Cross-domain Sentiment Analysis

Christian Di Buó

April 2022

Abstract

In questo progetto si utilizzano implementazioni dell'algoritmo Perceptron al problema dell'analisi del sentimento su recensioni di farmaci, come descritto in Gräßer et al. 2018, riproducendo risultati analoghi a quelli riportati in tabella 3 dell'articolo. Il progetto è stato interamente scritto in *python*, l'ambiente di sviluppo utilizzato è *PyCharm* (community edition). Il dataset utilizzato è disponibile sul repository di UCI.

1 Introduction

L'analisi del sentimento consiste nell'elaborazione del linguaggio naturale e nella costruzione di sistemi per l'identificazione ed estrazione di opinioni dal testo.

Per prima cosa è necessario importare il proprio Dataset e suddividerlo in due parti, una destinata al Trainig dell'algoritmo ed una al Testing. Generalmente il 70% viene destinato al training e il 30% al testing, ma in questo caso il Dataset del repository era già suddiviso nelle suddette parti.

Successivamente, come fatto nell'articolo, si è normalizzato l'attributo rating del Dataset sostituendo i valori superiori o uguali a 7 con 1, quelli compresi tra 4 e 7 con 0, e i restanti con -1.

Il Dataset conteneva molti più informazioni di quelle necessarie e sono pertanto stati estratti solo i due attributi necessari all'analisi del sentimento in questione, le recensioni ed i rating.

Prima di procedere con il training dell'algoritmo è necessario trasformare le recensioni, poiché l'algoritmo prende come input esclusivamente vettori numerici. Le recensioni vengono trasformate in Bag-of-world e, a tal fine, è stata utilizzata la classe CountVectorizer di scikit-learn.

Il termine Bag-of-world racchiude i processi di tokenizzazione, conteggio e normalizzazione. La tokenizzazione consiste nel suddividere il testo in tokens. I token possono variare a seconda che si scelga una suddivisione basata su unigrammi Table 1, bigrammi Table 2 o trigrammi Table 3, che comporterà poi una variazione sulla correttezza della classificazione dell'algoritmo.

Unigram example						
likely	perscriptions	unsure	concise	gotta	sports	walnut

Table 1: Esempio unigrammi

Bigram example						
burst or	039 ll	sudden acting	also right	039 exposed	reaction they	was tested

Table 2: Esempio bigrammi

trigram example					
the normal stuff	foods but	pill 10x	yay gained	of some common	able to and

Table 3: Esempio trigrammi

Gli unigrammi sono token che consistono in una sola parola di testo, mentre i bigrammi sono tutte le possibili combinazioni date da due parole del testo. Nella tokenizzazione viene poi assegnato un intero ad ogni token. La frequenza di ogni token verrà calcolata nella fase di conteggio, e nella normalizzazione verrà assegnato a ciascun token un peso inversamente proporzionale alla sua frequenza.

Non sono state rimosse le stop word, cioè le parole che in genere compaiono con una frequenza così elevata da non influenzare la classificazione dell'algoritmo, per cercare di riprodurre fedelmente i risultati dell'articolo. È stata utilizzata anche la classe TfidfTransformer, sempre di scikitlearn, che diminuisce l'impatto dei token che compaiono più frequentemente nel documento per dare maggior risalto a quelli con frequenza minore, che si presumono essere di maggior rilievo.

Dopo aver trasformato il testo in Bag-of-word è ora possibile utilizzarlo come input del classificatore, dove è stata impiegata la classe Perceptron di scikit-learn. Nell'inizializzazione della classe sono stati utilizzati gli attributi con i valori di default, poiché anche impostando, ad esempio, un numero di epoche inferiori i risultati non avrebbero subito variazioni eccessive. Tramite il metodo fit è stato effettuato il train del classificatore. Le istanze del metodo sono state le recensioni dei pazienti, mentre i rating costituivano le label. È stato effettuato il train su un dominio medico alla volta, facendo poi il testing sia nello stesso campo medico che nei rimanenti quattro. Le predizioni dell'algoritmo sono poi state valutate sulle effettive label assegnate ai dati del testing, tramite l'utilizzo del metodo accuracy_score.

2 Risultati

I risultati ottenuti mostrano la capacità dell'algoritmo di classificare correttamente le reviews tra i vari campi medici selezionati.

Dando in input al Perceptron i dati suddivisi in bigrammi, i risultati ottenuti sono mostrati nella Table 4.

I risultati differiscono leggermente rispetto a quelli riportati nell'articolo in quanto è stato utilizzato un classificatore diverso dal Perceptron mentre i risultati con input unigrammi e trigrammi sono mostrati in Table 5 e Table 3 rispettivamente. Come si può notare, mentre nel passaggio da unigrammi a bigrammi vi è un miglioramento sensibile nell'accuratezza del perceptron, lo stesso non avviene con l'utilizzo dei trigrammi, che può essere ricondotto alla dimensione molto ridotta del dataset.

		Train Data				
		Birth Control	Depression	Pain	Anxiety	Diabetes, Type 2
Test Data	Birth Control	92.11	61.41	57.36	58.02	59.04
	Depression	58.64	92.28	74.83	75.8	61.81
	Pain	63.24	74.71	91.62	79.57	50.57
	Anxiety	59.7	79.25	77.62	90.57	62.05
	Diabetes, Type 2	66.71	72.65	65.59	72.15	91.96

Table 4: Bigram Cross-domain Sentiment Analysis

		Train Data				
		Birth Control	Depression	Pain	Anxiety	Diabetes, Type 2
Test Data	Birth Control	80.3	56.18	54.48	56.17	50.39
	Depression	46.62	88.59	70.57	73.47	54.73
	Pain	53.05	65.14	88.76	77.24	47.43
	Anxiety	49.48	74.48	74.69	88.78	58.28
	Diabetes, Type 2	50.62	59.9	62.0	65.1	90.35

Table 5: Unigram Cross-domain Sentiment Analysis

		Train Data				
		Birth Control	Depression	Pain	Anxiety	Diabetes, Type 2
Test Data	Birth Control	92.05	60.58	56.59	56.47	59.65
	Depression	56.35	91.99	75.12	76.25	64.78
	Pain	59.57	76.76	91.9	79.86	52.67
	Anxiety	56.24	80.45	79.82	91.19	66.19
	Diabetes, Type 2	66.34	71.41	69.06	72.4	91.58

Table 6: Trigram Cross-domain Sentiment Analysis

Nella Table 7 sono riportati i risultati ottenuti mediante la classe Tf-idf. Come si può notare non differiscono di molto rispetto a quelli in Table 4 anzi, si può osservare un leggero peggioramento. Questo è probabilmente dovuto all'utilizzo dei termini medici in modo non adeguato da parte dei pazienti, come viene ampiamente descritto nell'articolo.

		Train Data				
		Birth Control	Depression	Pain	Anxiety	Diabetes, Type 2
Test Data	Birth Control	91.64	60.35	57.36	57.77	55.41
	Depression	57.42	91.89	73.86	75.61	59.29
	Pain	64.14	72.71	90.86	78.95	49.33
	Anxiety	58.23	79.51	77.36	91.35	58.96
	Diabetes, Type 2	66.21	70.42	67.82	73.64	91.21

Table 7: Tf-idf Cross-domain Sentiment Analysis