Cross-domain Sentiment Analysis

Intelligenza Artificiale - Elaborato per l'esame finale

Niccolò Benedetto MAT. 7024656

Docente: Paolo Frasconi

1 Ambiente di sviluppo

Il progetto è stato sviluppato attraverso il servizio cloud *Google Colab*, nella sua versione standard, quindi con accesso limitato alle risorse di calcolo fornite direttamente da Google (GPUs, RAM). Per eseguire il codice in linguaggio Python (nella sua ultima versione, Python 3), si sfruttano dunque i Jupyter Notebook, un'applicazione web open-source che consente di creare e condividere documenti che contengono codice live e altre svariate risorse multimediali.

2 Analisi e implementazione

Il dataset di riferimento, su cui il modello dovrà operare, si presenta in questa forma:

	Unnamed: 0	drugName	condition	review	rating	date	usefulCount
0	206461	Valsartan	Left Ventricular Dysfunction	"It has no side effect, I take it in combinati...	9.0	May 20, 2012	27
1	95260	Guanfacine	ADHD	"My son is halfway through his fourth week of ...	8.0	April 27, 2010	192
2	92703	Lybrel	Birth Control	"I used to take another oral contraceptive, wh...	5.0	December 14, 2009	17
3	138000	Ortho Evra	Birth Control	"This is my first time using any form of birth...	8.0	November 3, 2015	10
4	35696	Buprenorphine / naloxone	Opiate Dependence	"Suboxone has completely turned my life around...	9.0	November 27, 2016	37

Figure 1: original dataset

L'obbiettivo ultimo è quello di effettuare un'analisi del sentimento cross-domain con riferimento alla colonna *review* del dataset sopra. Per analisi del sentimento cross-domain s'intende analizzare come il classificatore reagisce quando viene addestrato e testato su reviews appartenenti a domini diversi. Nel nostro caso i domini a cui viene fatto riferimento sono le condizioni che si presentano con maggior frequenza all'interno del dataset: *Anxiety*, *Birth Control*, *Depression*, *Diabetes Type 2* e *Pain*. Si è deciso inoltre di addestrare il modello, oltre che sui dati "grezzi" della colonna *review*, anche su una nuova colonna che viene aggiunta al dataset, detta questa *cleaned_review*, che contiene i dati della colonna originale elaborati dopo una fase di pre-processing del testo. Questa fase è stata realizzata mediante l'utilizzo della suite di librerie e programmi per l'analisi simbolica e statica nel campo dell'elaborazione dei linguaggi naturali (principalmente in lingua inglese), anche conosciuta come NLTK. Al fine di preparare il testo a una fase di analisi "pulita", il pre-processing adoperato prevede diversi steps tra cui la rimozione dei caratteri non alfanumerici, rimozione delle stop-words, rimozione dei tags HTML, la tokenizzazione e la conversione in minuscolo. In **Figure 2** è riportato una riga del dataset con l'aggiunta della colonna *cleaned_review*.

I testi delle reviews rappresenteranno dunque il contenuto delle variabili del modello. I valori invece delle etichette vengono generati a seguito di un'analisi

review	rating	date	usefulCount	cleaned_review
"It has no side effect, I take it in combinati...	9.0	May 20, 2012	27	side effect take combination bystolic 5 mg fis...
"My son is halfway through his fourth week of ...	8.0	April 27, 2010	192	son halfway fourth week intuniv became concern...
"I used to take another oral contraceptive, wh...	5.0	December 14, 2009	17	used take another oral contraceptive 21 pill c...
"This is my first time using any form of birth...	8.0	November 3, 2015	10	first time using form birth control 039 glad w...
"Suboxone has completely turned my life around...	9.0	November 27, 2016	37	suboxone completely turned life around feel he...

Figure 2: dataset after pre-processing phase on *review* column

dei testi eseguita attraverso il tool VADER, strumento inlcuso nella libreria NLTK, che sfrutta un vasto dizionario di parole etichettate con punteggi di sentimento. VADER valuta la polarità di un testo, i.e. effettua un'analisi del sentimento di questo, generando un dizionario con 4 chiavi:

- **negative key**, indica la proprorzione di negatività espressa dal testo.
- **positive key**, indica la proporzione di positività espressa dal testo.
- **neutral key**, indica la proporzione di neutralità espressa dal testo.
- **compound key**, si tratta di un valore composto che è generato per combinazione dei tre valori sopra.

I primi tre valori di questo dizionario sono compresi tra 0 e 1, dove 0 indica l'assenza del sentimento corrispondente e 1 la massima presenza di questo. Il quarto valore invece varia nel range [-1,1], dove valori > 0 indicano un sentimento positivo, < 0 un sentimento negativo e valori vicino a 0 indicano neutralità o assenza di un chiaro sentimento positivo o negativo. Si è pensato, in relazione al significato dei valori delle chiavi di compound definite dall'analisi VADER, di esprimere il sentimento di una review attraverso i seguenti intervalli: se il valore di compound key rientra nell'intervallo [-1;-0.3] allora il rispettivo valore di etichetta vale 0 (e corrisponde a un sentimento negativo), se cade in (-0.3;0.3) il valore di etichetta è 1 (sentimento neutro), mentre se è contenuto in [0.3;1], etichetta prende il valore 2 (sentimento positivo).

review	rating	date	usefulCount	cleaned_review	cleaned_review-dict-VD	cleaned_review-numberlist-VD	standard_review-dict-VD	standard_review-numberlist-VD
"It has no side effect, I take it in combinati...	9.0	May 20, 2012	27	side effect take combination bystolic 5 mg fis...	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...	(0.0, 1.0, 0.0, 0.0)	{'neg': 0.121, 'neu': 0.879, 'pos': 0.0, 'comp...	(0.121, 0.879, 0.0, -0.296)

Figure 3: dataset after VADER analysis

Si aggiunge quindi una nuova colonna al dataset, detta questa *rating_model*, che mappa i valori di compound key, tenendo riferimento alla logica sopra esposta.

review	rating	date	usefulCount	cleaned_review	cleaned_review-dict-VD	cleaned_review-numberlist-VD	key_neg	key_neu	key_pos	key_compound	rating_model
"It has no side effect, I take it in combinati...	9.0	May 20, 2012	27	side effect take combination bystolic 5 mg fis...	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...	(0.0, 1.0, 0.0, 0.0)	0.0	1.0	0.0	0.0	1

Figure 4: adding column *rating-model*

Il dataset, dopo esser stato manipolato in maniera tale da contenere tutte le informazioni necessarie per addestrare il modello viene suddiviso con un rate test_size = 0.33 (i.e. il 67% dei dati viene utilizzato per l'addestramento e il restante 33% per il test). Il modello dispone di 4 variabili: X_train (reviews di addestramento), X_test (reviews di test), y_train (etichette di addestramento, che corrispondono appunto a valori della nuova colonna *rating_model*), e y_test (etichette di test). La logica secondo cui si è pensato di estrarre le lexical-features dai testi delle reviews, che devono essere fornite all'algoritmo Perceptron come input, sfrutta il principio di funzionamento della classe TfidfVectorizer del modulo scikit-learn, che estrae le caratteristiche da un testo per convertirle in una rappresentazione vettoriale basata sulla frequenza delle parole e sulla loro importanza nel corpus complessivo. Vengono dunque considerate le parole delle reviews come unità di analisi, più dettagliatamente si valutano gli unigrammi (le singole parole) e i bi-grammi (coppie di parole adiacenti), differentemente dal lavoro descritto in Gräßer et al. 2018, in cui si considerano anche i tri-grammi. La scelta è stata obbligata causa le limitate risorse offerte dalla versione standard di *Google Colab*.

3 Studio dei risultati

L'analisi cross-domain ci impone poi di utilizzare iterativamente Perceptron per addestrare e testare il modello creato su reviews appartenenti a porzioni del dataset originale, ottenute considerando tutte le possibili combinazioni conseguibili incrociando coppie di domini.

```
[PERCEPTRON ON CLEAN DF]
------------------------------------------------------------
['X_BC_trainCl', 'y_BC_trainCl', 'X_BC_testCl', 'y_BC_testCl']
------------------------------------------------------------
Accuracy of Perceptron is 0.875381538785391
------------------------------------------------------------
['X_D_trainCl', 'y_D_trainCl', 'X_BC_testCl', 'y_BC_testCl']
------------------------------------------------------------
Accuracy of Perceptron is 0.6879275865698348
------------------------------------------------------------
```

Figure 5: partial output of Perceptron

Ad esempio il primo valore di *accuracy* della **Figure 5** indica il risultato dell'esecuzione di Perceptron quando il modello è addestrato su reviews appartenenti al dominio *Birth Control* e testato su etichette sempre appartenenti

allo stesso dominio. Il secondo valore di *accuracy* è riferito al risultato predittivo di Perceptron quando il modello è addestrato su reviews appartenenti al dominio *Depression* e testato su etichette appartenenti al dominio *Birth Control*.

train \ test	Anxiety	Birth Control	Depression	Diabetes, Type 2	Pain
Anxiety	0.800924	0.676981	0.710992	0.659549	0.678501
Birth Control	0.731657	0.875382	0.753091	0.696323	0.700690
Depression	0.707542	0.687928	0.850317	0.666667	0.689349
Diabetes, Type 2	0.612622	0.638354	0.650184	0.792408	0.646450
Pain	0.636737	0.652247	0.655530	0.625148	0.813116

Figure 6: cross-domain analysis on clean dataset

train \ test	Anxiety	Birth Control	Depression	Diabetes, Type 2	Pain
Anxiety	0.765008	0.622566	0.649516	0.604982	0.592209
Birth Control	0.651616	0.856962	0.678249	0.655991	0.659763
Depression	0.659826	0.652563	0.821918	0.601423	0.631164
Diabetes, Type 2	0.574654	0.612988	0.586702	0.744958	0.610454
Pain	0.598769	0.607515	0.634815	0.589561	0.784024

Figure 7: cross-domain analysis on dataset without text pre-processing

La tabella in **Figure 6** dunque raccoglie i valori di accuratezza dell'algoritmo quando il modello viene addestrato su reviews di uno specifico dominio (le righe della tabella) e viene testato su etichette di reviews di un altro specifico dominio (le colonne della tabella). Si osservi come i valori numerici della predizione siano decisamente elevati quando si addestra e si testa il modello su stessi domini, quella che prende il nome di analisi in-domain (ad esempio si prenda in considerazione il valore 0.875382 ottenuto per *Birth Control*), mentre l'accuratezza si può abbassare anche a 0.612622 (Perceptron predice correttamente con probabilità $\approx 60\%$) quando addestro sul dominio *Diabetes, Type 2* e testo su *Anxiety*. I numeri asseriscono inoltre che c'è un'alta correlazione tra i domini *Birth Control* e *Anxiety*, ossia si può pensare che coloro che hanno redatto reviews in cui viene fatto riferimento ad esempio a metodi contraccettivi con buona probabilità vi sarà un qualche accenno a sensazioni di ansia, così come tra *Depression* e *Anxiety*. Bisogna sottolineare che questi risultati sono strettamente legati alle

scelte intraprese durante la fase di analisi del progetto. Si consideri che il range degli intervalli con cui definisco le etichette del modello sono congrui con il significato del quantificatore generato dal tool VADER dopo la valutazione della polarità del testo, ma molto probabilmente non sono ottimizzati per raggiungere la migliore accuratezza predittiva. Manipolandoli diversamente andremo incontro sicuramente a una diversa produzione dei risultati.
Si concluda osservando l'analisi cross-domain ottenuta sul dataset in assenza della fase di pre-processing del testo (**Figure 7**). Se l'andamento dei valori e le relazioni tra domini ricalcano quello di **Figure 6**, l'accuratezza predittiva dell'algoritmo appare più imprecisa.