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 creare un modello che effettui 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 appunto contiene i dati della colonna originale a seguito di 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*.

La fase di analisi vera e propria del sentimento viene a questo punto implementata attraverso l'utilizzo del tool VADER, strumento inlcuso nella libreria

	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

NLTK che sfrutta un vasto dizionario di parole etichettate con punteggi di sentimento per valutare un testo. VADER valuta la polarità di un testo 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.

	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

Continuando con la preparazione del dataset per l'addestramento del modello, viene adesso aggiunta una nuova colonna, detta questa *rating_model*, che mappa i valori della colonna *rating*. La funzione che permette di eseguire la mappatura è stata così elaborata: il valore della colonna *rating_model* contiene un intero compreso nell'intervallo [0,2] e in particolare assume il valore 0 se il corrispondente valore di *rating* rientra nell'intervallo [1,4], 1 se rientra in (4,7) e 2 se è contenuto nel range [7,10].

review	rating	date	usefulCount	cleaned_review	cleaned_review-dict-VD	cleaned_review-numberlist-VD	key_neg	key_neu	key_pos	key_compound	ratings_cond	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	rating >= 7	2

Figure 4: adding column *rating-model*

L'aver eseguito questa mappatura della colonna *rating* permette una notevole semplificazione del modello, che a fronte della suddivisione del dataset di un rate test_size = 0.33 (i.e. il 67% dei dati viene utilizzato per l'addestramento e il restante 33% per il test), dispone di 4 variabili: X_train (review di addestramento), X_test (review di test), y_train (etichette di addestramento, che corrispondono appunto a valori della nuova colonna *rating_model*; si osservi dunque che in tal modo si raggiunge quella riduzione di complessità del modello prima accennata invece di utilizzare come etichette direttamente i valori della colonna *rating*) 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 a Perceptron(), 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.

```
['X_BC_trainCl', 'y_BC_trainCl', 'X_BC_testCl', 'y_BC_testCl']
Accuracy of Perceptron is 0.8790653615408904
['X_D_trainCl', 'y_D_trainCl', 'X_BC_testCl', 'y_BC_testCl']
Accuracy of Perceptron is 0.6147773918534891
```

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 ap-

partenenti 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*.

test train	Anxiety	Birth Control	Depression	Diabetes, Type 2	Pain
Anxiety	0.872242	0.589306	0.723689	0.665480	0.784024
Birth Control	0.508979	0.879065	0.524223	0.593120	0.574951
Depression	0.814264	0.614777	0.855997	0.686833	0.719921
Diabetes, Type 2	0.510518	0.580044	0.521550	0.855279	0.478304
Pain	0.758338	0.580149	0.690611	0.635824	0.874260

Figure 6: cross-domain analysis on clean dataset

test train	Anxiety	Birth Control	Depression	Diabetes, Type 2	Pain
Anxiety	0.877373	0.595516	0.755763	0.703440	0.792406
Birth Control	0.594664	0.882433	0.600401	0.657177	0.650394
Depression	0.824525	0.642353	0.861343	0.708185	0.771203
Diabetes, Type 2	0.630067	0.613199	0.617107	0.846975	0.547830
Pain	0.785531	0.594464	0.730037	0.682088	0.874753

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 0.874260 ottenuto per *Pain*), mentre l'accuratezza si può abbassare anche a 0.508979 (Perceptron() predice correttamente con probabilità $\approx \frac{1}{2}$) quando addestro sul dominio *Birth Control* e testo su *Anxiety*. I numeri asseriscono inoltre che c'è un alta correlazione tra i domini *Anxiety* e *Pain*, ossia si può pensare che coloro che hanno redatto reviews in cui viene fatto riferimento ad ansia allora molto probabilmente ci si aspetta che in quest'ultime ci siano accenni a sensazioni di dolore, così come tra *Depression* e *Anxiety*.

Si concluda osservando l'analisi cross-domain ottenuta sul dataset in assenza della fase di pre-processing del testo (**Figure 7**). I valori ricalcano quelli di **Figure 6**, anzi riportano quasi sempre una maggiore accuratezza. Questo potrebbe far pensare che se si tratta di classificare il sentimento su testi brevi e informali (come tweet), il pre-processing potrebbe non essere così critico quanto su testi più formali e strutturati.