Sentiment Analysis del dataset Amazon Fine Food Reviews

Luca Gandolfi 807485, Stefano Sacco 807532

July 2019

1 Introduzione

L'evoluzione delle reti sociali e il crescere di utenti in grado di lasciare una propria opinione in merito a qualcosa ha portato alla necessità di sviluppare tecniche di Natural Language Processing (NLP) sempre più affidabili. Sentiment Analysis è un task fondamentale nel NLP, permettendo non solo di analizzare in modo automatico le informazioni che l'utente vuole comunicare, ma anche di definire delle vere strategie di mercato in base ai feedback.

Il progetto svolto, consiste nell'effettuare una Sentiment Analysis sul dataset Fine Food Reviews di Amazon. I principali task che verranno trattati in questa analisi sono i seguenti:

- analisi preliminare del dataset
- analisi del sentiment delle recensioni utilizzando Afinn e vaderSentiment
- aspect-based analysis utilizzando LDA
- aspect and sentiment analysis utilizzando ASUM

Infine, verranno discussi i risultati ottenuti.
La repository **gitlab** del progetto si trova a questo indirizzo.

2 Analisi del dataset

Il dataset utilizzato è un estratto del famoso dataset FineFoods di Amazon avente oltre 500.000 record. Nell'estratto, si è scelto di tenere soltanto 9 attributi e 100.000 righe. In particolare, ogni riga rappresenta una review di prodotti Fine Food da Amazon. Le review sono state raccolte dal 1999 al 2012, tuttavia, nell'estratto, non sono presenti review del 1999. Le informazioni che abbiamo a disposizione sono:

- productid: id univoco associato ai prodotti recensiti
- userid: id univoco associato all'utente che sta effettuando la recensione

- profile: nome utente associato ad un userid
- helpFulNum: numero di valutazioni positive sulla review
- helpFulDen: numero di valutazioni totali sulla review
- score: classica valutazione da 1 a 5 stelle del prodotto Amazon da associare alla recensione
- text: contenuto della review
- time: yyyy-mm-dd. Data in cui è stata pubblicata la recensione
- year: anno in cui è stata pubblicata la recensione. Questa informazione è ridondante, ma in fase di estrazione del dataset è stato valutato essere più comoda averla a disposizione in una colonna separata.

ogni record è caratterizzato da una coppia ¡productid; userid¿ come chiave identificativa.

2.1 Statistiche sulle review

Nel dettaglio, abbiamo 12.560 prodotti e 70396 utenti. Il numero di votazioni non è omogeneo, infatti molti utenti hanno recensito 1 solo prodotto mentre quello che ne ha recensiti di più ne ha recensiti 86. La cosa strana è che alcuni utenti hanno recensito più volte lo stesso prodotto. Basti pensare che il prodotto più recensito è stato recensito 629 volte ma solo da 609 utenti diversi. Vi sono poi prodotti che hanno ricevuto una sola recensione.

2.2 Pre-processing del testo

Per poter svolgere alcuni task di analytics, è stato necessario processare il testo, poichè il linguaggio naturale presenta molte eterogeneità. Il processo seguito è stato il seguente:

1. Conversione di tutte le parole in lower-case.
2. Rilassamento delle forme contratte. E.g. isn't -¿ is not, it's -¿ it is.
3. Applicato la funzione word_tokenize() della libreria nltk al fine di ottenere una lista di token per ogni recensione.
4. Rimozione delle *stopwords*.
5. Rimozione della *punteggiatura*.

Al termine di questa fase, è stata aggiunta una nuova colonna al dataset contenente le parole tokenizzate per ogni review.

2.3 Parole più frequenti e stemming

Dopo aver svolto i task di pre-processing, è stato deciso di visualizzare le parole più frequenti utilizzate nelle review. In particolare, le 3 parole più frequenti sono state *like* con 14.262 apparizioni, *good* con 11.269 apparizioni e *coffee* con 9.773 apparizioni. Questo lascia immaginare come l'argomento caffè sarà molto discusso, ma questa anlisi verrà fatta in seguito. I risultati sono stati visualizzati attraverso una Wordcloud, che, tuttavia, ha mostrato i risultati dal quinto in poi. Infatti, come è possibile osservare in Figura 1, la parola più grande visualizzata è *one* e la seconda più grande è *taste*, che appaiono rispettiva in 5a posizione con 9.589 e 6a posizione con 9.464 utilizzi.

Figure 1: Wordcloud parole più frequenti

Per un'analisi più approfondita, è stato applicato ai token un processo di *Stemming*, nel dettaglio Lancaster's Stemming. Questo processo permette di mantenere soltanto la radice delle parole. A questo punto, è stata visualizzata nuovamente una Wordcloud contenente le parole stemmate più frequenti, la quale è visibile in Figura 2. In questo caso, i token più frequenti sono stati mostrati. Infatti, i 3 token più frequenti sono *lik*, che viene utilizzato 161.111 volte, *tast*, che viene utilizzato 14.658 volte e *us*, che viene utilizzato ben 12.500 volte. Come possiamo notare, la parola good è stata scavalcata dalla famiglia di parole con radice "tast" e "us". A grandi linee però possiamo dire che il trend di parole più frequenti è stato mantenuto.

Notiamo inoltre, che le parole più frequenti siano comunque parole positive, questo lascia intendere che la maggior parte delle recensioni avrà una opinione positiva. Nel prossimo sotto-paragrafo andremo a valutare gli score per vedere se questa idea fatta è confermata.

Figure 2: Wordcloud parole più frequenti dopo lo stemming

2.4 Analisi dei punteggi

L'analisi dei punteggi è sicuramente una delle analisi più importanti in questo dataset, perchè fornisce una idea generale dell'andamento dei prodotti presenti nel dataset. Gli score sono una valutazione da 1 a 5 espressa dall'utente che sta scrivendo la review. I risultati ottenuti sono mostrati in *Table 1*.

score	counts
1	9.318
2	5.568
3	8.059
4	14.643
5	62.412

Table 1: Numero di review in base allo score assegnato.

Come è possibile notare, la maggior parte delle recensioni ha portato una valutazione positiva. Questo risultato era già ipotizzabile osservando le parole più frequenti, anch'esse tutte positive. La valutazione è quindi sproporzionata. A fronte di questi risultati, è possibile svolgere diverse analisi:

- individuare i prodotti che hanno una media di stelle alta e quelli che hanno una media di stelle bassa.
- individuare gli utenti che hanno valutato più positivamente e quelli che hanno valutato più negativamente.
- valutare gli score in relazione al punteggio sulle recensioni ottenuto con una tecnica Lexicon-Based.

2.4.1 Media punteggi per prodotti

Come prima analisi approfondita, è stata ricercata la media degli score ottenuti per ogni prodotto. La media ha riportato 378 valori diversi per 4.224 prodotti. Inoltre, ben 1.689 prodotti hanno ricevuto una media di 5 stelle. Allo stesso modo, ben 221 prodotti hanno ricevuto una media di 1 stella.

2.4.2 Media punteggi per utenti

In seguito, è stata effettuata un'analisi sul comportamento dei vari utenti. In particolare, si è cercato di individuare gli utenti più "buoni" e quelli più "cattivi", dove si intende la quantità di stelle assegnate ai vari prodotti che hanno recensito: un utente buono assegna sempre stelle alte, mentre un utente cattivo il contrario. Per fare questo è stata presa in considerazione la media di score per ogni utente. Dalle prime analisi però, ci si è resi conto che molti utenti avevano poche o addirittura solo una recensione, pertanto ottenere dei dati sul comportamento di questi utenti era ininfluente. E' stato deciso quindi, di concentrare la nostra analisi su utenti che hanno effettuato almeno 15 recensioni. La media degli score ottenuta è raffigurata in Figura 3, dove sono ordinati i **top10** e **worst10** utenti.

userid	counts	score_mean
A2E3WMF9RWW2X2	16	5.000000
A1GQAKL9CGQLP1	17	5.000000
ADS5APY1NKTL4	30	4.966667
A1LZJZIHUPLDV4	24	4.958333
A1Q7A78VSQ5GQ4	21	4.904762
A3OXRFCJI67IMN	19	4.894737
A281NPSIMI1C2R	64	4.890625
A1XGFW5016CGQI	17	4.882353
A1P2XYD265YE21	33	4.878788
A1HOXKR7OKJ1X1	16	4.875000

userid	counts	score_mean
AW41Q5K4R499D	16	1.000000
A3TVZM3ZIXG8YW	35	1.000000
AF3BYMPWKWO8F	16	2.000000
AKZKG2Z7CNV27	21	2.142857
A2TN9C5E4A0I3F	30	2.666667
A1SCANWWQTEG9I	16	2.750000
A1RRMZKOMZ2M7J	20	2.750000
A1AEQZM99LO9VA	19	3.105263
A3HJHV83O2U8BL	17	3.117647
AR7TAEEUDHMUB	16	3.187500

Figure 3: Top 10 utenti vs worst 10 utenti

Nella prima immagine è stata riportata la classifica dei 10 più "buoni" mentre nella seconda quelli più "cattivi". Da questi dati si può notare come i primi due utenti di ogni classifica abbiano sempre scelto di assegnare 5 e 1 stella per i prodotti recensiti. E' stato quindi deciso di analizzare nome e periodo di attività di questi 4 utenti e i risultati sono stati davvero interessanti. Per quanto riguarda i primi due utenti con 5 stelle, i loro nomi e il loro periodo di attività

sono, in ordine:

- K. Duvall first review: 2007-01-05 and last review: 2012-06-25
- L. M. Keefer first review: 2011-08-03 and last review: 2012-06-17

Nulla di strano per loro, sembrano utenti normali e molto positivi sui prodotti comprati. Tuttavia, osservando i valori per quanto riguarda i primi due utenti dell'altra classifica:

- mom of 2 first review: 2012-08-11 and last review: 2012-08-11
- christopher hayes first review: 2010-12-04 and last review: 2010-12-04

Possiamo notare come abbiano effettuato tutte le recensioni negative nello stesso giorno: marketing? concorrenza? E' evidente come questa situazione debba essere approfondita per comprendere meglio.

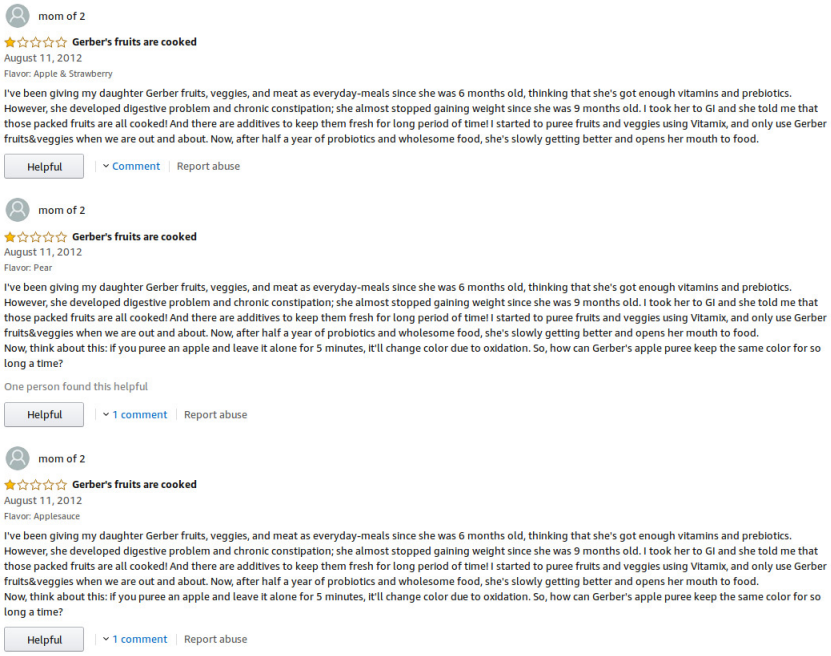
2.4.3 Analisi dei 2 peggiori utenti

Per effettuare questa analisi in modo approfondito, abbiamo scelto di utilizzare il dataset originale contenente 500.000 righe. Come prima cosa, è stato ritenuto importante capire l'orario in cui queste recensioni sono state effettuate, dato che sono nello stesso giorno. Il risultato ottenuto ha mostrato come l'orario sia sempre lo stesso per tutte le recensioni. In altre parole, l'utente *mom of 2* ha effettuato tutte le recensioni circa alle 2:00:00 del 2012-08-11, mentre l'utente *christopher hayes* ha effettuato tutte le recensioni circa alle 1:00:00 del 2010-12-04. La cosa più sensata è quindi pensare a due bot, tuttavia non sappiamo l'approssimazione che è stata tenuta riguardante l'orario. Per verificare se si trattassero di bot o meno, sono stati visualizzati i contenuti delle recensioni. Ovviamente, in entrambi i casi, la recensione scritta dall'utente era sempre la stessa e talvolta anche molto lunga. A questo punto, si è reso necessario controllare se il dataset fosse stato perturbato di proposito, con tanto di copia e incolla per aumentare il numero di score negativi. Il passo seguente è stato quindi controllare il product id di questi prodotti recensiti. I risultati sono stati i seguenti:

- 7 prodotti diversi per l'utente "mom of 2"
- 38 prodotti diversi per l'utente "christopher hayes "

Per il primo utente abbiamo che per ognuno dei prodotti sono state scritte in media 4 recensioni, mentre per il secondo utente abbiamo 18 recensioni in media per ogni prodotto.

Andando ad analizzare manualmente sulla scheda del prodotto amazon, tramite product id, è stato riscontrato che le recensioni del primo utente sono tutte rivolte a prodotti alimentari per la prima infanzia come omogenizzati della marca americana Gerber, come è possibile vedere in Figura 4.


mom of 2

Gerber's fruits are cooked

August 11, 2012

Flavor: Apple & Strawberry

I've been giving my daughter Gerber fruits, veggies, and meat as everyday-meals since she was 6 months old, thinking that she's got enough vitamins and prebiotics. However, she developed digestive problem and chronic constipation; she almost stopped gaining weight since she was 9 months old. I took her to GI and she told me that those packed fruits are all cooked! And there are additives to keep them fresh for long period of time! I started to puree fruits and veggies using Vitamix, and only use Gerber fruits&veggies when we are out and about. Now, after half a year of probiotics and wholesome food, she's slowly getting better and opens her mouth to food.

Helpful | Comment | Report abuse

mom of 2

Gerber's fruits are cooked

August 11, 2012

Flavor: Pear

I've been giving my daughter Gerber fruits, veggies, and meat as everyday-meals since she was 6 months old, thinking that she's got enough vitamins and prebiotics. However, she developed digestive problem and chronic constipation; she almost stopped gaining weight since she was 9 months old. I took her to GI and she told me that those packed fruits are all cooked! And there are additives to keep them fresh for long period of time! I started to puree fruits and veggies using Vitamix, and only use Gerber fruits&veggies when we are out and about. Now, after half a year of probiotics and wholesome food, she's slowly getting better and opens her mouth to food.
Now, think about this: if you puree an apple and leave it alone for 5 minutes, it'll change color due to oxidation. So, how can Gerber's apple puree keep the same color for so long a time?

One person found this helpful

Helpful | 1 comment | Report abuse

mom of 2

Gerber's fruits are cooked

August 11, 2012

Flavor: Applesauce

I've been giving my daughter Gerber fruits, veggies, and meat as everyday-meals since she was 6 months old, thinking that she's got enough vitamins and prebiotics. However, she developed digestive problem and chronic constipation; she almost stopped gaining weight since she was 9 months old. I took her to GI and she told me that those packed fruits are all cooked! And there are additives to keep them fresh for long period of time! I started to puree fruits and veggies using Vitamix, and only use Gerber fruits&veggies when we are out and about. Now, after half a year of probiotics and wholesome food, she's slowly getting better and opens her mouth to food.
Now, think about this: if you puree an apple and leave it alone for 5 minutes, it'll change color due to oxidation. So, how can Gerber's apple puree keep the same color for so long a time?

Helpful | 1 comment | Report abuse

Figure 4: Esempio di review dell'utente mom of 2.

Per quanto riguarda il secondo utente invece, non sono state trovate le recensioni descritte, probabilmente si è trattata di una campagna denigratoria contro il marchio americano Hill che produce cibo per gatti. In conclusione, in entrambi i casi è possibile pensare a due bot che hanno selezionato una serie di prodotti delle stesse marche in tempi davvero ravvicinati e hanno spammato lo stesso commento. Il secondo soprattutto per il numero esagerato registrato, anche se sulle pagine dei prodotti non sono state trovate tracce.

2.5 Evoluzione nel tempo

Un aspetto importante che è possibile considerare, è l'evoluzione degli score nel tempo. Amazon ha iniziato la sua crescita economica a partire dal 2003, superando lo scetticismo iniziale. Infatti, a partire dal 2003, è possibile notare un forte aumento di utenti e, di conseguenza, di recensioni. La crescita del numero di review ha portato non solo ad una crescita esponenziale del numero di valutazioni con 5 stelle, segno di un possibile abuso del sistema, ma anche ad una decresita dello score medio, considerando che l'utente tende ad abituarsi ad un servizio man mano che lo utilizza diventando più critico. L'utilizzo di bot, sia per generare valutazioni positive che negative, come abbiamo visto nel precedente paragrafo è in realtà un problema abbastanza grave. Per far fronte a questa situazione, Amazon, negli ultimi anni, sta inserendo una verifica degli account.

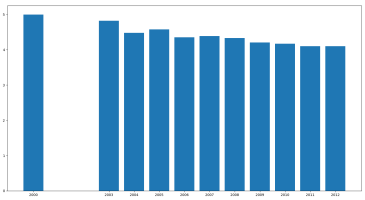


Figure 5: Media degli score totali per anno

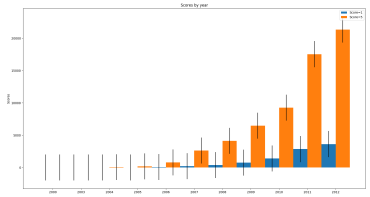


Figure 6: Crescita numero valutazioni 5 stelle e 1 stella

3 Sentiment analysis con Afinn e vaderSentiment

Dopo aver effettuato una prima analisi dei dati, si è scelto di analizzare il sentiment delle review con due *lexicon-based* tools con l'obiettivo ultimo di individuare eventualmente frasi potenzialmente ironiche confrontando lo score fornito dall'utente con lo score ottenuto dai tool.

3.1 Afinn

AFINN lexicon è uno dei più famosi lexicon contenente oltre 3.300 parole con uno score di polarità, positivo o negativo, associate. Questo lexicon lavora parola per parola, associando alle parole utilizzate nelle review presenti nel dizionario lo score, che può variare da -5 (super negativo) a +5 (super positivo). Il punteggio della review è dato dalla somma di tutti gli score associati alle parole presenti. Il problema principale di questa tecnica è l'eterogeneità dei sentiment espressi nelle review, nonchè la possibilità che pareri positivi su una parte di review vadano ad annullarsi contro pareri negativi espressi su un argomento diverso.

Per poter effetuare un confronto concreto con vaderSentiment, i risultati sono stati appiattiti su 3 valori: -1, 0 e +1. Processo che però ha contribuito a perdere molta informazione, osservando che i risultati ottenuti variavano da un minimo di -49 a un massimo di +164.

3.2 vaderSentiment

vaderSentiment è un tool lexicon e rule-based molto popolare in ambito social. La sua popolarità è dovuta al fatto che tiene conto anche della punteggiatura, oltre che delle parole. Inoltre, l'output fornito per ogni review associa una probabilità ad ogni polarità possibile: positiva, neutra, negativa. Inutile dire che la somma delle probabilità deve dare 1. Questo "peso" associato può essere molto utile, come è utile anche una quarta informazione contenenuta nell'output ovvero il valore di *compound*, che indica un valore tra -1 e +1 per l'intera review.

Per poter effettuare un confronto concredo con AFINN lexicon, è stato scelto di tener conto del risultato compund e, seguendo le linee guida, è stato assegnato ad ogni review un valore tra -1, 0 e +1. In particolare, i valori compresi tra -1

e -0.05 sono stati assegnati alla classe -1 e allo stesso modo alla classe 1 sono stati assegnati tutti gli score compresi tra +0.05 e +1, mentre alla classe 0 gli score compresi tra -0.05 e +0.05.

3.3 Confronto dei risultati

Da una prima analisi è possibile osservare che numericamete, le distribuzioni legate agli score ottenuti sono molto simili per le review positive mentre iniziano a discostarsi per quanto riguarda le altre due classi.

tool	-1	0	+1
AFINN	8.174	4.088	87.738
vaderSentiment	10.416	2.050	87.534

Table 2: Confronto distribuzioni di polarità tra AFINN e vaderSentiment.

Andando ad analizzare alcune review che presentano uno score di 5 stelle e valutazione sia di Afinn sia di vader negativa, possiamo osservare come:

- alcune recensioni presentano un velo di ironia, come ad esempio la seguente recensione: *"this was sooooo deliscious but too bad i ate em too fast and gained 2 pds! my fault"*.
- alcune recensioni citano altre che criticano alcuni aspetti, dicendo infine che non si sono verificati i problemi descritti. Una review così è molto complessa da analizzare con metodi a dizionario.
- alcune recensioni come il prodotto sia ottimo nel caso in cui un altro prodotto non sia di tuo gradimento, come ad esempio nel seguente estratto di una review: *"if you're feeling tired during the day and hate coffee, give this a shot"*. Anche in questo caso è molto complesso.

Allo stesso modo, analizzando alcune review che presentano uno score di 1 stella e valutazione sia di Afinn sia di vader positiva, possiamo osservare come:

- due aspetti discordanti siano un problema per entrambi i tool, come ad esempio si nota leggendo la seguente recensione: *"I'm not sure why Amazon is selling it for $9.99 for a box of 24 singles. Hazelnut coffee creamer is my favorite, but truly this is not a good buy."*
- la presenza di keyword positive non è sempre sinonimo di score positivo, basti vedere una recensione come: *"Would never reccommend this. Paid a good chunk of cash for nothing"*. In questo caso, la parola good è presente come rafforzativo del costo.

Problemi di questo tipo potrebbero essere risolti attraverso tecniche di machine learning, che però non verranno analizzate. In tabella 3 è possibile osservare un confronto rispetto al numero di falsi positivi e falsi negativi dei due tool. con

falsi positivi si intende tutte le volte che il tool assegna 1 mentre la recensione ha score 1, mentre con falsi negativi si intende tutte le volte che il tool assegna -1 quando in realtà la recensione ha score 5. Osservando la tabella, si può notare come vader sia leggermente più performante di Afinn ad individuare recensioni negative.

tool	falsi negativi	falsi positivi
AFINN	2.127	4.928
vaderSentiment	2.244	4.626

Table 3: Confronto falsi negativi e falsi positivi tra AFINN e vaderSentiment.

4 LDA

Il Topic Modeling è un tipo di modeling statistico per la scoperta di "topic" (argomenti) che si presentano in una serie di documenti. **Latent Dirichlet Allocation (LDA)** è un esempio di topic model ed è utilizzato per classificare testo in un documento riguardo a uno specifico topic.

LDA è quindi in grado di rilevare quali sono i topic più frequenti tra una serie di documenti e le parole più frequenti per i determinati topic, costruendo un modello basato sulle distribuzioni Dirichlet. Sostanzialmente una volta che viene fornito all'algoritmo il numero di topic che vogliamo trovare, esso si occupa di riordinare le distribuzioni dei topic tra i documenti e la distribuzione delle parole chiave tra i topic per ottenere un buon mapping topic-keywords. É stato deciso di applicare questa tecnica su un dataset ridotto, composto da circa 35.000 record.

4.1 Preprocessing dei documenti

Come azione preliminare alla costruzione del modello LDA, è stato necessario:

- effettuare un preprocessing dei documenti. Questo comprende tokenization delle frasi, portare tutto in lower case, rimozione di stopwords e punctuation.
- sono stati creati **bigram** e **trigram** sulle nostre frasi.
- una volta ottenuti i bigram sulle nostre frasi, è stata importata la libreria **sPacy** necessaria per il task di **Natural Language Processing (NLP)** attraverso il quale è stata performata la **lemmatization** delle frasi considerando solo i **NOUN, ADJ, VERB e ADV**.
- come ultimo passo sono stati creati i due input per la costruzione del modello LDA. Il primo è un dizionario **id2word** creato a partire dalle frasi lemmatizzate ed il secondo è il **corpus**, ottenuto sempre dalle frasi lemmatizzate e dal dizionario id2word appena realizzato.

Di fatto avremo una struttura con un id per ogni parola nei documenti. Il corpus è composto come mostrato qua sotto. Abbiamo un mapping tra **(word_id, word_frequency)**: [[(0, 1), (1, 2), (2, 1), (3, 1), (4, 1), (5, 1), (6, 5), (7, 1), (8, 1), (9, 2), (10, 1), (11, 1), (12, 1), (13, 1), (14, 1), (15, 1), (16, 1), (17, 1), (18, 1), (19, 1), (20, 1)]]

Per esempio, (0,1) implica che la parola con id 0 compare una volta nel primo documento, così come la parola con id 1 compare 2 volte e così via.

4.2 Costruzione del modello LDA

Per la costruzione del modello LDA sono richiesti questi ultimi due documenti appena creati, **Corpus** ed il **Dizionario**. É stato scelto di effettuare la ricerca di 10 topic differenti nel nostro dataset. Dopo un analisi dei risultati si è giunti alla conclusione che i due topic predominanti riguardano:

- bevande e in particolar in modo positivo sul *the*, sono presenti infatti parole come *taste, like, love, flavor*. Questo fa immaginare che probabilmente alcuni articoli sul the siano molto graditi dagli utenti in questione.
- cibo per per animali e in particolar modo i cani, infatti tra le parole principali compare *dog*. Tra le altre parole sono presenti anche *healty, natural*, pertanto si potrebbe pensare che gli utenti ritengano salutari gli articoli di cibo per cani presenti.

In generale i primi **cinque topic** sono rappresentati dalle seguenti parole. In ordine è riportato numero del topic e la lista della parole con la probabilità associata. In particolare le prime 5 sono quelle più rappresentative del topic:
(0,
'0.055*"tast" + 0.055*"good" + 0.053*"lik" + 0.045*"tea" + 0.043*"lov" + 0.041*"flav" + 0.036*"gre"'),
(1,
'0.109*"dog" + 0.080*"food" + 0.069*"eat" + 0.040*"tre" + 0.025*"help" + 0.024*"healthy" + 0.023*"nat"'),
(2,
'0.103*"keep" + 0.055*"cle" + 0.043*"dry" + 0.030*"fee" + 0.027*"continu" + 0.026*"send" + 0.026*"min"),
(3,
'0.100*"cup" + 0.041*"larg" + 0.034*"big" + 0.027*"expect" + 0.023*"bar" + 0.021*"husband" + 0.021*"sav"),
(4,
'0.086*"ad" + 0.041*"receiv" + 0.033*"whit" + 0.031*"top" + 0.024*"bold" + 0.023*"oz" + 0.023*"mapl_syrup"),

5 ASUM

Aspect and Sentiment Unification Model, *ASUM*, è una tecnica di Sentiment Analysis, in particolare una estensione di *Sentence-LDA* che unisce la ricerca dei topic più frequenti nelle Online Review divise per frasi con l'individuazione di un sentiment associato ai topic. Questa tecnica permette un'analisi molto più approfondita delle review non strutturate, grazie soprattutto alla divisione in frasi. Le normali analisi della polarità di una review possono essere meno attendibili a causa di pareri discordanti su aspetti (topic) differenti, il che porterebbe ad un bilanciamento di polarità in realtà non vero. La divisione in topic permette di individuare i topic dei quali si parla male (in media) e quelli dei quali si parla bene.

Per consistenza di dati anche in questo caso è stato deciso di applicare ASUM al dataset ridotto, composto da circa 35.000 record, per motivi computazionali. E' inutile dire che utilizzando un dataset più esteso i risultati potrebbero essere più concreti e reali.

5.1 Parametri

ASUM richiede in input:

- un file di testo contenente tutte le singole parole presenti nelle review, una per riga. Le parole richiedono un pre-processing per rimuovere stopwords, punteggiatura e applicare stemming.
- un file di testo contentente informazioni in formato numerico su ogni review. Precisamente, è necessario specificare in una riga il numero di topic presenti in quella review , nelle righe successive viene indicato l'indice di apparizione nel file descritto precedentemente di ogni token presente nel topic.
- un file di testo contenente alcune parole riconosciute come polarità positiva
- un file di testo contenente alcune parole riconosciute come polarità negativa

Per un migliore processing delle parole negate, es. *not good*, durante la fase di pre-processing tutte le negazioni sono state incollate con la parola successiva, se presente. Es. *notgood*.

5.2 Train e risultati

L'obiettivo di questa analisi era quello di migliorare ulteriormente le informazioni ottenute con LDA, andando quindi a formulare considerazioni in merito al trend dei prodotti Fine Foods di Amazon. Per questo motivo, si è scelto di impostare i parametri di ASUM come quelli utilizzati precedentemente in LDA. In particolare, il modello è stato allenato per individuare 10 topic positivi e 10 topic negativi, processando il dataset con 1000 iterazioni.

I risultati ottenuti hanno mostrato come, nel periodo tra il 2000 e il 2012 su Amazon, i topic che hanno mostrato sentiment più positivo riguardano bevande come il *thè* e il *caffè*, oltre a prodotti da sgranocchiare come *patatine e snack salati*. Anche il *cibo per cani* ha ricevuto un sentiment positivo. I topic che hanno mostrato sentiment più negativo invece riguardano ancora cibo per animali domestici, in particolare per *gatti*, e, restando in tema di animali, anche *giocattoli per cani*. Anche uno dei topic più positivi si è rivelato essere uno dei più negativi, ovvero quello del *caffè*.

coffee(p)	coffee(n)	chip(p)	tea(p)	dog food(p)	pet food(n)
coff	cup	chip	tea	dog	food
flav	coff	flav	tast	tre	cat
cup	us	tast	flav	food	dog
tast	mak	best	lik	lov	eat
lik	wat	lov	drink	good	problem

Table 4: Alcuni topic inferiti con ASUM. (p) indica il sentiment positivo, (n) il sentiment negativo.

Come è possibile notare, se per alcuni topic sono presenti entrambi i sentiment, per altri è presente soltanto un solo sentiment. La tabella 4, inoltre, mostra i topic più evidenti, con probabilità delle prime parole molto alta. Analizzando l'intero output, è possibile notare come non sempre parole che prese singolarmente sono classificate positive, sono utilizzate effettivamente per esprimere positività, ad esempio patatine troppo dolci (sweet) possono non piacere, così come in alcune review si parla di come altri prodotti simili/dello stesso marchio erano piaciuti (lik) mentre questo no.

5.3 Confronto con LDA

Dopo aver analizzato i risultati ottenuti con ASUM ed LDA è possibile effettuare un confronto veloce. Utilizzando entrambi i modelli con l'obiettivo di trovare 10 topic, molti risultati appaiono simili. Infatti, tra i topic individuati da entrambi i modelli, è possibile notare il topic riguardante il *thè*, quello riguardante il *cibo per cani* e quello riguardante il *caffè*. Grazie ad ASUM è stato possibile approfondire alcuni aspetti di questi topic, come ad esempio è stato possibile osservare che per quanto riguarda il cibo per animali domestici, quello per i gatti tenda ad essere sempre molto negativo, mentre quello per cani è sbilanciato verso il sentiment positivo.

ASUM fornisce una visione più reale dei fatti, analizzando il sentiment associato a ciascun topic, permettendo di differenziare in sotto-topic individuando cosa è più criticato e cosa più apprezzato. Inoltre, ASUM ha trovato topic meno astratti di LDA, basti pensare infatti che il topic su patatine e quello sui popcorn non sono stati individuati da quest'ultimo.

6 Conclusioni

In questo progetto sono state valutate diverse tecniche di Natural Language Processing, con l'obiettivo di comprendere il sentiment espresso dagli utenti della categoria Fine Foods di Amazon. Il percorso svolto per raggiungere il nostro scopo è partito analizzando ogni review con tecniche a dizionario come Afinn e vaderSentiment, individuando i loro limiti e la loro accuratezza. Quindi, abbiamo intrapreso la strada del topic modelling, analizzando quali sono gli aspetti e i topic più discussi in questa categoria di Amazon, arrivando a considerare di questi il sentiment espresso.

In media, è stato possibile notare come la lunghezza media delle review sia rimasta abbastanza costante negli anni, con uno sbilanciamento significativo del numero di caratteri nelle review con 5 stelle. Questo risultato è stato mantenuto nonostante la crescita esponenziale del numero di review dopo il 2003.

Come confermato dai risultati ottenuti con Afinn e vaderSentiment, il dataset si è rivelato abbastanza sbilanciato verso review positive, questo è stato un problema durante l'utilizzo del tool di ASUM, per quanto riguarda il setting dei parametri ottimali. Analizzando i topic più discussi con due tecniche diverse, è stato possibile studiare come non sempre un topic molto discusso sia in realtà discusso positivamente, nonostante lo sbilanciamento del dataset, e come sia possibile che un topic molto discusso sia però molto apprezzato come molto criticato, in base agli stessi parametri data la varietà di clientela che Amazon offre. Quest'ultimo punto è possibile osservarlo guardando i risultati ottenuti con ASUM nella Tabella 4, leggendo le *most common words* per il topic sul caffè.

Grazie a queste analisi potrebbe essere possibile svolgere dei veri e propri ragionamenti in ottica di marketing. Infatti, individuare le categorie, o meglio, i prodotti dove la clientela è più propensa a lasciare valutazioni positive, consentirebbe ai produttori di questi articoli di commercializzarli sulla piattaforma riscuotendo in media un discreto successo. Allo stesso modo, per i produttori di articoli con clientela molto critica, potrebbe essere conveniente analizzare in modo approfondito gli aspetti negativi per migliorarli.

Infine, osservando i risultati, se siete produttori di *tea* o *patatine*, e avete intenzione di lanciare un nuovo prodotto sul mercato di Amazon, questo potrebbe andare alla grande. Invece, se siete produttori di cibo per animali, questo potrebbe ottenere poco successo, in particolare per quanto riguarda il cibo per gatti, dove i clienti discutono spesso di problemi intestinali, rischiando di rovinare l'opinione comune del marchio a causa delle tante lamentele.