Data Analytics
Relazione Progetto

A.A. 2019-2020, Appello 06/07/2020

Componenti gruppo:

- Christian Bernasconi 816423
- Gabriele Ferrario 817518
- Riccardo Pozzi 807857
- Marco Ripamonti 806785

Link al repository: `https://gitlab.com/unizzan/dataanalyticsproject`

Indice

1 Introduzione

In questo elaborato vengono affrontati i temi dell'analisi di una rete di un sottoinsieme di prodotti presenti in Amazon e della sentiment analysis sulle relative recensioni.

Al capitolo 2 viene presentato il dataset con una breve descrizione delle tabelle e dei dati. Nel capitolo 3 è illustrata la parte di costruzione e di analisi della rete. Proseguendo al capitolo 4, si passa alla sentiment analysis delle recensioni. Infine, al capitolo 5 viene riportata una sintesi dei risultati raggiunti.

Gli strumenti utilizzati per svolgere le diverse parti del progetto sono i seguenti:

- esplorazione dataset: Python
- analisi rete prodotti: Python
- visualizzazione rete prodotti: Cytoscape
- sentiment analysis: Python + R

2 Esplorazione dataset

Il dataset oggetto di studio in questo progetto è relativo ad una partizione di prodotti di Amazon e dalle rispettive recensioni rilasciate dagli utenti. Esso è composto dalle due tabelle *product* e *reviews* descritte nel seguito sezioni 2.1 e 2.2. I testi presenti in entrambe le tabelle sono in lingua italiana.

2.1 Prodotti

La tabella *product* è quella su cui si baserà l'analisi di rete e si compone dei seguenti campi:

- **_id**: codice alfanumerico identificativo del prodotto
- **title**: titolo del prodotto; rappresenta già di per sé una breve descrizione delle sue caratteristiche
- **category**: categoria di appartenenza del prodotto
- **price**: prezzo del prodotto in euro
- **avg_rating**: media delle stelle delle recensioni sul prodotto; valore nell'intervallo [1.0, 5.0]
- **reviews_number**: numero totale di recensioni ricevute dal prodotto
- **questions_number**: numero totale di domande poste dagli utenti sul prodotto
- **pictures**: url delle immagini relative al prodotto
- **description**: descrizione dettagliata del prodotto
- **features**: caratteristiche del prodotto
- **versions**: lista degli identificativi dei prodotti che rappresentano altre versioni dello stesso
- **bought_together**: lista degli identificativi dei prodotti che vengono più spesso acquistati con il prodotto in questione nello stesso ordine; rappresenta una relazione tra prodotti
- **also_bought**: lista degli identificativi dei prodotti che vengono più spesso acquistati in combinazione con il prodotto in questione; non necessariamente l'acquisto avviene nello stesso ordine; rappresenta una relazione tra prodotti

- **also_viewed**: lista degli identificativi dei prodotti che vengono spesso più visualizzati da chi visualizza il prodotto in questione; rappresenta una relazione tra prodotti

Questa tabella raccoglie un totale di 20,459 prodotti, i quali si dividono in 37 categorie. La distribuzione nelle categorie è mostrata in figura 1 e si può notare come essa sia particolarmente sbilanciata.
Le relazioni *bought_together*, *also_bought* e *also_viewed* che legano i prodotti sono fondamentali per effettuare un'analisi di rete e saranno approfondite nella sezione 3. Riguardo a queste relazioni è bene far notare che molti degli identificativi fanno riferimento a prodotti ignoti in quanto non compaiono nel dataset disponibile.

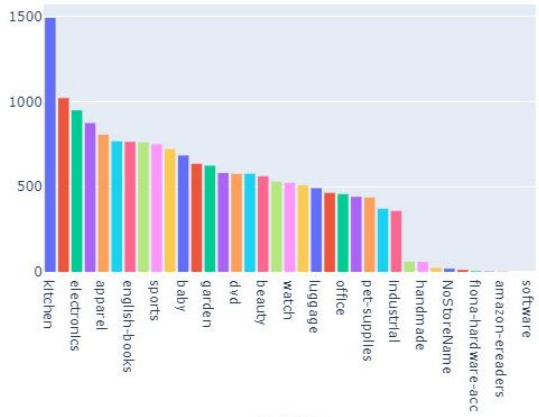

Figura 1: Distribuzione dei prodotti tra le categorie

2.2 Recensioni

La tabella *reviews*, che sarà oggetto di una sentiment analysis, si compone dei seguenti campi:

- **_id**: codice alfanumerico identificativo della recensione
- **product**: codice alfanumerico identificativo del prodotto a cui si riferisce la recensione
- **title**: titolo della recensione
- **author-id**: codice alfanumerico identificativo dell'autore della recensione
- **author-name**: nome dell'autore della recensione

- **date**: data in cui è stata scritta la recensione
- **rating**: voto in stelle dato al prodotto in questione; valore intero nell'intervallo [1, 5]
- **helpful**: numero di volte in cui la recensione è stata segnalata come utile da altri utenti
- **verified**: flag che rappresenta se la recensione sia relativa ad un acquisto verificato
- **body**: testo della recensione

La tabella raccoglie un totale di 1,988,854 recensioni. In figura 2 è possibile osservare una distribuzione fortemente sbilanciata verso le valutazioni positive, per questo motivo sarà importante bilanciare i dati nella fase di sentiment analysis. Come verrà mostrato alla sezione 4, gli attributi *helpful* e *verified* assumeranno un ruolo importante nel filtraggio del numero di recensioni. L'attributo *body* costituirà invece il testo oggetto dell'analisi.
Nelle figure 3 e 4 sono riportati rispettivamente i termini più frequenti per le recensioni positive ($rating > 3$) e negative ($rating < 3$).

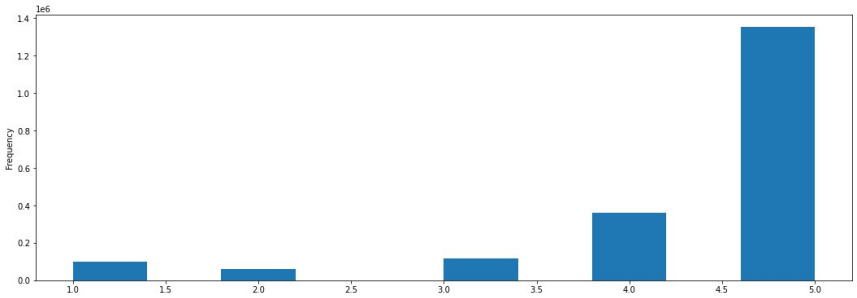

Figura 2: Distribuzione del numero di stelle delle recensioni

Figura 3: Word cloud delle parole che compaiono nelle recensioni positive

Figura 4: Word cloud delle parole che compaiono nelle recensioni negative

3 Rete prodotti

L'analisi condotta sulla rete ha lo scopo di individuare gruppi di prodotti che interagiscono tra loro e di andare ad identificare quelli più rilevanti all'interno di essa tramite un'analisi di centralità. Nel seguito vengono descritte le scelte di costruzione della rete, definito il concetto di prodotto rilevante ed infine presentate le analisi che partendo dalla globalità della rete scendono nei dettagli delle sottoreti più significative.

3.1 Costruzione rete

Inizialmente sono state generate 3 reti diverse tenendo conto delle singole relazioni *bought_together*, *also_bought* e *also_viewed*. Da uno studio generale è emerso che la relazione *also_viewed* risulti poco significativa per gli obiettivi preposti, per cui è stata scartata in quanto collega prodotti prettamente simili tra loro. Le relazioni *bought_together* e *also_bought*, invece, permettono di portare alla luce prodotti diversi il cui impiego effettivo può dipendere l'uno dall'altro e ciò permette di stabilire se esistono prodotti particolarmente strategici all'interno della rete. Si è inoltre notato che le due relazioni in questione riportano spesso lo stesso prodotto sia in *bought_together* che in *also_bought* e che inoltre la direzionalità della relazione *also_bought* sia indipendente dall'ordine temporale di acquisto.
Al netto di queste motivazioni, è state considerato ragionevole effettuare un'unione delle due relazioni in un'unica relazione di acquisto adirezionale.

Una volta stabilito il nuovo insieme di relazioni è stata necessaria una fase di pulizia per andare a scartare gli archi e i nodi superflui per la costruzione della rete. Per prima cosa è stato necessario rimuovere tutti gli archi di cui non sono presenti le informazioni relative al nodo destinazione. Dopodiché, sono state unite tutte le relazioni simmetriche in quanto la direzionalità delle relazioni non è utile ai fini di questa analisi. Infine, sono stati rimossi dalla tabella dei prodotti tutti i nodi rimasti isolati dopo il filtraggio delle relazioni. Al termine del processo di pulizia si è passati da 82628 a 26572 per il numero di archi e da 20459 a 17914 per il numero di nodi.
A partire dal nuovo insieme di archi è stata quindi costruita la rete dei prodotti tramite un grafo non orientato.

3.2 Concetto di prodotto rilevante

In questa analisi con il termine **prodotto rilevante** si intende quel prodotto che, se pubblicizzato in base a quanto emerge dalla rete, può costituire delle vendite

facili verso chi acquista altri prodotti correlati ad esso.
I prodotti strategici vengono individuati tramite il calcolo delle centralità dei nodi. In particolare, è la *degree centrality* quella che permette di individuare i prodotti che hanno maggiori interazione con altri. Trovare un prodotto con centralità di grado alta significa trovare un prodotto con un ampio vicinato. Per questo motivo se si dovesse scegliere quali prodotti pubblicizzare a partire da uno specifico nodo, una scelta strategica potrebbe essere quella di dare priorità ai prodotti del vicinato con grado più alto. Questa misura di centralità sarà particolarmente utile per stabilire i prodotti più importanti all'interno delle sottoreti relative a categorie e communities.

Un'ulteriore misura di centralità che invece può essere utile per un'analisi sulla rete nella sua globalità è quella della *closeness centrality*. Tramite essa è possibile individuare i prodotti centrali meglio posizionati per influenzare la rete più velocemente. Per esempio aggiungendo un annuncio pubblicitario sulla pagina del prodotto, esso sarà più velocemente raggiungibile dagli altri prodotti della rete seguendo una catena di acquisti.

3.3 Analisi generale

La rete che è stata costruita conta un totale di 962 componenti connesse, tra cui una giant component (figura 5) costituita da 11,937 nodi e 18,991 archi.
In figura 6 è mostrata la distribuzione dei gradi. Computando l'average degree (2.966) si osserva che mediamente un prodotto tende ad essere connesso con altri 3 prodotti. Il coefficiente di clustering medio dell'intera rete è pari a 0.286, quindi mediamente per un nodo si ha che due prodotti vicini siano connessi con il 30% di probabilità. La densità della rete ha un valore molto basso, pari a 0.000165 .

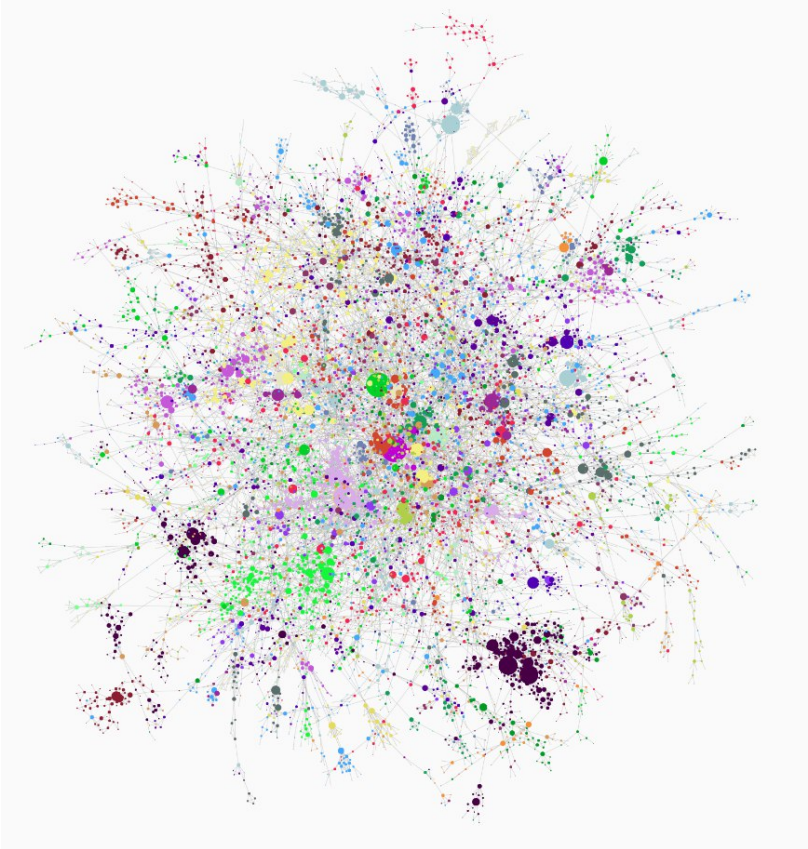

Figura 5: Giant component colorata per categorie e con dimensione dei nodi data dalla degree centrality

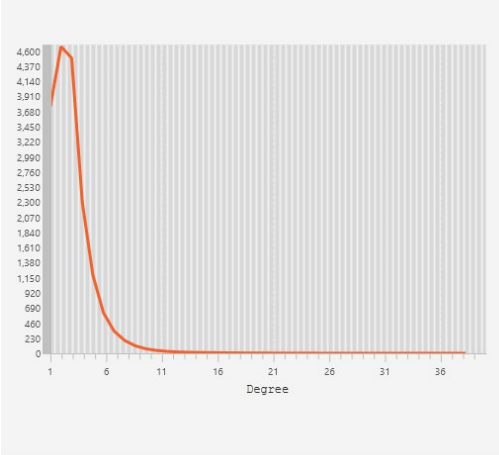

Figura 6: Distribuzione dei gradi nella rete completa

Calcolando il coefficiente di assortatività sull'intera rete si ottiene un valore di -0.0194 che, essendo negativo, permette di asserire che la rete sia di tipo disassortativo. È infatti possibile osservare nel grafico in figura 8 come i nodi con grado più alto, seppur in modo non nettamente marcato, tendano a connettersi a nodi con grado basso. A conferma di ciò, in figura 7 si può osservare come i nodi con più interazioni[1] non tendano a connettersi tra loro. È possibile quindi supporre che nella rete sono più frequenti gli acquisti di prodotti con centralità di grado alta in combinazione con prodotti di bassa centralità. Data la natura della rete, la rimozione di un prodotto hub potrebbe quindi portare ad isolare i prodotti con grado basso.

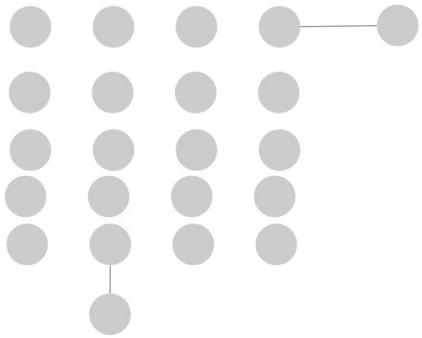

Figura 7: Nodi hubs con grado almeno pari a 20 isolati dal resto della rete

[1]sono stati considerati come nodi gli hubs di grado almeno pari a 20

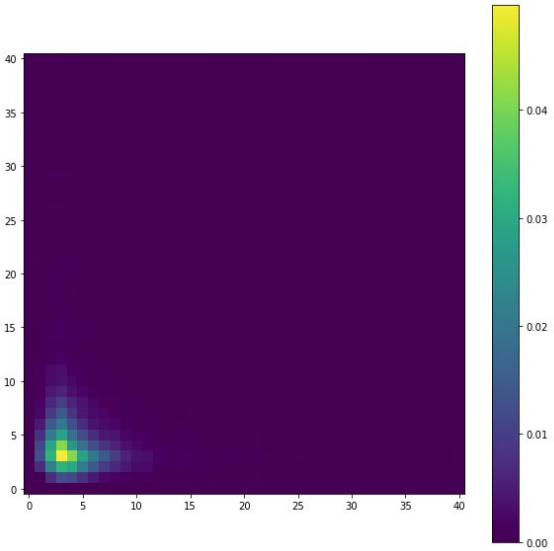

Figura 8: Grafico di degree correlation

In tabella 1 sono riportati i prodotti centrali della rete in termini di grado, ossia quelli definiti precedentemente come prodotti rilevanti.
In tabella 2 sono invece riportati i prodotti centrali della rete in termini di closeness.

id	**title**	**degree**	**closeness**
B0792HCFTG	Echo Dot (3 generazione)	0.002233	0.055939
B07P24R81S	Gfone 8 Set di strumenti per manicure per unghie da 8 pezzi	0.002121	0.057762
B07DMJPV31	FIFA 19 - PlayStation 4	0.002065	0.052239
B07PYGPZZF	RFID Blocking Porta Carte di Credito	0.001730	0.060146
B00UFGMVFG	Bruzzzler 200100001066 Ciminiera con impugnatura di sicurezza	0.001618	0.039824

Tabella 1: Tabella top 5 prodotti per degree centrality

id	title	degree	closeness
B07PYGPZZF	RFID Blocking Porta Carte di Credito	0.001730	0.060146
B07PBFJTZG	Willful Cuffie Bluetooth 5.0	0.001116	0.057851
B07P24R81S	Gfone 8 Set di strumenti per manicure per unghie da 8 pezzi	0.002121	0.057762
B07NQ7JJKJ	Feob Orologio Fitness Tracker Pressione Sanguigna Cardiofrequenzimetro da Polso	0.000223	0.057703
B07GFT2NS8	Rolsac - Cagliplast Rotolo con Scatola Esterna	0.001451	0.057644

Tabella 2: Tabella top 5 prodotti per closeness centrality

3.4 Analisi categorie

Dopo aver osservato la rete nella sua globalità si è deciso di osservare le interazioni tra i prodotti dividendo la rete rispetto alle categorie. I risultati mostrati sono quelli relativi alle categorie più grandi in quanto, avendo a disposizione una porzione non bilanciata del dataset dei prodotti di Amazon, sono le più significative.

In tabella 3 sono disponibili le statistiche estratte per le prime 5 categorie più popolate. Per ogni categoria viene indicato il totale dei prodotti da cui è composta e il prodotto più centrale rispetto alla degree centrality con il rispettivo grado.

category	cardinality	id max degree product	max degree	title max degree product
kitchen	1784	B00M94FVO0	12	Vileda Asse da Stiro Smart
electronics	1143	B07MJKSZFM	8	iBetter per Xiaomi Redmi Note 7 Cover
tools	1125	B0076VLVXG	10	BLACK+DECKER A7188-XJ Set per Forare ed Avvitare
books	993	8806240986	12	La versione di Fenoglio
apparel	946	B00I0DF9I2	10	Navigare 512 Maglietta intima Uomo

Tabella 3: Tabella statistiche delle top 5 categorie

Prendendo per esempio il nodo centrale della categoria *kitchen*, esso corrisponde al prodotto di un asse da stiro. Verificando i prodotti del suo vicinato, ossia quelli comprati spesso in combinazione con lo stesso, si osserva che sono per lo più ferri da stiro. Questo è un chiaro esempio di come il prodotto centrale in termini di grado sia in realtà un prodotto accessorio rispetto al suo vicinato. Essendo quindi un prodotto comune per diversi prodotti principali può essere considerato un prodotto strategico data la sua più facile vendibilità.
Un comportamento analogo è stato riscontrato anche per molti altri prodotti tra quelli con degree centrality maggiore.

Osservando come si dispongono nella rete le diverse categorie, si può notare che una stessa categoria vada a formare gruppi distinguibili di prodotti. In figura 9 è riportato l'esempio per la categoria *musical-instruments*. Per questo motivo, come si vedrà nella sezione successiva, è stato effettuato in seguito un processo di community detection sulla rete per individuare i reali gruppi di prodotti che si formano in base alle relazioni di acquisto.

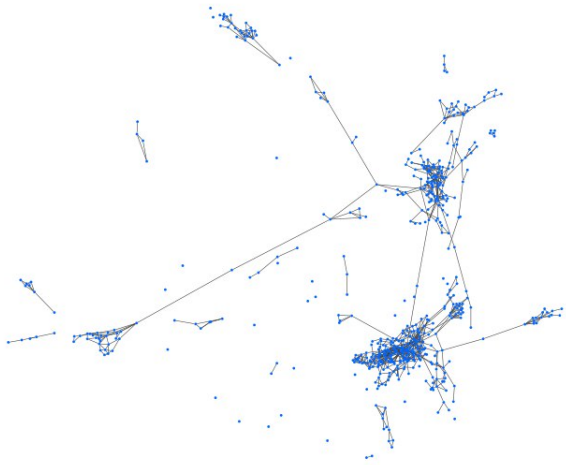

Figura 9: Sottorete dei prodotti della categoria musical-instruments

3.5 Analisi communities

Le communities della rete sono state individuate mediante l'algoritmo Clauset-Newman-Moore greedy modularity maximization. Tramite questo algoritmo sono state trovate 1064 communities. Il numero elevato era prevedibile data la grande quantità pari a 962 delle componenti connesse. Si è osservato, inoltre, che le communities più rilevanti fanno tutte parte della giant component, la quale ne contiene

103.

Come per le categorie, i risultati mostrati sono quelli relativi alle communities più numerose.

In tabella 4 sono disponibili le statistiche estratte per le prime 10 communities. Oltre a fornire la cardinalità e gli elementi centrali come per le categorie, per ogni community è stato calcolato il coefficiente di clustering medio per determinare il livello di coesione. Inoltre ci si è posti l'obiettivo di identificare la sfera di interesse di ogni gruppo di prodotti e di darne una sorta di etichetta che possa riassumerne il contenuto. A tale scopo sono stati calcolati i seguenti campi:

- **Dominant Category**: categoria dominante della community; estratta osservando con quale frequenza le categorie compaiono tra i prodotti della community tenendo in considerazione anche lo z-score delle frequenze.
- **Categories Distribution**: distribuzione di frequenze relative delle categorie principali presenti nella community.
- **Top Words**: parole più frequenti presenti nei titoli dei prodotti di una community; i titoli sono stati divisi in token, ogni token è stato reso lowercase e filtrato eliminando *stop-words*, punteggiatura e parole poco rappresentative della community come colori e numeri.
- **Top Entities**: entità più frequenti estratte dai titoli dei prodotti di una community tramite entity recognition e in seguito filtrate in modo da eliminare quelle poco rilevanti come i colori.

id	dominant category	cardinality	max degree	max degree product title	avg clust. coeff.	top words	top entitites
1	videogames	520	36	FIFA 19 - PlayStation	0.19467	playstation, nintendo, switch	PlayStation 4, Nintendo Switch, Regno Unito
2	music	450	21	Fleurs .(Vinile Ross...	0.36605	vinile, esclusiva, amazon.it	Esclusiva Amazon.it, Notti Brave, Playlist
3	appliances	383	22	Wpro SKS101 Kit di a...	0.22555	lavatrice, libera, installazione	Litri, Ciarra, Argento
4	electronics	363	26	RFID Blocking Porta...	0.12357	led, luce, impermeabile	iPhone, Confezione, Lampada Solare
5	lighting	313	38	Echo Dot (3 generazione)	0.27407	led, smart, wifi	Google Home, Alexa, Amazon Alexa
6	tools	313	37	Gfone 8 Set di strum...	0.18641	led, lampada, batteria	Batteria, Tablet, Lampada
7	musical-instruments	274	29	Cordes pour guitare...	0.25264	chitarra, corde, acustica	Corde, Chitarra, Basso
8	baby	240	21	jane, Mutande post-p...	0.27871	chicco, pezzi, bambini	Confezione, Grigio, 2 Pezzi
9	luggage	240	14	Travel Buddy, Lucche...	0.15322	viaggio, bagaglio, mano	Bagaglio, Pollici, Ryanair
10	jewelry	239	17	Fablcrew - Bracciale...	0.25089	argento, regalo, braccialetto	Sterling 925, Orecchini, San Valentino

Tabella 4: Tabella statistiche delle top 10 communities

Nelle figure 10 e 11 si possono vedere rispettivamente le word clouds delle communities 5 e 7. Dalle immagini è facilmente intuibile che la prima verte su prodotti smart per la casa, mentre la seconda comprende accessori e prodotti legati alla chitarra. Ciò è confermato anche dalle entità più frequenti.

Figura 10: Word cloud della community 5

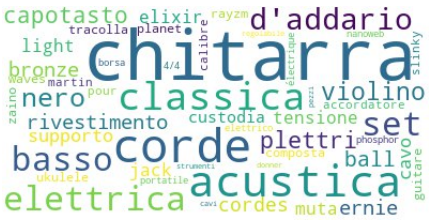

Figura 11: Word cloud della community 7

4 Sentiment analysis recensioni

La sentiment analysis sulle recensioni è stata affrontata sotto diversi aspetti per raggiungere tre obiettivi principali. Il primo obiettivo è stato quello di verificare la corrispondenza tra la positività individuata dal numero di stelle e la polarità ottenuta tramite un approccio basato sul lessico. Il secondo obiettivo è stato quello di capire quale fosse il sentimento medio delle 3 stelle, le quali dovrebbero corrispondere ad un voto neutro. Questo obiettivo nasce dal fatto che una recensione a 3 stelle, seppur neutra, possa presentare delle sfumature più tendenti al positivo o negativo (es: "Prodotto di qualità mediocre, ma per il prezzo che ha fa il suo lavoro... Consigliato per chi vuole spendere poco" e "Non è il massimo... ci sono alternative migliori sul mercato e non credo che lo ricomprerei." potrebbero essere recensioni da 3 stelle che esprimono sentimento diverso). Infine, come ultimo obiettivo è stato reputato interessante lo studio dell'evoluzione nel tempo del sentimento, rapportato anche all'evoluzione del numero di stelle, di alcuni prodotti d'esempio.

Nel seguito viene presentato il lavoro attraverso due approcci in quanto il primo (4.1), basato sul lessico, ha portato a scarsi risultati, per cui ad esso è stato preferito un approccio basato sul machine learning (4.2). In entrambi gli approcci l'attributo della recensione su cui viene condotta l'analisi è quello del *body*. È stato inoltre deciso di etichettare come positive le recensioni con $rating > 3$ e come negative quelle con $rating < 3$ per poter dare una stima dell'accuratezza dei risultati ottenuti.

4.1 Approccio Lexicon Based

La lingua italiana del dataset porta ad avere una cerchia più ristretta di alternative per effettuare una sentiment analysis. L'approccio che è stato adottato fa affidamento al lessico di Sentix, grazie al quale è stato possibile assegnare alle recensioni una polarità ottenuta dalla somma delle polarità delle parole considerate. Ogni record del lessico è costituito da un lemma, dal tipo (nome, verbo, aggettivo, ecc.) e da diversi scores[2].

Questo metodo è stato testato su un campione casuale di 1000 recensioni etichettate come positive e 1000 come negative. Le recensioni in questione sono state estratte da un sottoinsieme attendibile, ottenuto tramite l'intersezione di quelle di acquisti verificati ($verified = true$) e di quelle considerate utili da almeno un utente ($helpful > 0$). Considerando come positive le recensioni con polarità > 0,

[2] `http://valeriobasile.github.io/twita/sentix.html`

come negative quelle con polarità < 0 e scartando quelle per cui Sentix non ha riscontrato parole nel lessico, si ottiene un'accuratezza pari a 0.605 . In tabella 5 è disponibile la matrice di confusione, dalla quale si può notare che la maggior parte delle recensioni venga classificata come positiva. Questo risultato è coerente con quanto si era potuto osservare dalla word cloud in figura 4 mostrata nel capitolo 2, dalla quale si non si nota un netto spicco delle parole negative.

	NEG	**POS**
NEG	279	718
POS	68	929

Tabella 5: Matrice di confusione classificazione recensioni con Sentix

È stato inoltre verificato che non ci fosse una correlazione tra i valori della polarità rispetto al rating, sia aggregato per 1-2 e 4-5 stelle che preso singolarmente. Questo per stabilire se le polarità potessero essere coerenti con il rating considerando un intervallo di valori diverso per distinguere recensioni positive e negative. I valori ottenuti come coefficiente di correlazione di Pearson sono rispettivamente di 0.369 e 0.347, quindi si può concludere che non ci sia correlazione.

Gli scarsi risultati hanno portato a scartare l'approccio e a prenderne in considerazione uno alternativo per poter perseguire gli obiettivi preposti sulla tendenza di utilizzo delle 3 stelle e sull'andamento della polarità nel tempo.

4.2 Approccio Machine Learning

Anche in questo caso, avendo a disposizione un numero di recensioni molto elevato, si è deciso di utilizzare il sottoinsieme filtrato su quelle che sono sia verificate che considerate utili da almeno un utente in quanto considerate più affidabili per addestrare il modello. Per bilanciare il numero di recensioni positive molto più alto di quelle negative da dare in input al modello è stato fatto undersampling ed è stato ottenuto un insieme di 321,642 esempi divisi in due tra le classi. Tutte le recensioni sono state divise in token e ripulite da: stopwords, punteggiatura (sono stati aggiunti alcuni simboli molto ricorrenti nel testo delle recensioni), numeri, tokens di lunghezza 1 ed è stato effettuato lo stemming. Questo passaggio è essenziale per consentire l'eliminazione di rumore e quindi un migliore funzionamento dell'algoritmo di machine learning.
Data la natura binaria della variabile target, il modello scelto è stato quello di *LogisticRegression*, un modello di regressione non lineare. Il modello utilizzato per

rappresentare il testo delle recensioni in un linguaggio interpretabile dal calcolatore è il *CountVectorizer* (*Bag of Words*).
Gli esempi da utilizzare per il modello sono stati divisi con proporzione 80% train e 20% test. Il modello è stato costruito sfruttando l'inseme di train e successivamente è stato valutato sull'insieme di test producendo un valore di accuratezza pari a 0.90. Nella tabella 6 è stata riportata la matrice di confusione ottenuta.
Successivamente il modello è stato validato tramite la tecnica 10-Fold Cross-Validation per verificare la veridicità dei risultati ottenuti precedentemente. Tramite quest'ultima tecnica sono stati ottenuti risultati in linea con quelli ottenuti nel test.

	NEG	**POS**
NEG	28866	3253
POS	2808	29402

Tabella 6: Matrice di confusione classificazione recensioni con modello di regressione logistica

Con il modello ottenuto è stato poi possibile calcolare il sentimento delle recensioni su alcuni prodotti d'esempio per osservarne l'andamento nel tempo rispetto al trend delle stelle. In questo caso, per avere un insieme di valori più ampio e tracciare un grafico migliore, sono state considerate anche le recensioni non verificate e non marcate come utili. Dopo aver effettuato la previsione del sentimento, è stata calcolata la media dei valori raggruppati per mese ed è stata tracciata la curva nel tempo. Come si può vedere in figura 12, la curva del sentimento e quella del numero medio di stelle[3] seguono lo stesso andamento. Inoltre, è stato calcolato un valore per rappresentare il trend della curva del sentimento, il quale corrisponde al coefficiente angolare della retta che meglio approssima la curva. Dato un prodotto, è quindi possibile verificare se il sentimento delle recensioni sia in crescita o meno.

[3] i valori medi del numero di stelle sono stati portati in una scala compresa tra 0 e 1

Figura 12: Andamento di sentimento e numero di stelle nel tempo per il prodotto "Kingston Dtig4/32Gb Datatraveler Memoria Flash, US"

Come ultima cosa, è stato utilizzato il modello per verificare il sentimento medio rispetto all'utilizzo delle 3 stelle. I risultati ottenuti mostrano come il 46.8% (figura 13) delle recensioni sia positivo, dimostrando quindi che non ci sia una tendenza dominante verso un sentimento.

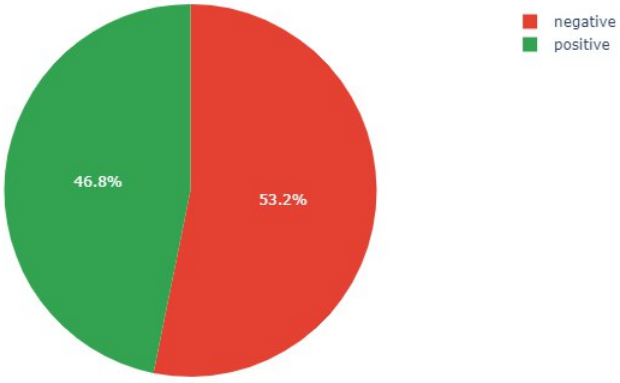

Figura 13: Risultati ottenuti applicando il modello costruito sulle recensioni neutre

5 Conclusioni

In questa analisi sono stati affrontati aspetti legati alla rete dei prodotti e al sentimento delle recensioni.

Per quanto riguarda la rete dei prodotti, essa è stata costruita sulla base delle relazioni di acquisto che li lega. Effettuandone l'analisi, come prima cosa è stato possibile identificare i prodotti che più spesso compaiono in combinazione con l'acquisto di altri beni. Inoltre sono stati individuati i prodotti che garantirebbero una migliore raggiungibilità partendo da altri acquisti e che quindi hanno una maggiore visibilità all'interno della rete.
È stata poi rilevata la presenza di communities formate da gruppi di prodotti spesso acquistati in combinazione. Ciò ha permesso quindi di mostrare come una divisione guidata dagli acquisti sia più significativa rispetto al semplice raggruppamento per categorie. Per ogni community sono stati individuati i prodotti più rilevanti e fornite le sfere d'interesse.

Passando alla parte relativa alla valutazione del sentimento delle recensioni è stato riscontrato che un approccio basato sul lessico sia risultato poco efficace. Ha raggiunto risultati migliori, invece, un approccio di machine learning supervisionato che utilizza il modello della regressione logistica, il quale è stato utilizzato per raggiungere gli obiettivi preposti. Tramite questo modello è stato possibile fornire l'andamento del sentimento nel tempo per specifici prodotti, notando una forte coerenza con l'andamento del numero di stelle. Inoltre, è stato dimostrato che l'utilizzo medio delle 3 stelle da parte degli utenti non avesse tendenze verso una specifica polarità.