Intelligenza Artificiale
Analisi dei Sentimenti in Recensioni di Farmaci

Martina Buccioni

3 settembre 2023

Abstract

Questa relazione descrive il lavoro svolto nell'ambito del progetto di analisi dei sentimenti nelle recensioni di farmaci. L'obiettivo principale è stato esaminare le prestazioni dei modelli costruiti su dati di una condizione, ossia il dominio di origine, e valutarli su dati relativi ad altre condizioni, ossia il dominio di destinazione. Vengono presentati i dettagli della metodologia utilizzata, i risultati sperimentali e un'analisi critica dei risultati ottenuti.

1 Introduzione

L'analisi dei sentimenti in recensioni di farmaci è una parte fondamentale dell'elaborazione del linguaggio naturale (NLP) con applicazioni importanti in medicina, farmacovigilanza e raccolta di opinioni degli utenti. In questo progetto, abbiamo esaminato come i modelli addestrati su un dominio specifico influiscano sulle prestazioni quando applicati a domini diversi. In particolare, ci siamo concentrati sull'analisi di dominio (domain) e cross-domain.

2 Metodologia

2.1 Preprocessing dei Dati

I dati sono stati pre-processati inizialmente rimuovendo le righe con valori float dalla colonna "commentsReview" per garantire che solo le recensioni testuali fossero considerate. Successivamente, abbiamo applicato la tokenizzazione alle recensioni e rimosso le stopwords. Questa fase di preprocessing è stata eseguita su entrambi i dataset "DrugLib" e "DrugsCom".

2.2 Etichettatura dei Sentimenti

Le recensioni sono state etichettate in base al rating associato. Abbiamo etichettato come "negativa" le recensioni con un rating inferiore o uguale a 4, come "neutrale" quelle con un rating superiore a 4 ma inferiore a 7, e come "positiva" quelle con un rating maggiore o uguale a 7.

2.3 Estrazione delle Feature

Per rappresentare le recensioni in modo che potessero essere utilizzate come input per il modello, abbiamo utilizzato la tecnica della "Bag of Words" (BoW) con l'ausilio del CountVectorizer di scikit-learn. La rappresentazione BoW assegna un vettore a ciascuna recensione, dove ogni elemento del vettore corrisponde a una parola unica nel corpus di testo. Successivamente, abbiamo addestrato il modello Random Forest Classifier utilizzando le matrici BoW come input. Sono stati addestrati due modelli separati per i dataset "DrugLib" e "DrugsCom".

2.4 Random Forest

Nel nostro progetto, abbiamo scelto di utilizzare il Random Forest Classifier come modello di machine learning per l'analisi dei sentimenti nelle recensioni di farmaci. Il Random Forest è un modello basato su ensemble che combina i risultati di più alberi decisionali per migliorare la previsione e la generalizzazione. Questa scelta è stata motivata dalla sua robustezza, flessibilità e capacità di gestire sia dati numerici che testuali.

Abbiamo rappresentato le recensioni dei farmaci come vettori numerici utilizzando la tecnica della "Bag of Words" (BoW) in combinazione con il CountVectorizer di scikit-learn. Questo ci ha permesso di trasformare le recensioni testuali in dati strutturati che potessero essere utilizzati come input per il Random Forest Classifier.

3 Risultati Sperimentali

3.1 Analisi Cross-Dominio (Cross-Domain Analysis)

Nell'analisi di dominio, abbiamo valutato le prestazioni del modello addestrato su un dominio specifico (condizione) quando applicato a dati relativi a altre condizioni (domini di destinazione). I domini selezionati sono stati estratti da diverse discipline mediche e includono "Contraception", "Depression", "Pain", "Anxiety" e "Diabetes, Type 2". I risultati sono stati riportati in termini di accuratezza e Cohen's Kappa.

I risultati [Tabella 1] mostrano che il dominio di addestramento ha un impatto considerevole sulle prestazioni del classificatore quando applicato a dati di altri domini. L'addestramento e il test all'interno dello stesso dominio (in-domain) hanno chiaramente prestazioni superiori rispetto a tutte le configurazioni cross-domain. Questa scoperta sottolinea l'importanza del vocabolario specifico del dominio. Tuttavia, alcuni accoppiamenti di domini mostrano prestazioni migliori di altri, suggerendo coerenze sottostanti tra effetti collaterali o espressioni specifiche del dominio utilizzate dai pazienti. Nell'analisi cross-dominio, abbiamo valutato la trasferibilità dei modelli addestrati tra diverse fonti di dati. I modelli di soddisfazione complessiva dei pazienti sono stati addestrati su entrambi i set di dati di addestramento associati e valutati su recensioni di farmaci provenienti da un'altra fonte di dati indipendente. I risultati indicano che il trasferimento di un modello di sentiment addestrato su un set di dati significativamente più ampio può produrre risultati promettenti.

3.1.1 Tabella 1

	Birth Control	Depression	Pain	Anxiety	Diabetes, Type 2	Column Means
Birth Control	['Accuracy: 0.919051', "Cohen's Kappa: 0.859419"]	['Accuracy: 0.489176', "Cohen's Kappa: 0.138806"]	['Accuracy: 0.641905', "Cohen's Kappa: 0.194697"]	['Accuracy: 0.487421', "Cohen's Kappa: 0.117242"]	['Accuracy: 0.690594', "Cohen's Kappa: 0.352564"]	0.621559
Depression	['Accuracy: 0.550477', "Cohen's Kappa: 0.094754"]	['Accuracy: 0.907593', "Cohen's Kappa: 0.779572"]	['Accuracy: 0.780476', "Cohen's Kappa: 0.218655"]	['Accuracy: 0.785639', "Cohen's Kappa: 0.190575"]	['Accuracy: 0.706683', "Cohen's Kappa: 0.217505"]	0.6997738
Pain	['Accuracy: 0.522077', "Cohen's Kappa: 0.020739"]	['Accuracy: 0.705977', "Cohen's Kappa: 0.083465"]	['Accuracy: 0.896667', "Cohen's Kappa: 0.679238"]	['Accuracy: 0.769916', "Cohen's Kappa: 0.087551"]	['Accuracy: 0.668317', "Cohen's Kappa: 0.045157"]	0.7509526
Anxiety	['Accuracy: 0.522595', "Cohen's Kappa: 0.021403"]	['Accuracy: 0.710824', "Cohen's Kappa: 0.096146"]	['Accuracy: 0.77381', "Cohen's Kappa: 0.095927"]	['Accuracy: 0.898847', "Cohen's Kappa: 0.689079"]	['Accuracy: 0.684406', "Cohen's Kappa: 0.092854"]	0.7290354000000001
Diabetes, Type 2	['Accuracy: 0.593595', "Cohen's Kappa: 0.220115"]	['Accuracy: 0.685299', "Cohen's Kappa: 0.270467"]	['Accuracy: 0.661905', "Cohen's Kappa: 0.155624"]	['Accuracy: 0.703354', "Cohen's Kappa: 0.220188"]	['Accuracy: 0.935644', "Cohen's Kappa: 0.86195"]	0.7371288
Row Means	0.6456293999999999	0.7461736	0.7125908000000001	0.7180964000000001	0.7159594	nan

4 Conclusioni

In questo studio, abbiamo esaminato l'applicazione dell'analisi dei sentimenti basata su machine learning alle recensioni di farmaci generate dai pazienti. Abbiamo addestrato modelli di Random Forest utilizzando le recensioni preprocessate come dati di input e abbiamo esplorato la trasferibilità di tali modelli tra diversi domini e fonti di dati.

I risultati dei nostri esperimenti hanno evidenziato che l'addestramento e la valutazione in-domain, cioè all'interno dello stesso dominio, producono le prestazioni migliori. Questi risultati sono in linea con quelli riportati nella ricerca citata [1], che ha sottolineato l'importanza del vocabolario specifico del dominio nella classificazione delle recensioni dei farmaci. Tuttavia, abbiamo anche osservato che alcune combinazioni di domini mostrano prestazioni migliori rispetto ad altre, indicando la presenza di coerenze sottostanti tra gli effetti collaterali o le espressioni utilizzate dai pazienti nei diversi domini.

In conclusione, questo studio contribuisce a comprendere l'analisi dei sentimenti basata su machine learning nelle recensioni di farmaci e sottolinea l'importanza dell'addestramento in-domain per ottenere prestazioni ottimali. Sottolinea anche la necessità di ulteriori ricerche per sviluppare modelli più sofisticati e utilizzare set di dati più ampi per l'analisi dei sentimenti basata sugli aspetti. Questo lavoro ha il potenziale per promuovere la farmacovigilanza, lo sviluppo di sistemi di raccomandazione terapeutica e l'estrazione automatica di sentimenti legati agli aspetti delle recensioni dei pazienti sui farmaci.

References

[1] Felix Gräßer. "Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning". In: *Technische Universität Dresden* ().