DRUG REVIEW

Un'approccio basato su TF-IDF e Random Forest per la classificazione delle percezioni degli utenti

INTRODUZIONE

OBBIETTIVO DEL PROGETTO

Sviluppare un modello predittivo capace di classificare l'efficacia e la soddisfazione legate a specifici trattamenti farmacologici, basandosi sul linguaggio naturale espresso dagli utenti.

DATASET UTILIZZATO

Il "Drug Review Dataset" disponibile tramite il repository UCI Machine Learning, che combina dati testuali liberi (non strutturati) con variabili numeriche e categoriali (strutturate).

RECENSIONI

Linguaggio naturale e percezione utente

VALUTAZIONE

Scala da 1 - 10 di soddisfazione

CONDIZIONE

Patologia trattata

IL DATASET DRUG REVIEW

DATI NON STRUTTURATI

Recensione (review)
Contiene il linguaggio naturale e l'espressione diretta della percezione e dell'esperienza dell'utente, fondamentale per l'analisi del sentimento.

DATI STRUTTURATI

Valutazione Numerica (rating)
Variabile quantitativa (scala 1-10) che cattura la percezione complessiva di soddisfazione, fungendo da variabile target per la predizione.
Condizione (condition)
Il nome della patologia per cui il farmaco è stato utilizzato, fornendo un contesto medico essenziale.
Conteggio Utilità (usefulCount)
Il numero di utenti che hanno valutato positivamente l'utilità della recensione, aggiungendo una dimensione di validazione sociale.

PRE-ELABORAZIONE DEL TESTO

Prima di applicare algoritmi di analisi, il testo grezzo delle recensioni deve essere sottoposto a un'accurata fase di pulizia e normalizzazione per isolare i termini realmente significativi e ridurre la complessità del vocabolario.

TOKENIZZAZIONE

Il testo di ogni recensione viene scomposto in unità discrete, chiamate "token". Generalmente, questi token sono singole parole o sequenze di parole (n-grammi).

RIMOZIONE STOP WORDS

Vengono eliminate le parole comuni e funzionali della lingua (articoli, preposizioni, congiunzioni) che non aggiungono valore informativo alla classificazione.

LEMMATIZZAZIONE

Le parole vengono ridotte alla loro forma base o radice. Questo processo standardizza i termini, riduce la dimensionalità del vocabolario e migliora l'efficacia del modello.

ESEMPIO DI APPLICAZIONE

Testo originale: "Ho percepito un miglioramento significativo dopo aver preso il farmaco per la mia condizione."

Testo preprocessato: ["percepito", "miglioramento", "significativo", "farmaco", "condizione"]

VETTORIZZAZIONE TF-IDF

Cos'è il TF-IDF?
Il TF-IDF (Term Frequency-Inverse Document Frequency) è una tecnica statistica del Natural Language Processing, che risolve il problema di come dare un valore numerico all'importanza di una parola all'interno di un testo.

Come funziona?
Il punteggio TF-IDF è il prodotto di due metriche:

1. Frequenza del Termine (TF): Misura la rilevanza locale di una parola in un documento.
2. Frequenza Inversa del Documento (IDF): Misura l'importanza globale di una parola, penalizzando i termini comuni.

FORMULE MATEMATCHE

TF(t,d) = Frequenza di t in d / Numero totale di termini in d

IDF(t,D) = log(N / DF(t))
dove N è il numero di recensioni e DF(t) il numero di recensioni in cui appare t

TF-IDF(t,d,D) = TF(t,d) × IDF(t,D)

Matrice Termine-Documento
Applicando TF-IDF si genera una matrice:

RECENSIONI	FARMACO	DOLORE	EFFICACE
R1	0,2	0,5	0,8
R2	0,3	0,4	0,6

CLASSIFICAZIONE CON RANDOM FOREST

COSA È RANDOM FOREST?

Algoritmo di apprendimento supervisionato composto da un insieme di alberi decisionali per classificare i dati.

COME FUNZIONA

- Campionamento Bootstrap: Ogni albero lavora su un campione casuale del dataset
- Selezione Casuale: Considera una sotto-selezione casuale delle feature
- Voto di Maggioranza: Classe predetta quella con più voti

PROCESSO DI CLASSIFICAZIONE CON RANDOM FOREST

Ogni albero nel forest esprime il suo voto sulla classe

200 ALBERI

Configurazione con 200 alberi per una predizione robusta

CASUALITÀ

Meccanismi di casualità per prevenire l'overfitting

STIMA DELLA PERFORMANCE

Capacità di stimare l'accuratezza durante l'addestramento.

Bilanciamento

Migliora la capacità di classificare tutte le classi

PIPELINE DI MACHINE LEARNING

PREPARAZIONE DEI DATI

Dataset Drug Reviews suddiviso in Train (70%) e Test (30%) con stratificazione

VETTORIZZAZIONE TF-IDF

Trasformazione del testo in matrice termine-documento con pesi TF-IDF

CLASSIFICAZIONE RANDOM FOREST

Modello di classificazione con 200 alberi decisionali

BINARIZZAZIONE DEL RATING

- Classe Positiva (Efficace): Recensioni con rating ≥ 7
- Classe Negativa (Non Efficace/Neutrale): Recensioni con rating < 7

VANTAGGI DELLA PIPELINE

- Coerenza e riproducibilità nell'intero flusso di lavoro
- Gestione automatica delle trasformazioni dei dati
- Applicazione coerente di tutte le trasformazioni e modelli

RISULTATI E MATRICE DI CONFUSIONE

TN	FP	FN
Veri Negativi	Falsi Positivi	Falsi Negativi

ANALISI DEI RISULTATI

- I modello ha una buona capacità di identificare recensioni positive (573 previsioni corrette)
- Difficoltà nel classificare correttamente recensioni negative e neutre
- forte bias del modello verso la classe "positiva"

CLASSIFICAZIONE ERRATA

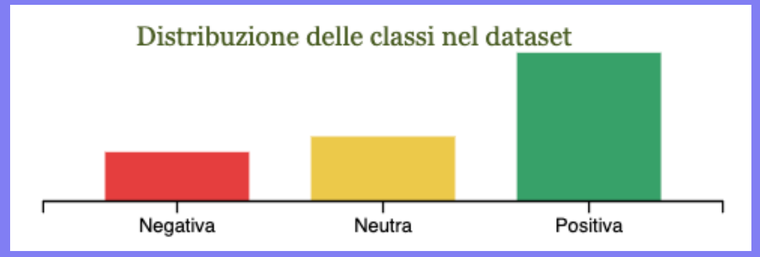
153
Recensioni negative classificate come positive

247
Recensioni neutre classificate come positive

CONCLUSIONI E SVILUPPI FUTURI

CONCLUSIONE DEL PROGETTO

- Dimostrata fattibilità di tradurre percezioni umane in linguaggio naturale in un modello matematico predittivo
- Integrazione di TF-IDF per vettorizzazione testo e Random Forest per classificazione efficace
- Limitazioni significative a causa di squilibrio nei dati che ha compromesso la capacità di identificare correttamente le classi minoritarie



PROPOSTE DI MIGLIORAMENTO

Pesatura delle Classi

Implementare class_weight='balanced' per assegnare peso maggiore agli errori sulle classi minoritarie

Oversampling

Utilizzare tecniche come SMOTE per generare campioni sintetici delle classi minoritarie

Modelli Più Sofisticati

Esplorare algoritmi avanzati come SVM o algoritmi di Boosting (Gradient Boosting, XGBoost)