote*, *overall*, *summary* e *unixReviewTime*.

Per semplicità gli attributi *reviewText*, *unixReviewTime* sono stati rinominati rispettivamente in *text*, *timestamp*.

asin	ID del prodotto
title	nome del prodotto
description	descrizione del prodotto
brand	marchio
date	data di disponibilità
price	prezzo in dollari (al tempo di crawl)
main_cat	categoria principale del prodotto
categories	lista delle categorie a cui appartiene
imageURL	URL delle immagini
imageURLHighRes	URL delle immagini ad alta risoluzione
rank	informazioni sul ranking di vendita
feature	lista di caratteristiche del prodotto
details	dettagli del prodotto
tech1	prima tabella dei dettagli tecnici del prodotto
tech2	seconda tabella dei dettagli tecnici del prodotto
also_view	prodotti anche visti
also_buy	prodotti anche acquistati
similar_item	prodotti simili

Tabella 1: Attributi dei prodotti

reviewerID	ID del recensore
asin	ID del prodotto
reviewerName	nome del recensore
verified	recensione con acquisto verificato
vote	#utenti che hanno trovato utile la recensione
style	metadati del prodotto
reviewText	testo della recensione
overall	valutazione del prodotto (da 1 a 5)
summary	breve descrizione della recensione
unixReviewTime	tempo della recensione (unix)
reviewTime	tempo della recensione (raw)
image	immagini del prodotto postate dall'utente

Tabella 2: Attributi delle recensioni

Per prima cosa è stata effettuata una pulizia dei dati per alcuni attributi rimuovendo eventuali parti di codice HTML/CSS.

Per l'attributo *price* sono stati rimossi il simbolo $ e le virgole che separano le migliaia. I valori con lunghezza uguale a 0 o superiore a 8 sono stati considerati non validi.

Dall'attributo *categories* sono state estratte le sotto-categorie di *Internal Components* e sono state rinominate rimuovendo la parola *Internal* che è già implicita nella sotto-categoria scelta in questo lavoro.

Per l'attributo *brand* inoltre è stata effettuata una normalizzazione tramite le seguenti operazioni, riducendo il numero di *brand* unici da 2842 a 828:

1. minuscolo
2. rimozione dei seguenti caratteri: ".", ",", "{", "}"
3. rimozione spazi in eccesso
4. rimozione delle seguenti parole: "by", "limited", "llc", "ltd", "inc", "co", "corp", "corporated", "corporation"
5. rinominazione dei brand composti da 1 solo carattere o da più di 7 parole in "unknown"

Successivamente sono stati identificati i valori mancanti e duplicati. Nella figura 1 sono state riportate le percentuali dei valori mancanti per ogni attributo. Gli unici attributi con valori mancanti sono *description*, *imageURLHighRes* e *price*. Questi attributi prima della pulizia presentavano parti di codice HTML/CSS in alcune recensioni, quindi si presume che molti di questi valori non siano stati estratti correttamente dal *crawler* e sono stati ignorati nel resto dell'analisi. Per quanto riguarda i duplicati sono state trovate e rimosse 67 recensioni con gli attributi *asin*, *summary*, *text* e *overall* uguali.

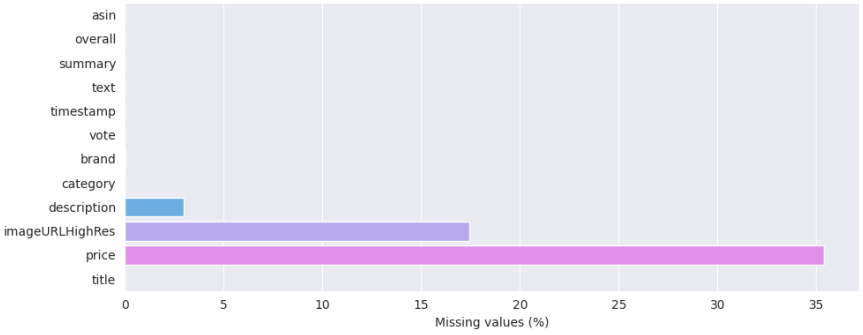

Figura 1: Percentuale dei valori mancanti per ogni attributo.

Tramite il modello FastText [2] sono state identificate e mantenute solo le recensioni in lingua inglese considerando l'attributo *text*, in cui sono stati rimossi gli URL e i tag HTML e infine è stato trasformato in minuscolo. Il numero di recensioni rimanenti è 28,983 che corrisponde al 96% del dataset.

2.1 Exploratory Data Analysis

Successivamente sono state effettuate diverse analisi esplorative del dataset pulito. Inizialmente sono state effettuate delle verifiche per controllare che non fossero presenti casi anomali, ad esempio verificare se un utente avesse fatto tutte le recensioni relative ad un prodotto.

Il dataset presenta un numero di 4,803 prodotti, e 25,374 recensori unici. Il numero medio di recensioni per prodotto è di 6, mentre il numero di recensioni medio per recensore è di 1.

Nella figura 2 viene mostrato il quantitativo di prodotti che possiedono un determinato numero di recensioni. Come si può notare la maggior parte dei prodotti ha al più 5 recensioni.

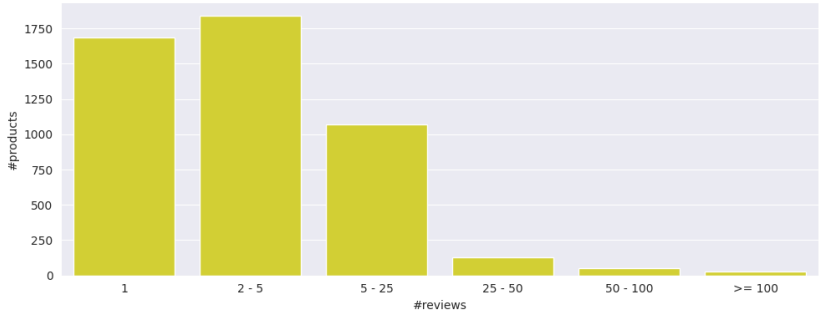

Figura 2: Numero di prodotti in base alla quantità di recensioni associate.

Successivamente è stato analizzato l'andamento delle recensioni nel tempo (figura 3), dove si può vedere un aumento considerevole negli ultimi anni probabilmente dovuto alla diffusione dell'utilizzo di Amazon; si trova conferma di questo anche nella figura 4 dove viene mostrato il numero di prodotti disponibili nel tempo.

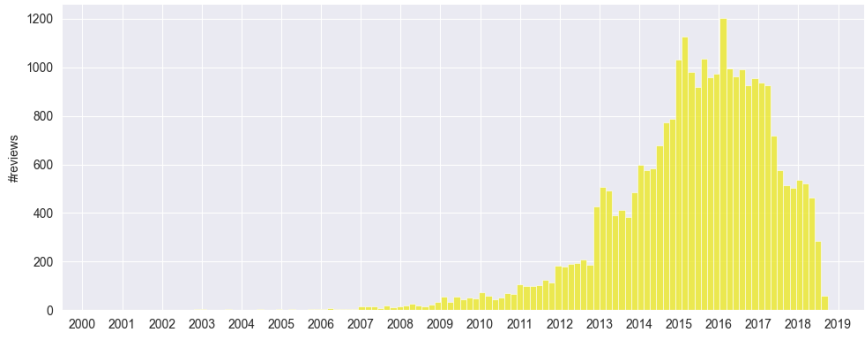

Figura 3: Numero di recensioni nel tempo.

Nelle figura 5 viene mostrato il numero di recensioni per i top-20 brand, mentre nella figura 6 viene mostrato il numero di recensioni per categoria del prodotto.

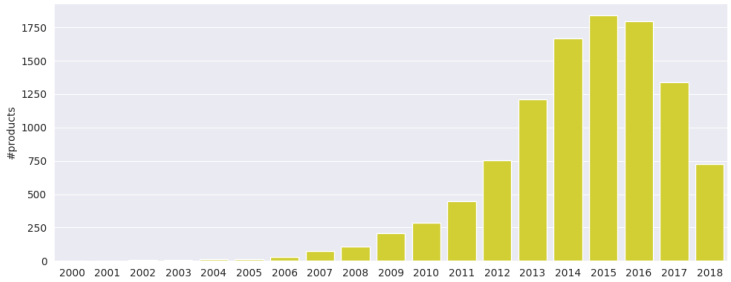

Figura 4: Numero di prodotti per anno.

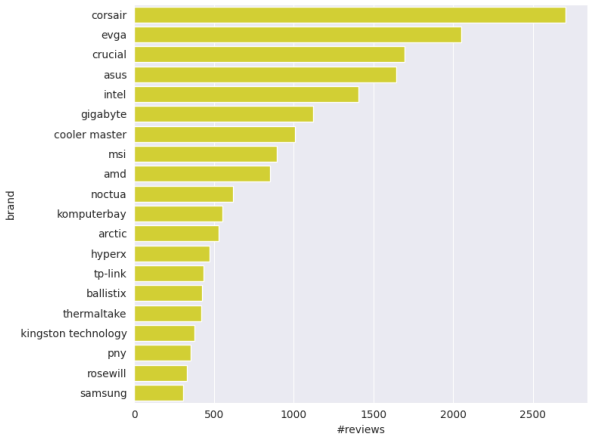

Figura 5: Distribuzione delle recensioni per brand.

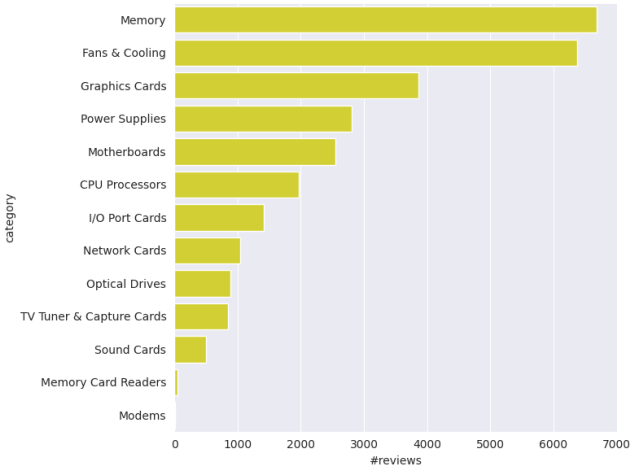

Figura 6: Distribuzione delle recensioni per categoria.

Una caratteristica importante da analizzare è il numero di stelle, che fornisce una prima e immediata valutazione del prodotto da parte dell'utente. Nella figura 7 viene mostrata la loro distribuzione in percentuale. Le recensioni presentano un numero di stelle prevalentemente positivo, infatti più dell'80% hanno una valutazione tra 4 e 5 stelle.

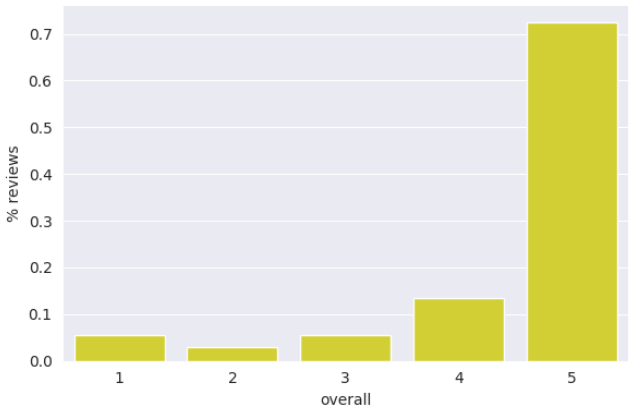

Figura 7: Distribuzione delle recensioni per numero di stelle.

Inoltre è possibile considerare il numero di voti assegnato a una recensione che indica, secondo gli utenti, quanto sia utile. Questo attributo può essere utilizzato per dare un peso alle recensioni, ma osservando la figura 8 si nota come gli utenti tendano a non votare le recensioni.

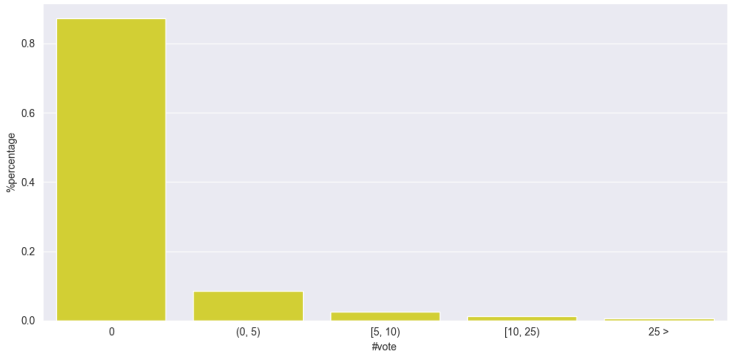

Figura 8: Percentuale di recensioni per numero di voti.

Infine è stata analizzata la parte testuale delle recensioni ovvero titolo e testo. Nelle figure 9 e 10 viene mostrata la percentuale delle recensioni in base alla lunghezza del titolo e del testo. Per quanto riguarda il titolo si può notare che più della metà delle recensioni hanno una lunghezza di 10-25 caratteri.

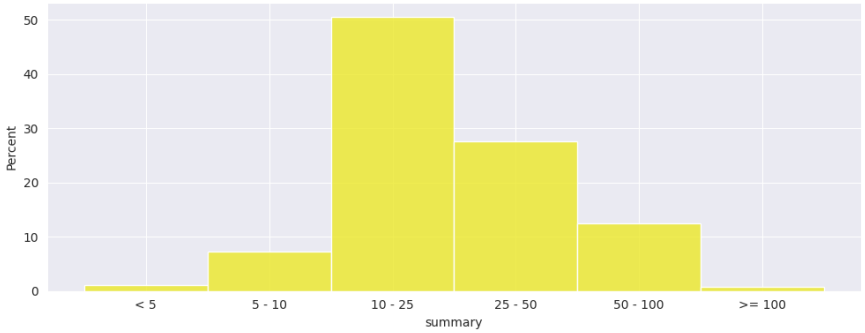

Figura 9: Percentuale di recensioni in base alla lunghezza del titolo.

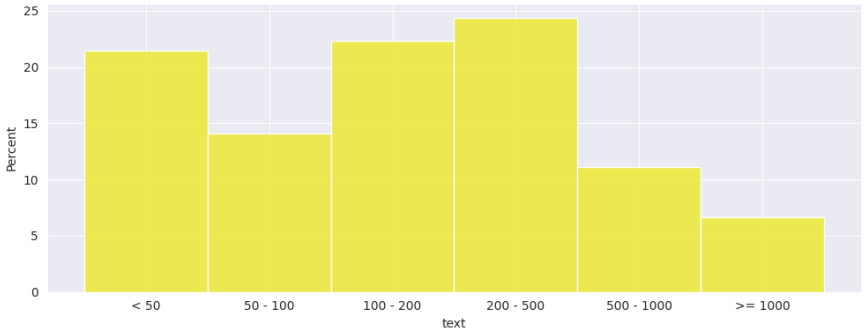

Figura 10: Percentuale di recensioni in base alla lunghezza del testo.

3 Pre-processing

Innanzitutto si è scelto di utilizzare il campo *text* e di escludere il campo *summary*, poiché come visto anche nel esplorazione dei dati questo campo prevede descrizioni troppo brevi e sommarie al fine di identificare degli aspetti.

Prima di passare il testo delle recensioni in input ai modelli vengono fatte le seguenti operazioni di preprocessing. Per prima cosa vengono rimossi gli URL tramite un'espressione regolare.

Poi vengono rimossi i tag HTML e sostituite le HTML entities (e.g. \amp; → &). Eventuali frasi o elenchi attaccati vengono separate aggiungendo uno spazio dopo punti e virgole. Successivamente i seguenti caratteri "+", "-", "_", "/", "\" vengono separati con uno spazio per dividere parole attaccate e vengono rimossi gli spazi in eccesso. Le contrazioni (e.g. i'm → I am) e gli slang (e.g. gotta → got to) vengono risolti.

Dopo queste operazioni il testo viene diviso in token mantenendo le frasi. In ogni frase viene identificato il POS associato ad ogni token che verrà usato per la lemmatization e vengono identificate le negazioni.

Per ogni token vengono rimossi i caratteri non alfanumerici e le ripetizioni (e.g. "greeeat" → "greeat", 'gooood' → "good") tramite espressioni regolari. I token vengono trasformati in minuscolo e poi viene applicato stemming o lemmatization.

Per le operazioni di preprocessing viene usata la libreria NLTK. Per lo stemming è stato usato *Snowball Stemmer*, per la lemmatization è stato usato *WordNet Lemmatizer*, come POS tagger è stato usato *Averaged Perceptron Tagger* e per la tokenization il *NLTKWordTokenizer*.

Le stopwords e i token con lunghezza pari a 1 o maggiore di 20 vengono rimossi. Le stopwords usate sono l'unione di quelle usate nelle librerie NLTK e SpaCy.

Ai token precedentemente identificati come negativi viene prefissato "not_". Come mostrato nella figura 11 vengono negati tutti i token che seguono un token identificato come negativo. Nell'esempio il risultato non è quello desiderato poichè "speed is pretty good" non dovrebbe essere negato. Per migliorare questo risultati si potrebbe cambiare il modo di dividere le frasi.

Text: This is a decent tablet for the price, i had no issues with this tablet, speed is pretty good [...]

Tokens: [decent tablet price] [not_issue not_tablet] [not_speed not_pretty not_good] [...]

Figura 11: Esempio della negazione dei token con lemmatization.

Dopo aver ottenuto i token è stata effettuata una rimozione in base alla *document frequency*. Sono stati rimossi i token presenti in più del 90% dei documenti e quelli presenti in meno di 4 documenti, mantenendo i token usati come seed (Tabella 3).

Nella figura 12 vengono mostrati i 25 token più frequenti al variare della normalizzazione: raw, lemmatization e stemming.

Normalization	Tokens before removal	Tokens in more than 90% documents	Tokens in less than 4 documents	Tokens after removal
Raw	30456	0 (0%)	19740 (65%)	10716
Lemmatization	25192	0 (0%)	16688 (66%)	8504
Stemming	20633	0 (0%)	13145 (64%)	7488

Tabella 3: Numero di token prima e dopo la rimozione in base alla *document frequency* al variare della normalizzazione.

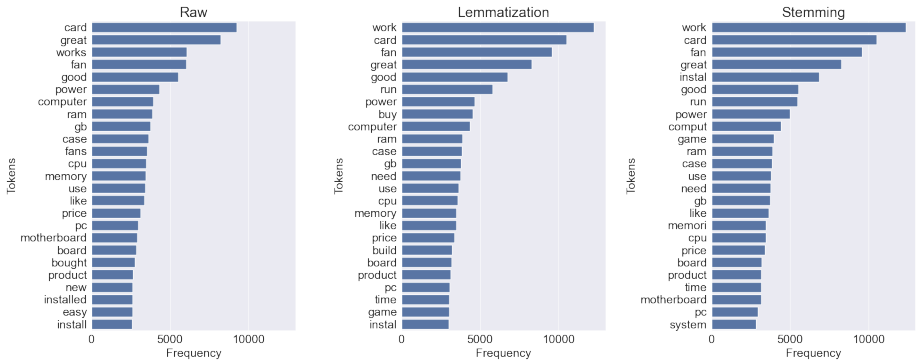

Figura 12: Top 25 token più frequenti al variare della normalizzazione.

Infine come effettuato anche nel lavoro originale di ASUM le frasi con numero di token superiore a 50 vengono ignorate.

4 Esperimenti

In questa sezione vengono riportati gli esperimenti effettuati per l'aspect-based sentiment analysis sui due modelli ASUM e JST. Per i due modelli sono state utilizzate le implementazioni degli autori[1][2].

I modelli sono stati valutati tramite il task di *sentiment classification*, poiché non è disponibile nessuna *groundtruth* per valutare i *senti-aspects*.

Come *groundtruth* per la *sentiment classification* sono state considerate le valutazioni in "stelle" date dai recensori. I valori 1 e 2 sono stati considerati avere *sentiment* negativi, mentre 4 e 5 positivi. I valori 3 sono stati esclusi dalla valutazione poiché il *sentiment* neutrale non è stato modellato.

Per ogni esperimento sono state effettuate 10 prove e sono stati mediati i risultati in termini di F1-macro, poiché il *sentiment* è sbilanciato.

Gli esperimenti che sono stati effettuati considerano diversi metodi di normalizzazione dei token (lemmatization, stemming, raw) e un diverso numero di topic (10, 20, 30, 50), mentre i parametri dei modelli (alpha, beta e gamma) sono stati fissati come nei lavori originali [1], [3].

Per limitazioni di risorse computazionali ci si è limitati a effettuare gli esperimenti sulla normalizzazione solo per ASUM, poiché nell'implementazione usata viene fornita la possibilità di eseguire in *multi-thread*, mentre per JST è stato considerato solo il caso con *stemming*.

4.1 ASUM

Come si può vedere nella figura 13 il metodo di normalizzazione migliore risulta essere *stemming* seguito da *lemmatization*, mentre il caso senza normalizzazione risulta quello peggiore. In generale all'aumentare del numero di topic si hanno valori di F1 più alti e una variabilità prove più bassa tra le 10.

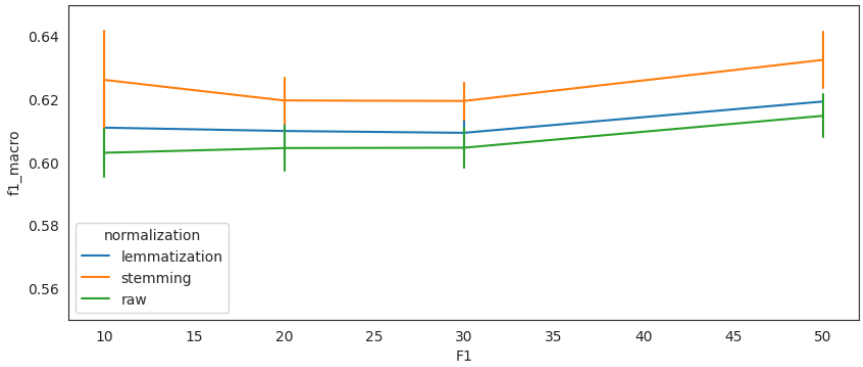

Figura 13: F1 al variare della normalizzazione e del numero di topic per ASUM.

[1] https://yohanjo.github.io/research/WSDM11/index.html
[2] https://github.com/linron84/JST

4.2 Comparazione ASUM e JST

Successivamente sono state effettuate le prove con JST al variare del numero di topic con solo stemming. Confrontando ASUM e JST figura (14) è possibile notare come a prescindere dal numero di topic considerato ASUM ottenga risultati superiori in termini di F1 rispetto a JST.

La prova con valore di F1 più alto viene scelta come modello finale ovvero ASUM considerando 50 topic effettuando stemming con 63.8% di F1.

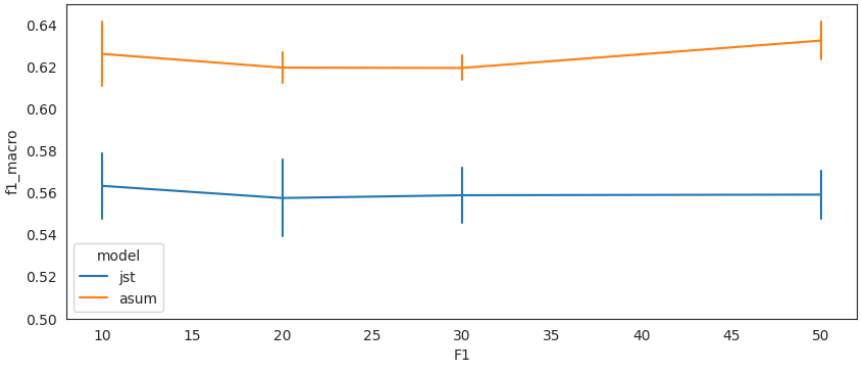

Figura 14: Comparazione di ASUM e JST sul task di sentiment classification al variare del numero di topic e usando stemming come normalizzazione.

4.3 Negazione dei token

Infine un'ulteriore prova è stata effettuata per cercare di migliorare i risultati di ASUM. Sotto assunzione che le congiunzioni avversative siano il punto in cui cambi l'opinione della frase, ogni frase è stata separata considerando le seguenti congiunzioni: "but, still, yet, while, however, nevertheless, whereas, notwithstanding, although e even though", in modo da migliorare la fase di negazione dei token.

Text: It's a very nice unit and great performance for the price point, however the fan noise was too loud for me.

Tokens (without split conjunctions): ['nice', 'unit', 'great', 'performance', 'price', 'point', 'fan', 'noise', 'loud']

Tokens (with split conjunctions): [['nice', 'unit', 'great', 'performance', 'price', 'point'], ['not_fan', 'not_noise', 'not_loud']]

Figura 15: Effetto della separazione di frasi utilizzando le congiunzioni sulla negazione dei token. A scopi visualizzativi in questo esempio viene utilizzata la lemmatization.

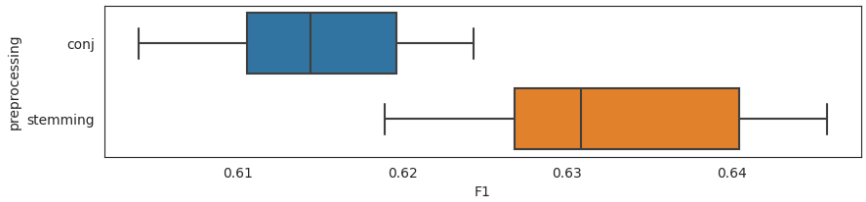

Figura 16: Comparazione del modello migliore con il metodo che sfrutta le congiunzioni avversative.

Considerando la figura 16 è possibile vedere che la modifica proposta addirittura comporta un peggioramento nella classificazione del sentiment, contrariamente da quanto aspettato, perciò tale modifica non verrà considerata per l'estrazione dei topic.

4.4 Identificazione Topics

Per ognuno dei 100 *senti-aspect* sono state assegnate delle etichette considerando le top-10 parole più probabili in ogni *senti-aspect*, successivamente verranno riportati alcuni esempi di topic con le relative etichette assegnate.

Alcuni *senti-aspect* che sono stati individuati corrispondono allo stesso concetto e hanno molte parole in comune, quindi sono state associate alla stessa etichetta, come ad esempio si può notare nelle figure 20 e 21.

Quindi questi *senti-aspect* sono stati aggregati in una partizione composta da 22 aspetti (figura 17).

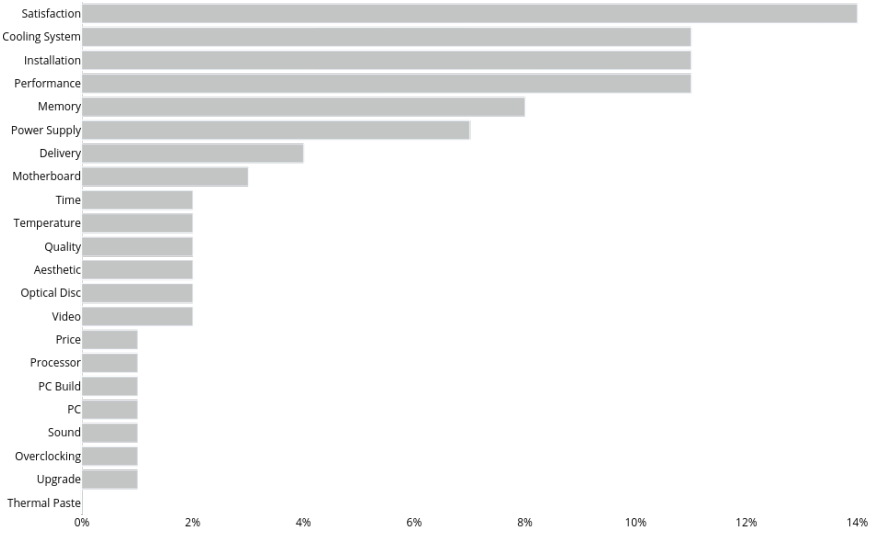

Figura 17: Distribuzione delle recensioni per gli aspetti identificati.

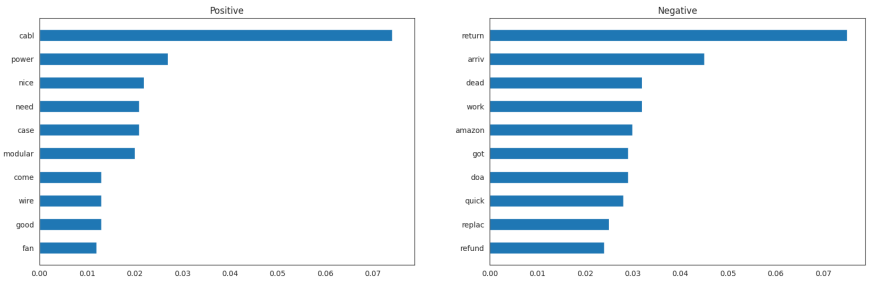

Figura 18: Power Supply

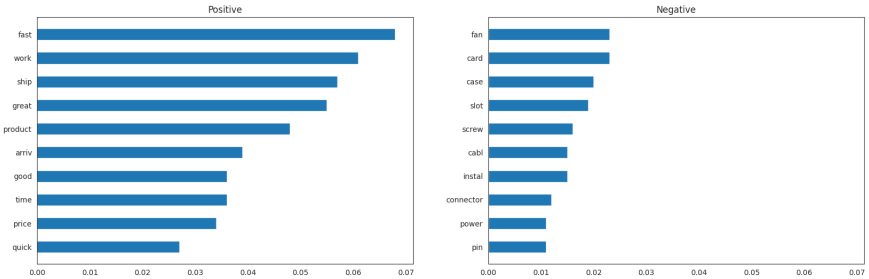

Figura 19: Delivery

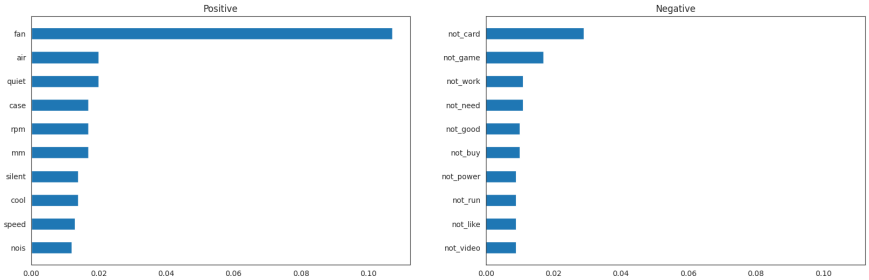

Figura 20: Cooling System

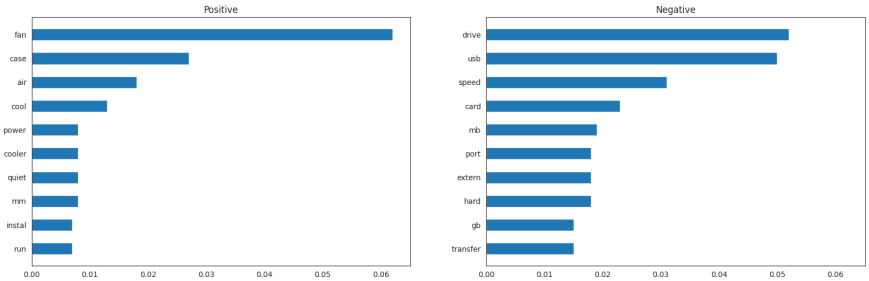

Figura 21: Cooling System

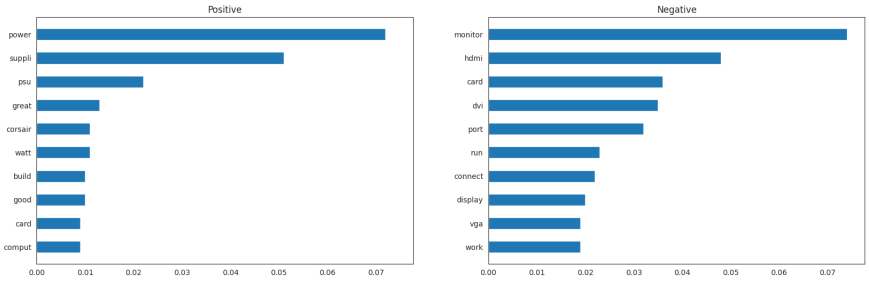

Figura 22: Power Supply

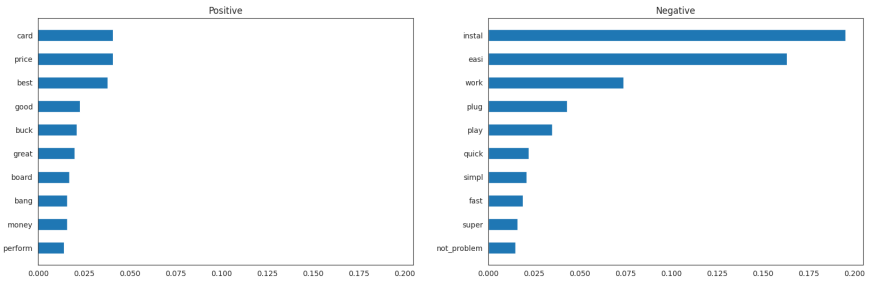

Figura 23: Price

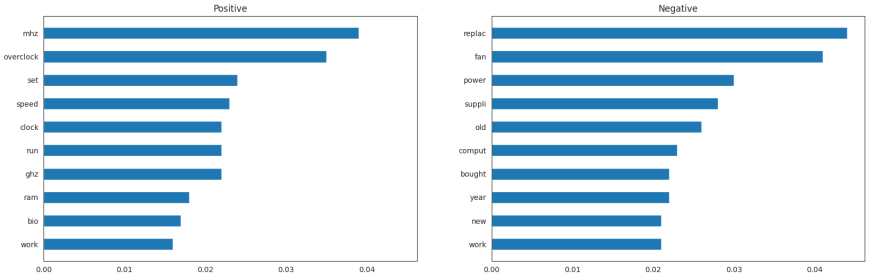

Figura 24: Overclocking

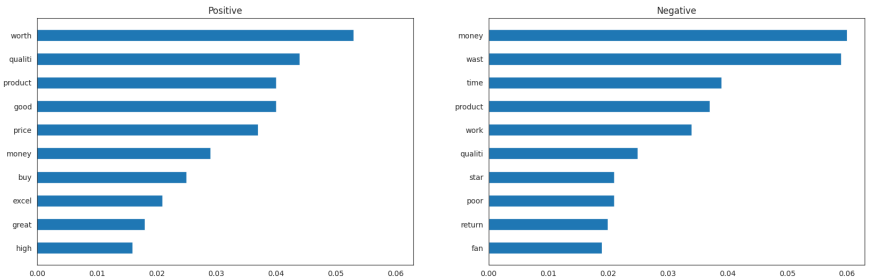

Figura 25: Quality

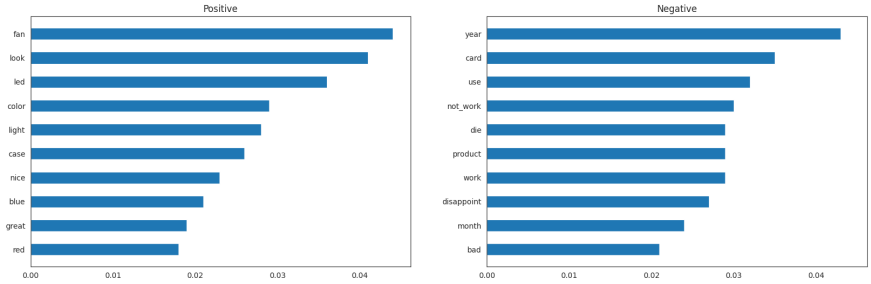

Figura 26: Aesthetic

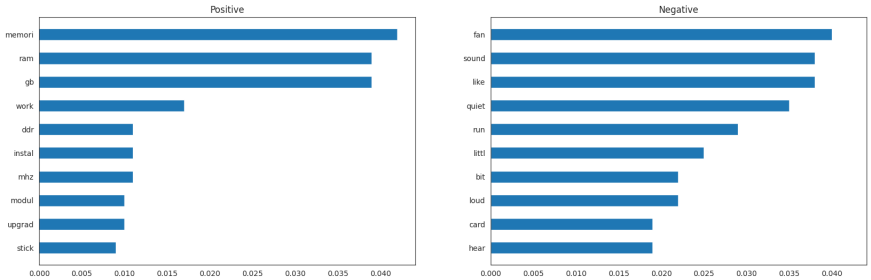

Figura 27: Memory

Per ogni recensione sono stati identificati i topic tramite una soglia T sulla distribuzione di probabilità document-aspect.

I *senti-aspect* che hanno una probabilità inferiore alla soglia T non verranno considerati. Inoltre le recensioni che presentano un'uguale probabilità per ogni topic fissato il sentiment non avranno assegnato alcun topic.

La scelta della soglia è stata effettuata manualmente in modo che abbia il valore più alto possibile tenendo topic con alta probabilità e che le recensioni abbiano un numero di topic maggiore possibile. Nella figura 28 viene comparato l'effetto di diverse soglie.

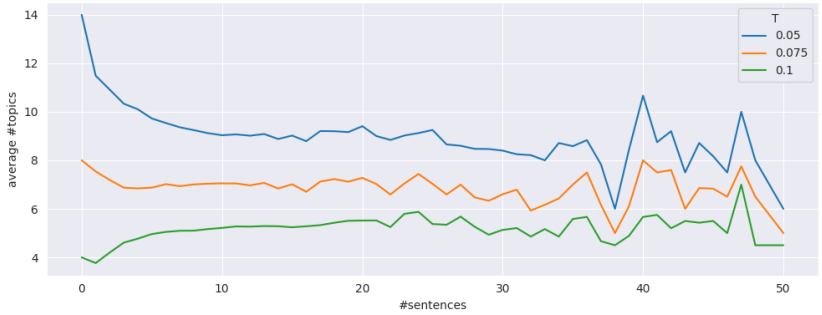

Figura 28: In questo grafico sono riportati il numero medio di topic assegnati a una recensione in base al numero di frasi al variare di diversi valori di soglia.

La soglia è stata scelta considerando che ASUM assume che a ogni frase venga associato un aspetto.

La soglia è stata scelta superiore a 0.04545, ovvero il caso in cui tutti i *senti-aspect* di una recensione hanno uguale probabilità, ed con il valore più alto possibile cercando di non scartarne troppi.

E' stata scelta una soglia di 0.075 che permette di avere un numero medio di circa 7 topic, considerando che il massimo numero di *senti-aspect* assegnabile a una recensione è 44 avendo individuato 22 topic tra positivi e negativi, ma che è molto improbabile che appaiano tutti i 22 topic anche in recensioni con 50 frasi.

5 Analisi

In questo capitolo verranno analizzati i dati raccolti per elaborare le informazioni che possono rispondere alla domande poste inizialmente (sezione 1).

Per poter fornire un'immagine del brand più realistica e aggiornata possibile sono stati considerati i dati dell'ultimo anno e sono stati confrontati con quelli dell'anno precedente.

Le informazioni più immediate sono il giudizio degli utenti espresso tramite le stelle; le stelle forniscono un indice approssimativo ma immediato di come il brand viene percepito.

La Corsair ha una votazione media di 4.5 stelle su un massimo di 5. Nella figura 29 è mostrata la distribuzione di stelle nelle recensioni.

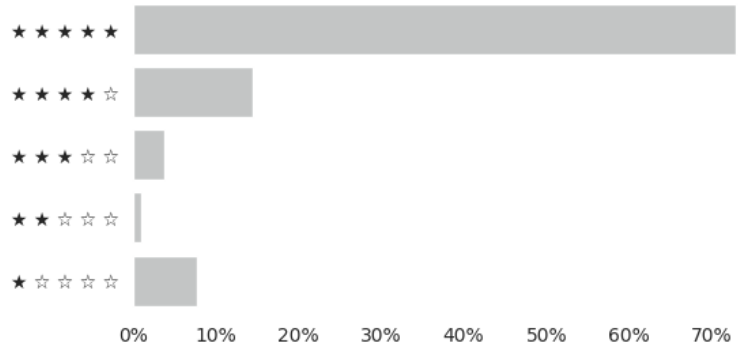

Figura 29: Distribuzione di stelle assegnate alle recensioni di Corsair.

Dalle recensioni è possibile capire quali sono i prodotti principali forniti da Corsair, come mostrato nelle figure 30 e 31. Le principali aree identificate associate al brand sono effettivamente quelle che sappiamo essere relative ai prodotti Corsair. Inoltre si può notare come appaiano parole e frasi che suggeriscono una valutazione del brand positiva, come "good", "great" e "i am very happy".

Figura 30: Questa wordcloud mostra le parole più ricorrenti all'interno delle recensioni dei prodotti del brand Corsair.

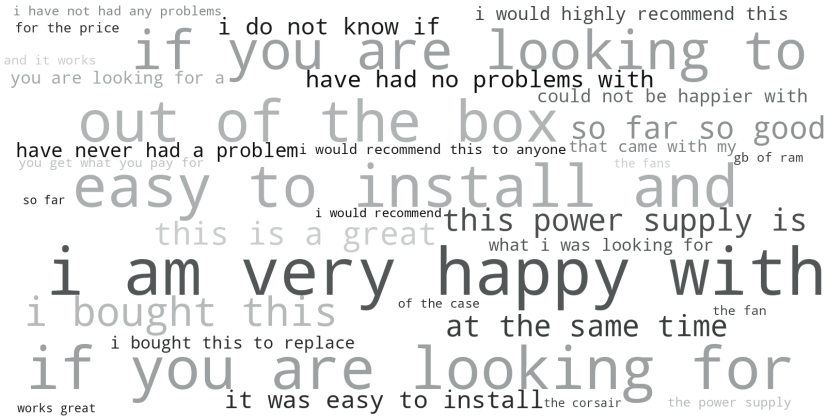

Figura 31: Questa wordcloud mostra le frasi più ricorrenti all'interno delle recensioni dei prodotti del brand Corsair.

Per avere informazioni più dettagliate che rispondano alle nostre domande è necessario usare analisi riguardanti gli aspetti e il *sentiment*, le quali permettono di sfruttare le informazioni presenti in ogni singola recensione.

Dall'analisi delle recensioni emerge che l'82.5% sono positive, mentre il 17.5% sono negative. Per andare a valutare in quali recensioni si ha un *sentiment* più negativo è utile analizzare come è distribuito il *sentiment* nelle categorie principali di Corsair.

Nella figura 32 viene mostrata la distribuzione delle recensioni per categoria per avere un'idea di quali categorie siano le più recensite dagli utenti. Nella figura 33 viene mostrato il *sentiment* per ogni categoria.

Si può notare come *Memory* e *Fans & Cooling* siano le categorie di punta del brand, mentre le categoria con *sentiment* più alto risultano *Power Supplies* e *Fans & Cooling*, seguite da *Memory* e per ultima *TV Tuner & Capture Cards* che però è quella con meno recensioni.

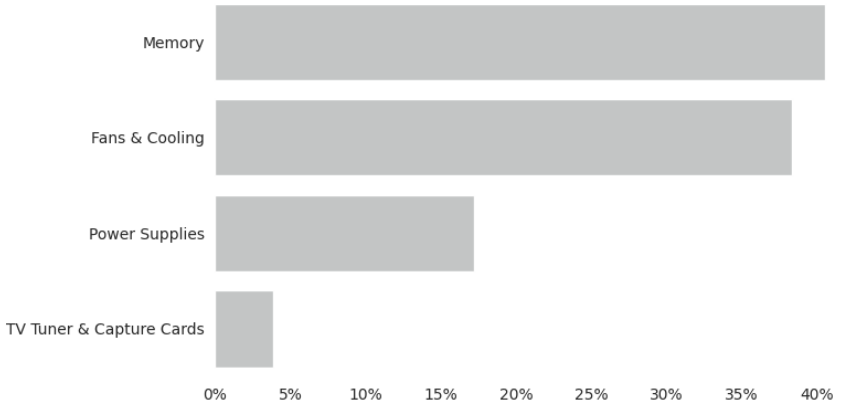

Figura 32: Principali categorie dei prodotti Corsair.

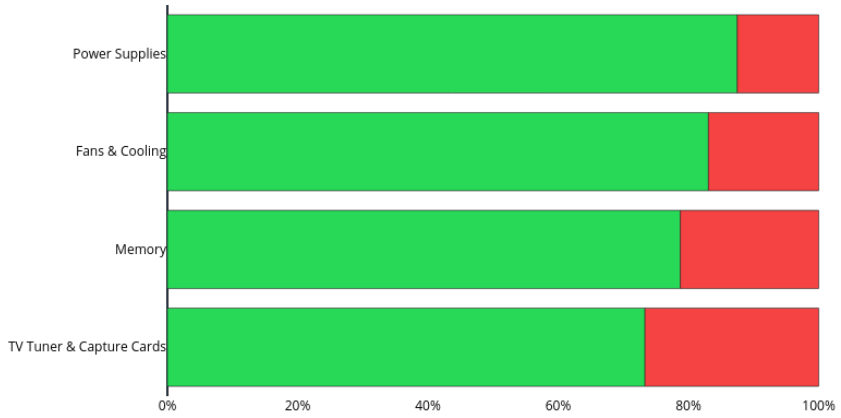

Figura 33: *Sentiment* delle principali categorie dei prodotti Corsair.

Un'ulteriore analisi che si può effettuare è quella degli aspetti più menzionati nelle recensioni (figura 34) e come per le categorie, andare a vedere la distribuzione del *sentiment* per ogni aspetto. Questo è utile per capire quali aspetti vengono associati al brand dagli utenti e di cosa si parla nelle recensioni.

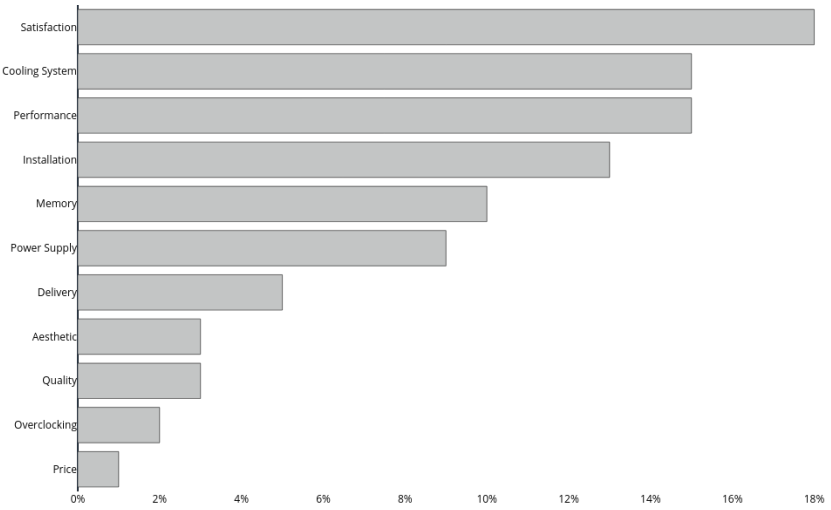

Figura 34: Aspetti più menzionati nelle recensioni.

Nella figura 35 si possono notare gli aspetti di forza del brand come *Aesthetic* e *Power Supply*; in corrispondenza si possono notare gli aspetti valutati in modo più negativo come *Price*, *Overclocking* e *Performance*.

L'aspetto *Price*, nonostante sia uno dei più negativi è menzionato in poche recensioni (figura 34), al contrario di *Performance* che determina quindi un fattore negativo per il brand.

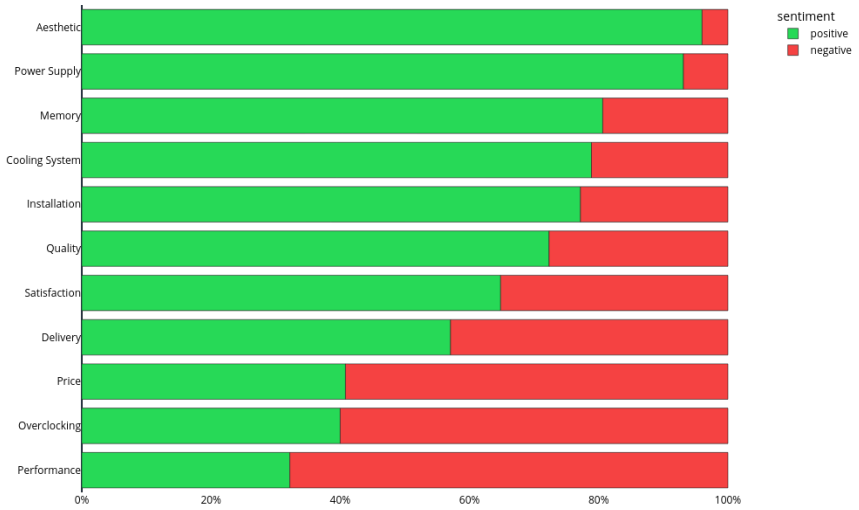

Figura 35: Distribuzione del *sentiment* degli aspetti più menzionati nelle recensioni.

Dopo aver effettuato queste analisi il passo successivo è stato individuare i principali competitor del brand Corsair in base alle categorie di prodotti più recensiti per poter effettuare un confronto. Nelle seguenti figure (36, 37, 38 e 39) viene mostrato il *sentiment* positivo associato a ciascun competitor per ogni categoria. Dalle figure emerge come tutti i competitor godano di una buona valutazione da parte degli utenti e inoltre le differenze tra i vari brand non sono significative.

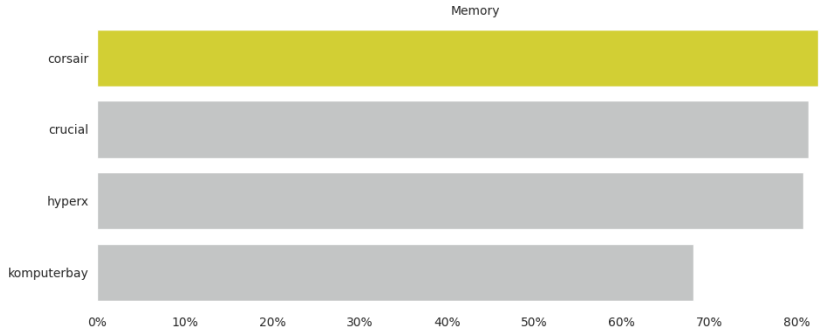

Figura 36: *Sentiment* positivo associato ai principali competitor del brand rispetto alla categoria *Memory*.

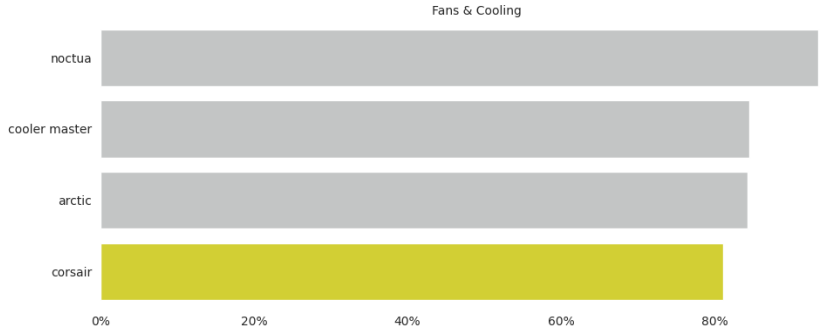

Figura 37: *Sentiment* positivo associato ai principali competitor del brand rispetto alla categoria *Fans & Cooling*.

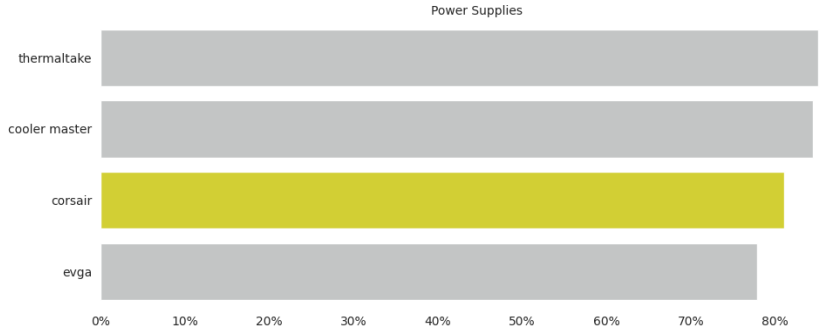

Figura 38: *Sentiment* positivo associato ai principali competitor del brand rispetto alla categoria *Power Supplies*.

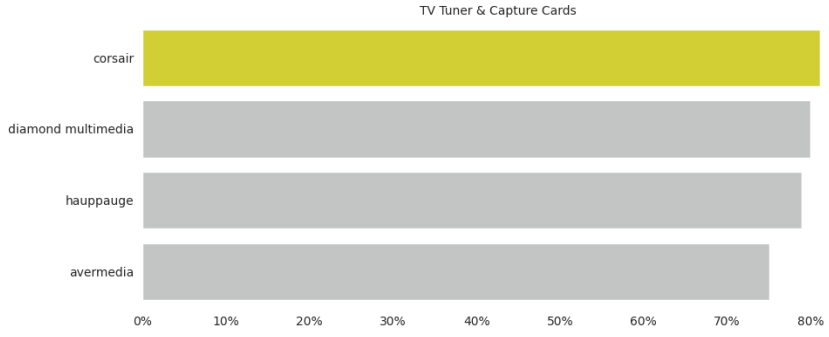

Figura 39: *Sentiment* positivo associato ai principali competitor del brand rispetto alla categoria *TV Tuner & Capture Cards*.

Nelle figure 40 e 41 è possibile notare un dettaglio ancora maggiore, in questi grafici viene rappresentato per ogni competitor e per le due categorie *Memory* e *Fans & Cooling* il *sentiment* positivo per ogni singolo topic. Questa analisi così dettagliata può permettere di capire quali aspetti possono essere causa di una migliore o peggiore valutazione da parte dell'utente e quindi capire su quali punti il brand dovrebbe focalizzarsi per poter migliorare i propri prodotti.

Nella figura 40 si può notare come Corsair e Crucial non solo abbiano una valutazione generale molto simile, ma anche la valutazione sul singolo topic è molto simile, a eccezione del topic *Performace*, che come già notato risulta avere un *sentiment* positivo basso.

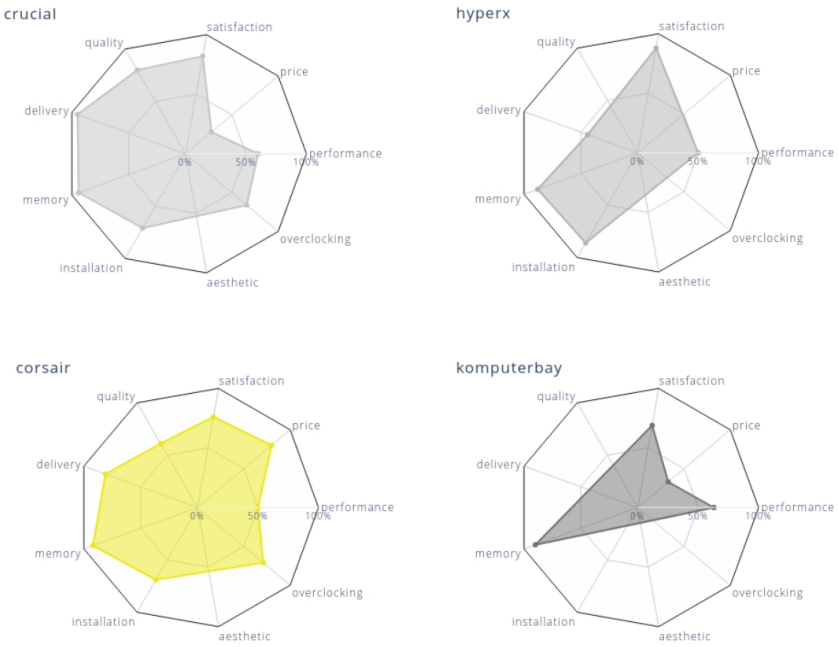

Figura 40: Confronto dei competitor del *sentiment* positivo per ogni singolo topic per la categoria *Memory*.

Nella figura 41 il valore di *Aesthetic* è molto elevato per Corsair mentre il valore di *Price* è il più basso fra i competitor, questo può indicare come una migliore estetica del prodotto sia apprezzata dagli utenti, ma pesi comunque di più un prezzo più elevato. Di conseguenza si ottiene un abbassamento della valutazione generale e quindi una minor competitività rispetto agli altri brand.

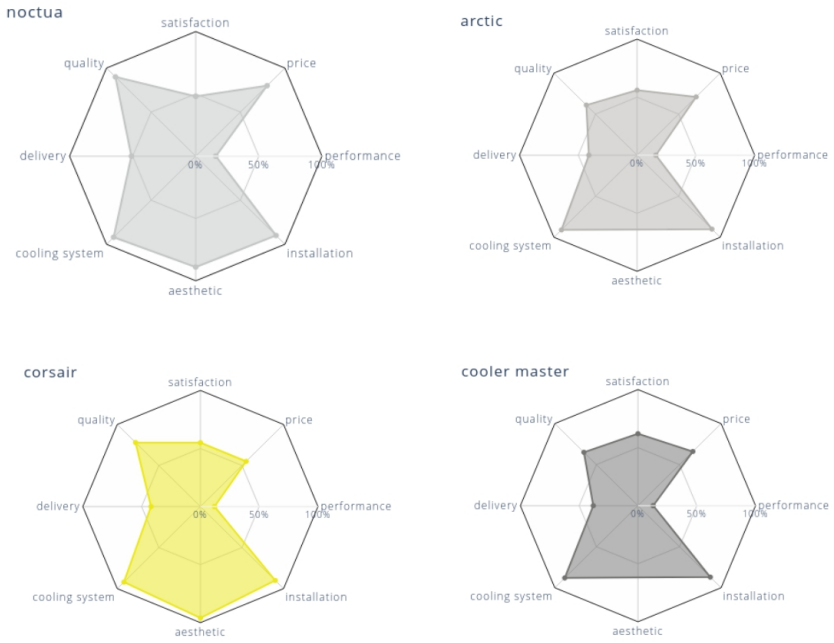

Figura 41: Confronto dei competitor del *sentiment* positivo per ogni singolo *topic* per la categoria *Fans & Cooling*.

6 Dashboard

In questa sezione vengono presentate le principali funzionalità della dashboard realizzata.

L'applicazione permette di visualizzare analisi per i brand più importanti del dataset specificando anche una determinata categoria, inoltre è possibile filtrare per uno specifico periodo di anni.

Nella pagina principale (figura 42) vengono mostrate le informazioni principali, lo storico di recensioni nel tempo, le valutazioni degli utenti, le categorie principali di un brand, i topic e le parole più frequenti.

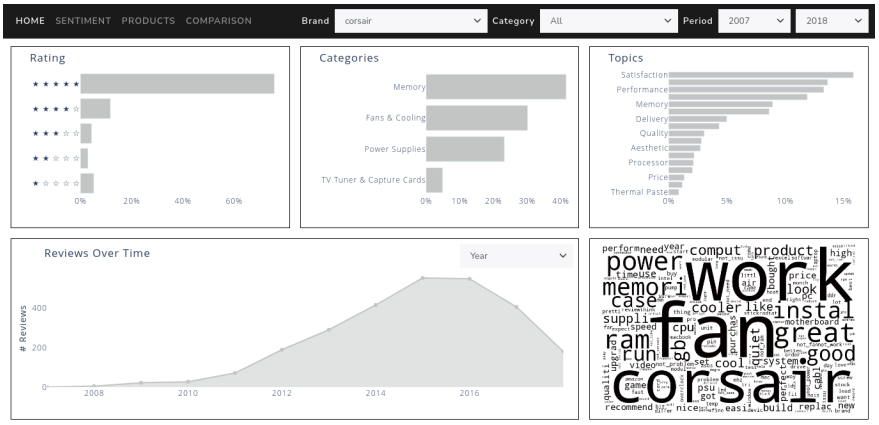

Figura 42: Screenshot che mostra la pagina principale della dashboard

Una pagina specifica dedicata all'analisi del sentiment (figura 43) è stata realizzata monitorando il sentiment generale nel tempo per categoria e per i topic.

Figura 43: Screenshot che mostra la pagina relativa all'analisi del sentiment della dashboard

Un'ulteriore pagina (figura 44) permette di mostrare i prodotti di un brand con le relative recensioni e con analisi sul sentiment e i topic.

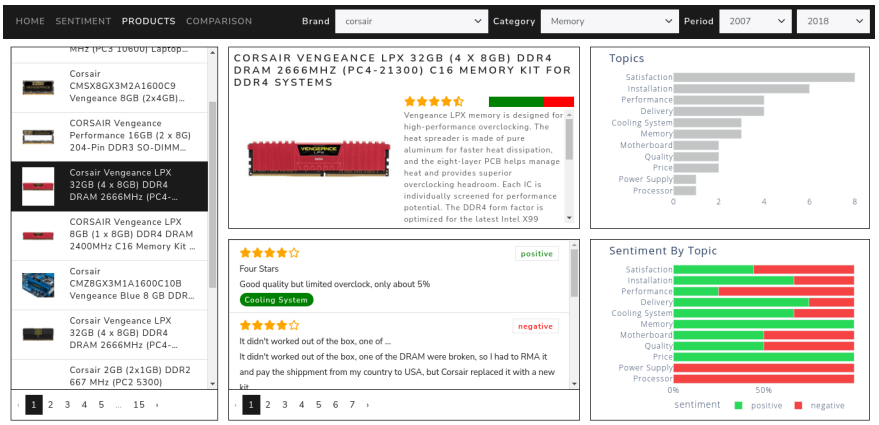

Figura 44: Screenshot che mostra la pagina relativa ai prodotti della dashboard

Infine scelto un brand è possibile compararlo con i competitor più importanti analizzando le differenze nel sentiment nel tempo e gli aspetti (figura 45).

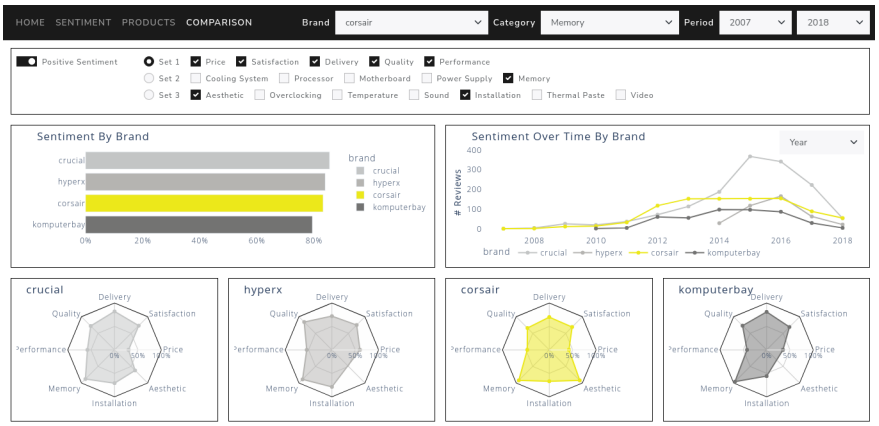

Figura 45: Screenshot che mostra la pagina relativa ai brand della dashboard

7 Conclusioni

Dopo aver eseguito un pre-processing sul testo, un'analisi iniziale sui dati e diversi esperimenti è stato possibile eseguire delle analisi aspect-based dei dati. Questo ci ha permesso di trovare risposta alle domande poste per poter compiere al meglio la richiesta posta dal brand nostro cliente, ovvero Corsair.

In sintesi quello che possiamo dire è che negli utenti si riscontra un alto gradimento verso il brand e i suoi prodotti, nei quali non è stato riscontrato nessun prodotto che danneggi l'immagine della Corsair.

I principali aspetti associati al brand e il relativo gradimento da parte del pubblico permettono di identificare i punti di forza e cosa più importante permette di identificare gli aspetti critici che richiedono più attenzione da parte di Corsair.

Grazie alle analisi è stato possibile identificare quali sono i principali competitor e analizzarne il *sentiment* rispetto al nostro brand. Sia rispetto ad una valutazione generale, sia rispetto a una valutazione dettaglia sul singolo topic.

In questo progetto due modelli ASUM e JST per l'aspect-based sentiment analysis sono stati addestrati e comparati con diverse prove effettuando il task di sentiment analysis. I risultati migliori sono stati ottenuti da ASUM considerando 50 topic ed effettuando lo *stemming*, con un valore di F1 del 63.8%.

Un'ulteriore prova è stata effettuata per cercare di migliorare la fase di negazione dei token, suddividendo le frasi dove presenti delle congiunzioni avversative, ma che non ha portato alcun miglioramento.

Sono molti i possibili miglioramenti tra cui diversi pre-processing, in particolare migliorare la suddivisione delle frasi permetterebbe una migliore negazione dei token aumentando le performance del riconoscimento del *sentiment*, inoltre si potrebbe considerare il *sentiment* neutrale, espandere i *seed* di partenza ed effettuare una valutazione dei topic estratti con una *groundtruth*.

Riferimenti bibliografici

[1] Yohan Jo e Alice H Oh. ≪Aspect and sentiment unification model for online review analysis≫. In: *Proceedings of the fourth ACM international conference on Web search and data mining*. 2011, pp. 815–824.

[2] Armand Joulin et al. ≪Bag of Tricks for Efficient Text Classification≫. In: *arXiv preprint arXiv:1607.01759* (2016).

[3] Chenghua Lin e Yulan He. ≪Joint sentiment/topic model for sentiment analysis≫. In: *Proceedings of the 18th ACM conference on Information and knowledge management*. 2009, pp. 375–384.

[4] Jianmo Ni, Jiacheng Li e Julian McAuley. ≪Justifying recommendations using distantly-labeled reviews and fine-grained aspects≫. In: *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*. 2019, pp. 188–197.