DATA ANALYTICS

Sentiment analysis on food reviews

Autori:

Antonio De Palma (874351)

Marco Di Capua (878295)

Michele Fontana (829658)

Sommario

Introduzione

Il progetto in descrizione ha come scopo quello di eseguire una sentiment analysis su un dataset contenente delle recensioni su dei prodotti di cibo venduti su Amazon. La valutazione dei prodotti su Amazon è una funzione cardine della piattaforma di e-commerce e l'analisi del sentimento di tali recensioni fornisce importanti informazioni alle compagnie per poter migliorare il proprio prodotto o, più in generale, analizzare come questo sia stato recepito dal pubblico.

Il progetto si pone come obiettivo quello di individuare degli approcci per effettuare delle previsioni sul sentiment, tali per cui sia possibile associare una valutazione ad una recensione, a partire dal contenuto del corpus della recensione stessa.

Nello specifico si utilizzeranno degli approcci basati su lessici e degli approcci basati su tecniche di apprendimento supervisionato. Per il primo tipo di approccio, saranno utilizzati i lessici di AFINN, YELP restaurant reviews e NRC emotion lexicon. Per il secondo tipo di approccio, invece, verranno addestrati due modelli di apprendimento supervisionato: Naive Bayes e regressione logistica. Verranno dunque valutate le misure di performance per ogni modello e verrà infine eseguito il confronto tra questi ultimi per stabilire quale sia il migliore da utilizzare in questo ambito.

Descrizione Dataset

Il dataset utilizzato contiene dei dati relativi alle recensioni di prodotti di cibo venduti su Amazon. In particolare, il dataset contiene al suo interno 35172 recensioni collezionate fino al 2012 con le seguenti colonne:

- **productid:** codice identificativo del prodotto Amazon
- **userid:** codice identificativo dell'autore della recensione
- **score:** valutazione del prodotto espressa in un punteggio compreso tra 0 e 5
- **text:** corpus della recensione del prodotto

Il dataset non possiede alcuna istanza contenente dei valori NA

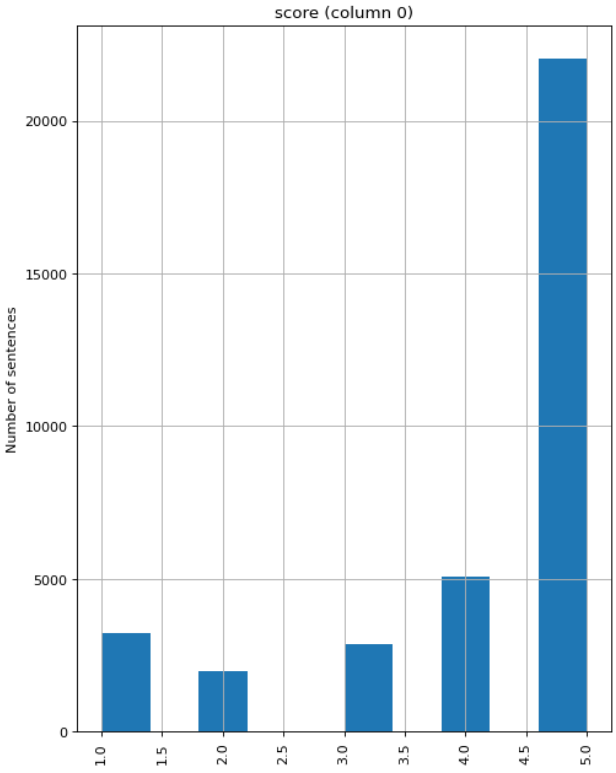
	productid	userid	score	text
0	B001E4KFG0	A3SGXH7AUHU8GW	5.0	I have bought several of the Vitality canned d...
1	B00813GRG4	A1D87F6ZCVE5NK	1.0	Product arrived labeled as Jumbo Salted Peanut...
2	B000LQOCH0	ABXLMWJIXXAIN	4.0	This is a confection that has been around a fe...
3	B000UA0QIQ	A395BORC6FGVXV	2.0	If you are looking for the secret ingredient i...
4	B006K2ZZ7K	A1UQRSCLF8GW1T	5.0	Great taffy at a great price. There was a wid...
5	B006K2ZZ7K	ADT0SRK1MGOEU	4.0	I got a wild hair for taffy and ordered this f...

La variabile target del nostro dataset è la variabile score che può assumere dei valori interi che hanno un range compreso tra 1 e 5.

Nel nostro caso d'uso è stato deciso di identificare come valutazioni positive tutte le recensioni che hanno ottenuto uno score superiore a 3, mentre negative tutte le recensioni con uno score minore o uguale a 3.

Analisi Esplorativa

Effettuiamo un'analisi preliminare del dataset. Analizzando la frequenza delle istanze della variabile score, si nota che il dataset non segue una distribuzione uniforme:



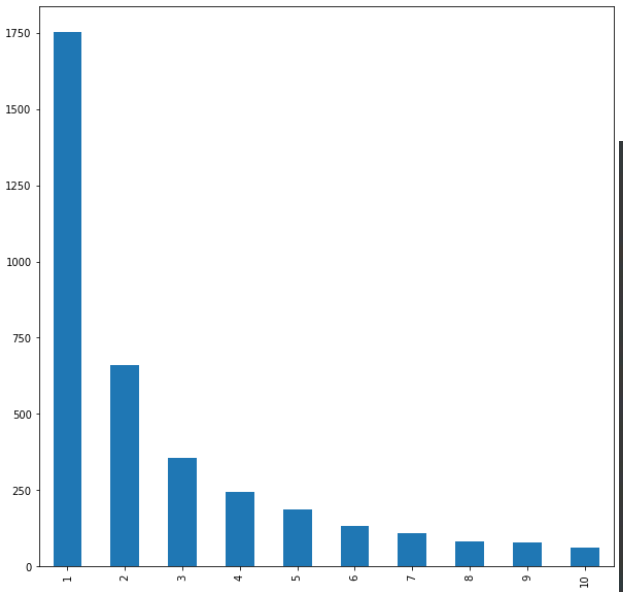
A fronte di queste informazioni, abbiamo deciso di non utilizzare l'accuracy come metrica, ma di utilizzare al suo posto l'AUC. Dato che l'AUC mostra come il modello è in grado di distinguere tra le due possibili classificazioni; quindi si avrà che maggiore sarà il valore dell'AUC, maggiore sarà la capacità predittiva del modello di apprendimento.

Tra le varie informazioni utili che possono essere estratte da un'analisi preliminare del contenuto del dataset, abbiamo estrapolato la frequenza di recensioni scritte per utente:

1	25563	28	1
2	2955	12	1
3	543	14	1
4	183	15	1
5	70	17	1
6	43	20	1
7	22	21	1
8	16	24	1
10	8	30	1
11	8		
9	7		
18	2		

Da questo elenco delle frequenze, si può notare come la maggioranza degli utenti - in genere - scrive una singola recensione rendendo così il dataset in oggetto particolarmente eterogeneo e, conseguentemente, adeguato ad un'analisi del sentimento che prende in considerazione una grande varietà di opinioni differenti.

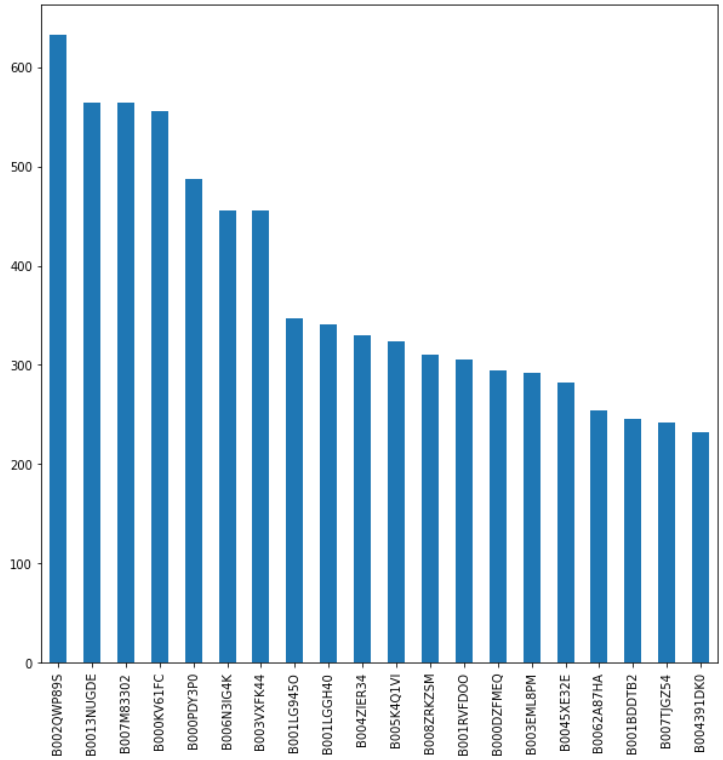
Una seconda informazione di particolare interesse riguardante il dataset, concerne il numero minimo e massimo di recensioni ricevute dai prodotti. È stato eseguito il plot di un primo grafico rappresentante il numero di prodotti che hanno ricevuto un numero basso di recensioni:



1	1751
2	659
3	356
4	244
5	186
6	131
7	109
8	81
9	78
10	63

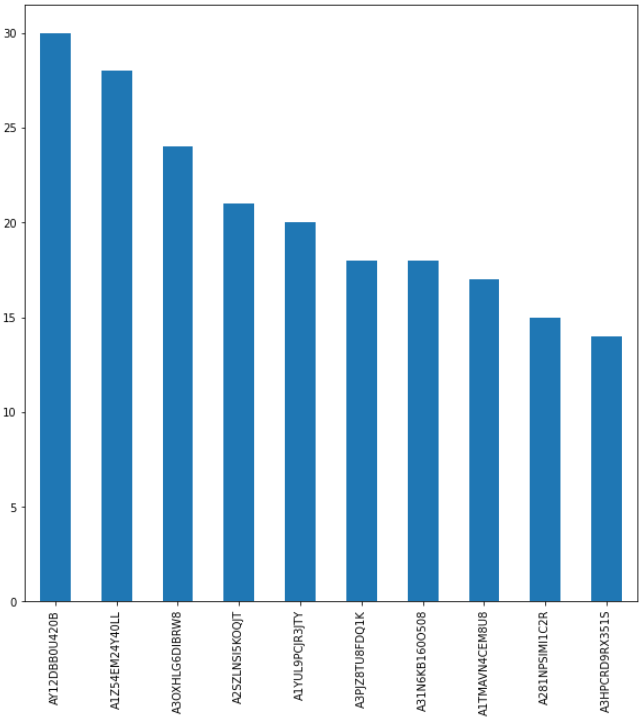
Dal grafico e dall'elenco delle frequenze risulta evidente che la maggior parte dei prodotti ha ricevuto un numero basso di recensioni, andando così ad aumentare l'eterogeneità del dataset.

Per quanto riguarda i prodotti che hanno ricevuto un numero maggiore di recensioni, è stato visualizzato il grafico che rappresenta i primi venti prodotti con il maggior numero di recensioni:



Tale grafico evidenzia come il prodotto di codice B002QWP89S che identifica del cibo per cani, risulta essere quello con il maggior numero di recensioni pari a 632.

Per concludere l'analisi esplorativa, è stato scelto di individuare gli utenti più attivi eseguendo un plot dei primi dieci utenti con più recensioni scritte:



Scelta degli approcci da utilizzare

Data la natura del dataset e dall'obiettivo preposto per lo svolgimento del progetto, sono stati selezionati due tipi di approcci: approcci basati sui lessici e approcci basati sui modelli di apprendimento supervisionato.

Gli approcci basati sui lessici sono utilizzati per poter classificare come positivo o negativo un testo, utilizzando un elenco di parole di cui si conosce la polarità. Nel nostro caso d'uso, le recensioni degli utenti verranno scomposte in token contenenti le singole parole che compongono il corpus della recensione e successivamente verranno processati dai lessici per individuarne la polarità. A tale scopo sono stati scelti tre tipi di lessici:

- **AFINN**: lessico generico contenente vocaboli della lingua inglese valutate con uno score compreso tra -5 e +5
- **YELP restaurant review**: lessico specifico contenente vocaboli della lingua inglese utilizzati per recensire i ristoranti
- **NRC emotion lexicon**: lessico generico contenente dei vocaboli della lingua inglese a cui sono associati delle emozioni di base e una valutazione positiva o negativa

Per quanto riguarda gli approcci basati sui metodi di apprendimento supervisionato sono stati scelti due modelli:

- **Naive Bayes**
- **Regressione logistica**

Tali algoritmi risultano essere particolarmente adatti per il nostro caso d'uso, in quanto sono tipicamente utilizzati per generare un modello in grado di prevedere eventi futuri, utilizzando dei dati riguardanti eventi passati. Nel dominio su cui tali algoritmi verranno applicati, lo scopo è quello di ottenere dei modelli in grado di classificare in maniera corretta lo score di future recensioni, a partire dal contenuto del loro corpus.

In tutti gli approcci adottati, i risultati ottenuti sono stati normalizzati per poter essere confrontati con gli score del dataset iniziale, i quali sono stati ricondotti a valutazioni positive e negative.

Natural language pre-processing

Da una prima analisi del contenuto delle recensioni, risulta evidente la necessità di una fase di pre-processing del corpus delle recensioni, al fine di andare a modellare degli esempi di minore ambiguità e maggiore apporto informativo possibile, da utilizzare per gli approcci basati su lessici e apprendimento automatico. Visualizzando quindi il corpus di alcune recensioni, ne sono state individuate alcune il cui corpus risulta essere particolarmente rumoroso, come l'esempio che segue:


Twizzlers, Strawberry my childhood favorite candy, made in Lancaster Pennsylvania by Y & S Candies, Inc. one of the oldest confectionery Firms in the United States, now a Subsidiary of the Hershey Company, the Company was established in 1845 as Young and Smylie, they also make Apple Licorice Twists, Green Color and Blue Raspberry Licorice Twists, I like them all

I keep it in a dry cool place because is not recommended it to put it in the fridge. According to the Guinness Book of Records, the longest Licorice Twist ever made measured 1.200 Feet (370 M) and weighted 100 Pounds (45 Kg) and was made by Y & S Candies, Inc. This Record-Breaking Twist became a Guinness World Record on July 19, 1998. This Product is Kosher! Thank You

Risulta evidente la necessità di una elaborazione del contenuto del corpus delle recensioni.

La prima tecnica di pre-processing attuata riguarda la sostituzione delle forme contratte delle parole, spesso utilizzata nella lingua inglese, tipicamente identificabile grazie alla presenza dei caratteri "\'". Tra le varie forme contratte, ad esempio, ne abbiamo riscontrato la presenza con la parola "not" che veniva espressa nella sua forma contratta con "n\'t". Dato che il significato delle parole risultava essere il medesimo, è stato deciso di ricondurre tutte le forme contratte dei vocaboli nella loro forma estesa, al fine di ridurre questi casi di simil sinonimia.

L'analisi del contenuto del corpus delle recensioni ha rivelato la presenza di altre tipologie di vocaboli che risulterebbero complicati o del tutto inappropriati all'applicazione delle tecniche basate sui lessici e apprendimento automatico. In particolare, si è notata la presenza di link a siti web facilmente individuabili dalla presenza di "http" prima di ogni link e, conseguentemente, altrettanto facili da rimuovere.

Infine, in questa prima fase di pre-processing, si è deciso di andare a trasformare il corpus delle recensioni in caratteri esclusivamente minuscoli; questa decisione è stata presa in quanto, nelle fasi successive, le parole contenenti dei caratteri maiuscoli sarebbero state considerate come due vocaboli differenti rispetto a delle parole equivalenti ma senza caratteri maiuscoli.

Dopo questa prima serie di procedure di pre-processing, il corpus delle recensioni risulta essere evidentemente migliorato rispetto alla visualizzazione precedente:

```
twizzlers strawberry my childhood favorite candy made in lancaster pennsylvania by y s candies inc one of the oldest confectionery firms in the united states now a subsidiary of the hershey company the company was established in as young and smylie they also make apple licorice twists green color and blue raspberry licorice twists i like them alli keep it in a dry cool place because is not recommended it to put it in the fridge according to the guinness book of records the longest licorice twist ever made measured feet m and weighted pounds kg and was made by y s candies inc this record breaking twist became a guinness world record on july this product is kosher thank you
```

```
[('the', 103444),
 ('i', 94880),
 ('and', 71568),
 ('a', 66115),
 ('it', 59594),
 ('is', 58050),
 ('to', 56015),
 ('of', 43049),
 ('this', 37900),
 ('not', 35625)]
```

Nonostante questa seconda rappresentazione risulti migliore è ancora evidente la necessità di una elaborazione del contenuto delle recensioni, in quanto la frequenza che individua le most common words denota la presenza, in grande quantità, di vocaboli non utili ad una elaborazione della sentiment analysis. Vocaboli come “the”, “and”, “this” e più in generale tutte le congiunzioni, pronomi, ecc.. sono ampiamente utilizzati nelle recensioni ma non apportano alcun tipo di beneficio nella valutazione del sentiment legato alle stesse. A fronte di questo problema, è stato deciso di rimuovere la presenza di questi vocaboli identificabili come stopwords.

Così come la presenza di stopwords non apporta alcun contributo alla definizione della polarità di una recensione, anche la punteggiatura segue lo stesso copione, perciò è stato deciso di rimuovere anch’essa dal corpus delle recensioni:

```
'twizzlers strawberry childhood favorite candy made lancaster pennsylvania candies inc one oldest confectionery firms united states subsidiary hershey company company established young smylie also make apple licorice twists green color blue raspberry licorice twists like alli keep dry cool place not recommended put fridge according guinness book records longest licorice twist ever made measu red feet weighted pounds kg made candies inc record breaking twist became guinness world record july product kosher thank'
```

```
[('not', 35625),
 ('like', 14454),
 ('good', 11466),
 ('coffee', 9993),
 ('great', 9878),
 ('one', 9722),
 ('taste', 9689),
 ('would', 9361),
 ('product', 8371),
 ('flavor', 8285)]
```

Il risultato ottenuto dalle tecniche di pre-processing attuate è stato ritenuto soddisfacente, tuttavia, per perfezionare ulteriormente i vocaboli da analizzare, si è deciso di utilizzare la tecnica di lemmatization per ottenere le forme canoniche delle parole. La tecnica di lemmatization riporta le parole alla loro forma canonica utilizzandone il lemma; ad esempio le parole “studies” e “studying” saranno ricondotte a “study”.

```
'twizzlers strawberry childhood favorite candy made lancaster pennsylvania candy inc one oldest confectionery firm united state subsidiary hershey company company established young smylie also make apple licorice twist green color blue raspberry licorice twist like alli keep dry cool place not recommended put fridge according guinness book record longest licorice twist ever made measured foot weighted pound kg made candy inc record breaking twist became guinness world record july product kosher thank'
```

```
[('not', 35625),
 ('like', 15116),
 ('taste', 12011),
 ('good', 11558),
 ('coffee', 10551),
 ('one', 10545),
 ('flavor', 10240),
 ('product', 9976),
 ('great', 9878),
 ('love', 9602)]
```

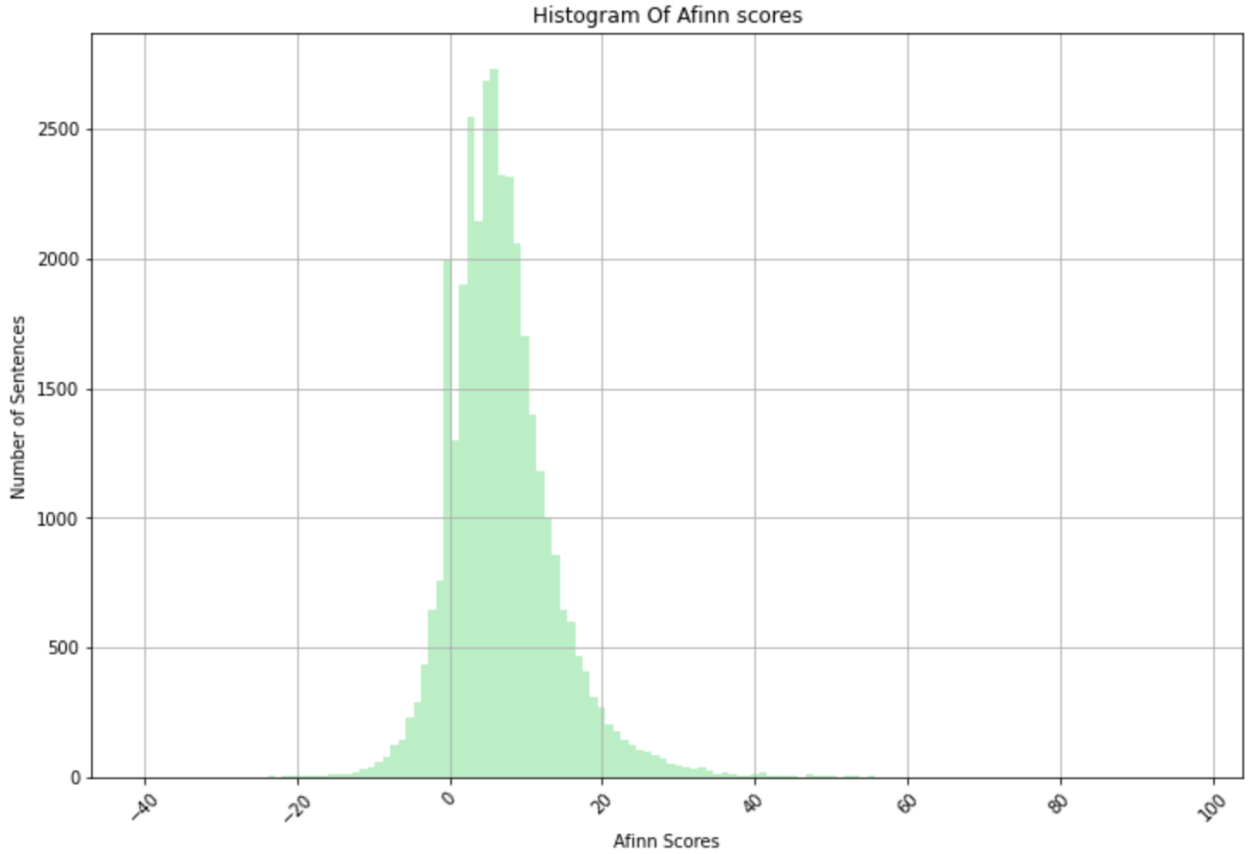
Una volta conclusa la fase di pre-processing, i dati risultano essere adeguati all’applicazione degli approcci di classificazione.

Approcci coi lessici

La prima tipologia di approccio adottata riguarda l’utilizzo dei lessici. La classificazione basata sui lessici processa il testo delle recensioni, individuando se ciascuna delle parole che le compongono appartengono ad un lessico di parole e ne attribuisce uno score; successivamente, una volta valutate tutte le parole di una recensione, si calcola la sommatoria di tutti gli score di tutte le parole di tale recensione per individuarne la

polarità. I lessici selezionati per questo esperimento sono: AFINN, YELP restaurant reviews e NRC emotion lexicon.

Il primo lessico utilizzato per la classificazione è AFINN. Il lessico AFINN è composto da un elenco di termini della lingua inglese a cui è stato affidato uno score compreso tra -5 e +5. Tale lessico è stato utilizzato per classificare gli interi corpus delle recensioni ottenendo i seguenti risultati:



Da questa rappresentazione si nota come la maggior parte delle recensioni abbia ottenuto uno score positivo; tale risultato risulta essere allineato con la distribuzione delle valutazioni del dataset iniziale.

Una volta normalizzati gli score di AFINN è stato possibile confrontare lo score del dataset iniziale con lo score assegnato da AFINN:

afinn	-1	1
score		
-1	2736	5348
1	2190	24898

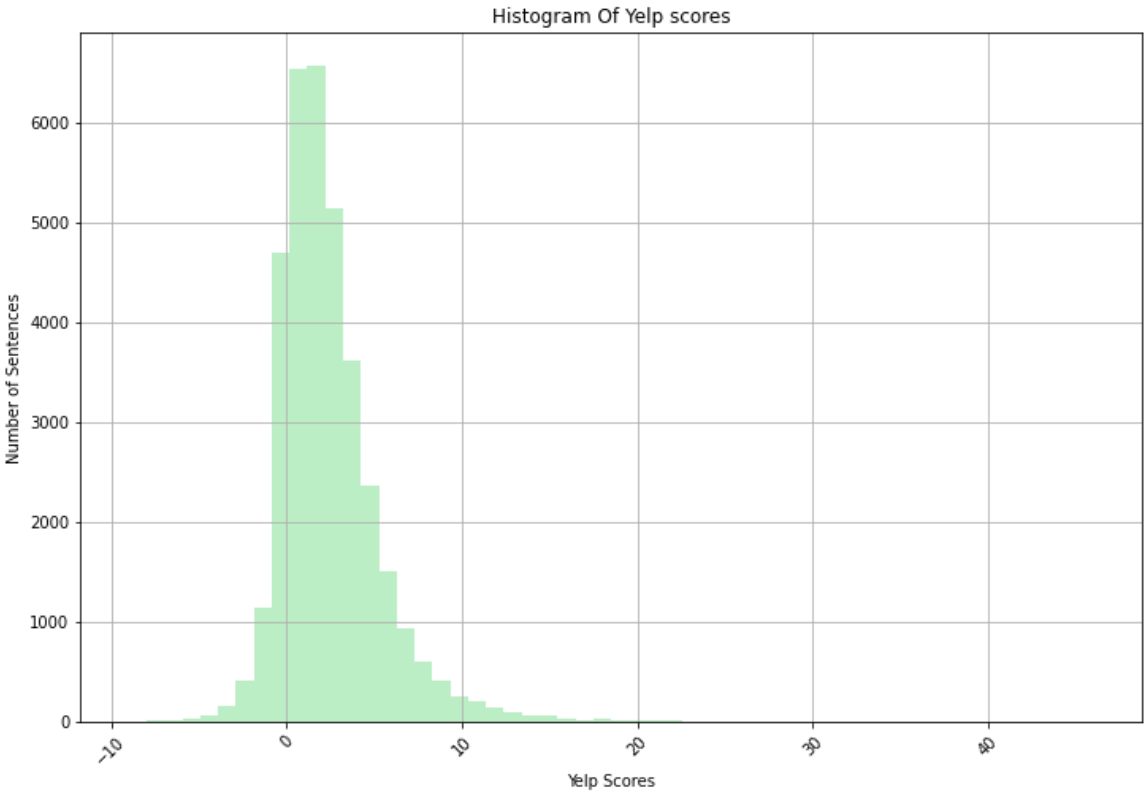
Infine, è stata calcolata l'accuracy del modello che corrisponde ad un valore di 0.785

Il secondo lessico selezionato per l'esperimento è il lessico "YELP restaurant reviews". Tale lessico è stato scelto in quanto potenzialmente più adatto alla classificazione delle recensioni del dataset, poichè possiede delle valorizzazioni di vocaboli tipicamente usati nelle recensioni dei ristoranti che risultano essere particolarmente affini con le recensioni di prodotti di cibo vendute su Amazon.

Tra i termini presenti nel lessico sono presenti dei vocaboli a cui segue la dicitura "_NEGFIRST"; tali vocaboli denotano la presenza di una negazione che antecede il vocabolo e dunque possiedono un valore di score differente dalla versione originale. Ad esempio, il vocabolo "disappoints" possiede uno score di -2,243, al contrario al vocabolo "disappoints _NEGFIRST" è stato assegnato uno score di +3,1. Grazie a queste diverse versioni di vocaboli, il lessico è in grado di distinguere con più precisione il valore dei vocaboli all'interno delle frasi e quindi di ottenere delle classificazioni più accurate.

A differenza del lessico AFINN, questa soluzione necessita di ricevere in input i token contenenti le singole parole che compongono il corpus delle recensioni.

Dopo aver classificato ogni singolo token di tutte le recensioni del nostro dataset, è stato eseguito un plot degli score assegnati dal lessico alle recensioni:



Confrontando gli score assegnati dal lessico con gli score originali tramite la matrice di confusione si ha che:

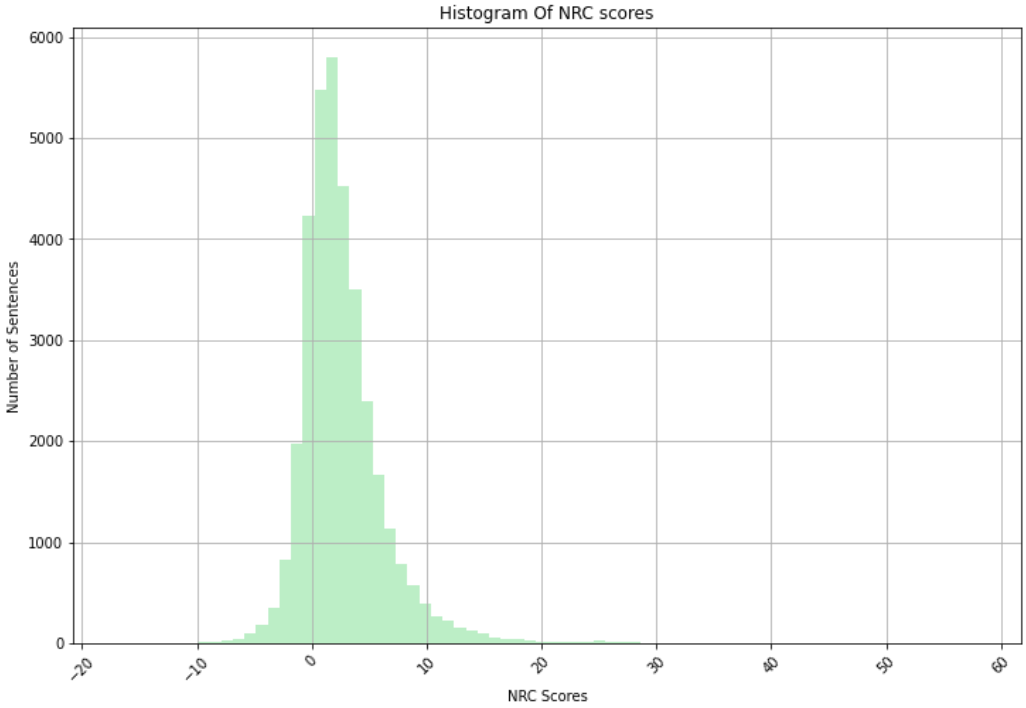
myscore_base_rest	-1	1
score		
-1	3611	4473
1	2910	24178

Da questa matrice è possibile notare come il lessico abbia delle difficoltà nel classificare correttamente le recensioni con polarità originale negativa.

Infine, l'accuracy ottenuta dal modello è pari a 0.7900

Il terzo ed ultimo lessico selezionato per la classificazione è il lessico NRC. NRC Word-Emotion Association Lexicon è un lessico generico contenente un elenco di vocaboli in diverse lingue di cui si conosce, oltre alla polarità, anche l'emozione associata. Le emozioni che possono essere associate ad un vocabolo sono le 8 emozioni base, ovvero: rabbia, anticipazione, disgusto, paura, gioia, tristezza, sorpresa e fiducia.

Gli score ottenuti dalla classificazione da parte del lessico risultano essere nuovamente conformi con la distribuzione delle valutazioni delle recensioni che tende verso delle valutazioni prevalentemente positive.



La matrice di confusione che segue mostra una evidente tendenza all'errore nella classificazione delle recensioni negative, mentre per le recensioni positive il risultato ottenuto è soddisfacente:

myscore_base_lemm	-1	1
score		
-1	2984	5100
1	4823	22265

La scarsa capacità del modello di classificare correttamente le recensioni negative si rispecchia anche nel valore della sua accuracy pari a 0.7178 che è la minore tra quelle dei tre approcci utilizzati.

Approcci supervisionati

Una volta terminati gli esperimenti basati sui lessici è stato deciso di utilizzare degli approcci basati sull'apprendimento supervisionato. Prima di procedere con l'addestramento dei modelli Naive Bayes e regressione logistica è necessario predisporre i dati, attraverso opportune

modellazioni, al fine di ottenere delle prestazioni soddisfacenti da parte dei modelli.

Il primo passaggio per predisporre i dati all'addestramento del modello consiste nel suddividere il dataset in training set e test set. I dati suddivisi, tuttavia, non risultano essere adeguati alla classificazione dei modelli di apprendimento, in quanto sono dati testuali; perciò è necessario convertire i dati in formato numerico. Per ottenere questo risultato è stato deciso di utilizzare tre tecniche:

- **Bag-of-Words**
- **Bag-of-Words Bigram**
- **TFIDF**

Bag-of-Words è una tecnica di NLP che si occupa di andare a processare e modellare del testo per estrarne le feature. Le operazioni svolte da BOW consentono quindi di convertire testi di dimensione variabile in dei vettori numerici di dimensione prefissata. I vettori vengono creati andando a rappresentare ogni singolo token che identifica una parola con la relativa occorrenza all'interno della frase in analisi.

Bag-of-Words Bigram è una variazione della classica BOW che consente di prendere in considerazione eventuali negazioni all'interno delle recensioni. Per fare ciò, il modello prende in considerazione delle coppie di token consecutivi consentendo in questo modo di poter individuare e, conseguentemente, interpretare in maniera corretta il significato di una frase contenente delle negazioni.

TFIDF, acronimo di Term Frequency Inverse Document Frequency, adotta un approccio differente dalle tecniche BOW sopracitate. Questo modello opera attraverso due fasi fondamentali: la prima fase si occupa di individuare la frequenza di una parola all'interno di un documento, mentre la seconda si occupa di andare a verificare la frequenza di tale parola all'interno di un set di documenti. Grazie a questo approccio, il modello individua tutte quelle parole con elevata frequenza che tuttavia non risultano essere particolarmente significative, come ad esempio il termine "the" della lingua inglese. Tanto più frequentemente la parola in

analisi è utilizzata all'interno di un set di documenti, tanto più il suo valore ne sarà penalizzato.

Una volta generati i training set e test set applicando i tre modelli sopracitati, il passo successivo prevede l'individuazione dei migliori iperparametri da fornire al modello di apprendimento. Gli iperparametri sono utilizzati come argomenti da trasmettere ai modelli, al fine di guidarne e ottimizzarne il processo di apprendimento.

Grid Search consente di individuare il migliore set di iperparametri attraverso una valutazione di tutte le possibili combinazioni. In questo modo sono stati estratti gli iperparametri che successivamente sono stati ordinati attraverso l'F1-score. Tale score indica un alto valore sia di recall che di precision ed è una tecnica di valutazione affine al confronto tra due o più modelli di apprendimento. Seguono i migliori iperparametri individuati da Grid Search per ognuno dei modelli utilizzati per generare i training e test set:

- **Bag-of-Words Unigram**: {'alpha': 1}
- **Bag-of-Words Bigram**: {'alpha': 1}
- **TFIDF**: {'alpha': 0.05}

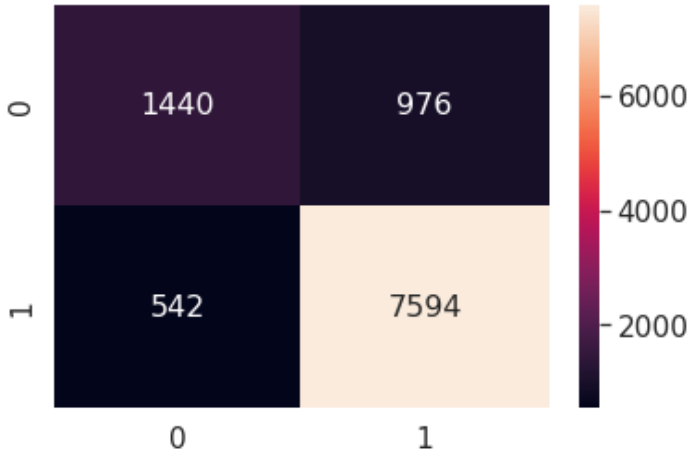
Una volta terminata la fase di preparazione dei dati, si procede con l'addestramento e la valutazione delle performance del primo modello: Naive Bayes.

Sono stati addestrati tre modelli sui training e test set ottenuti da BOW, Bigram e TFIDF ottenendo i seguenti risultati:

	AUC	Precision	Recall	F1-Score
BOW UNI	0.885	0.813	0.768	0.787
BOW BI	0.836	0.883	0.638	0.667
TFIDF	0.885	0.828	0.659	0.691

I dati appena riportati mostrano come, il modello addestrato sui training e test set ottenuti dall'applicazione di Bag of word unigram, risulta essere il più performante. Nel dettaglio è stato eseguito un plot della matrice di

confusione di tale modello per poter osservare come si comporta nella classificazione delle recensioni:



Come si può notare dalla matrice, il modello presenta delle difficoltà nella classificazione delle recensioni negative, tuttavia, la classificazione delle recensioni positive ha ottenuto dei risultati molto performanti.

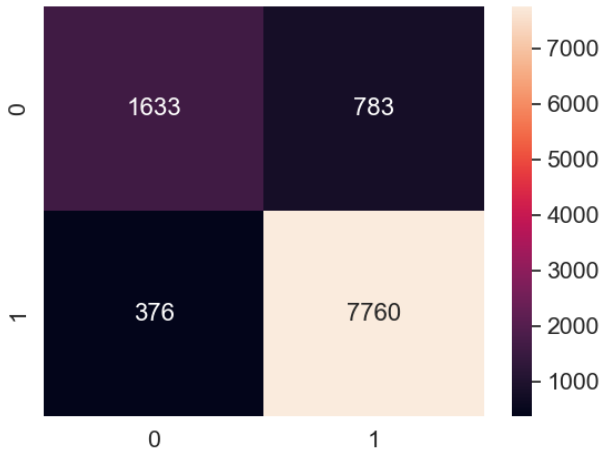
Analogamente al modello di apprendimento appena mostrato, si è addestrato un secondo modello di apprendimento supervisionato, ovvero la regressione logistica.

Sono stati nuovamente addestrati tre modelli basati sui training e test set ottenuti da BOW, Bigram e TFIDF con le seguenti performance:

	AUC	Precision	Recall	F1-Score
BOW	0.791	0.825	0.791	0.806
Bigram	0.814	0.861	0.815	0.834
TFIDF	0.796	0.834	0.796	0.812

Le performance ottenute da Bigram risultano essere migliori in tutte le misure effettuate, rendendolo quindi il modello più adatto tra i tre esaminati.

Analizzando la matrice di confusione del modello, si può notare come il problema riscontrato nella classificazione delle recensioni negative da parte del modello basato su Naive Bayes sia notevolmente ridotto nel caso del modello basato su regressione logistica:



Conclusioni

Le nostre analisi hanno portato a risultati ottimi sulla classe delle recensioni positive e a risultati non soddisfacenti su quelle negative. Una possibile motivazione a questo fenomeno riguarda la preponderanza di recensioni positive rispetto a quelle negative nel dataset. Una soluzione sarebbe quella di bilanciare la distribuzione delle classi tramite l'utilizzo di oversampling/undersampling. Per quanto riguarda l'approccio tramite lessici, i migliori risultati sono stati ottenuti utilizzando Yelp, principalmente per due motivi:

- Il lessico è più adatto per il nostro caso d'uso
- L'utilizzo dei termini con la dicitura NEGFIRST ha contribuito all'aumento di precisione nel calcolo della polarità

Allo stesso modo, il modello di apprendimento supervisionato con le migliori prestazioni, risulta essere la regressione logistica che utilizza bag of word bigram perché individua meglio le negazioni nel testo e garantisce migliori prestazioni sulla classe negativa.