Drug Review Rating Prediction - TF-IDF vs. MiniLM Embeddings

Abstract

This project explores automatic prediction of patient satisfaction scores from textual drug reviews. Two complementary text representation approaches were evaluated - traditional TF-IDF features and contextual transformer embeddings (MiniLM). Each was combined with a Logistic Regression classifier and evaluated on both (3-class sentiment) and fine-grained (10-class rating) prediction tasks.

1. Dataset Overview

The dataset originates from Drugs.com, containing over 200K patient reviews including drug name, condition, free-text review, and numerical rating (1–10). The dataset was divided into training (161,297) and testing (53,766) samples.

<table>
<tr><th>Train Samples</th><th>Test Samples</th><th>Unique Drugs</th><th>Unique Conditions</th><th>Average review length</th><th colspan="4">Rating</th></tr>
<tr><td rowspan="10">161,297</td><td rowspan="10">53,766</td><td rowspan="10">3,436</td><td rowspan="10">885</td><td rowspan="10">458.73</td><td>1</td><td>13.4%</td><td rowspan="4">$\text{rating} \leq 4$</td><td rowspan="4">24.85%</td></tr>
<tr><td>2</td><td>4.3%</td></tr>
<tr><td>3</td><td>4.03%</td></tr>
<tr><td>4</td><td>3.1%</td></tr>
<tr><td>5</td><td>4.96%</td><td rowspan="2">$4 < \text{rating} < 7$</td><td rowspan="2">8.9%</td></tr>
<tr><td>6</td><td>3.93%</td></tr>
<tr><td>7</td><td>5.9%</td><td rowspan="4">$7 \leq \text{rating} \leq 10$</td><td rowspan="4">66.25%</td></tr>
<tr><td>8</td><td>11.71%</td></tr>
<tr><td>9</td><td>17.06%</td></tr>
<tr><td>10</td><td>31.61%</td></tr>
</table>

2. Problem Definition

The objective is to predict patient rating from review text. We consider two tasks: (1) 3-class sentiment prediction (Negative ≤4, Neutral 5–6, Positive ≥7) and (2) 10-class fine-grained rating prediction (1–10).

3. Methodology

The process included several phases:

- Preprocessing: HTML cleaning, punctuation removal, stopword filtering, and lemmatization.
- Text construction: concatenating review + drug name + condition to provide context.
- Feature Representation: TF-IDF bigrams and SentenceTransformer MiniLM embeddings (384 dimensions).
- Modeling: Logistic Regression with GridSearchCV (3-fold CV, $C \in \{0.5, 1, 5, 10\}$).
  - Evaluation: Accuracy, Macro-F1, Weighted-F1.

4. Results and Evaluation

The table below summarizes model performance across validation and test sets:

Model	Task	Val. Accuracy	Test Accuracy	Test Macro-F1	Test Weighted-F1
TF-IDF + LR	3-class	0.872	0.874	0.787	0.874
TF-IDF + LR	10-class	0.654	0.653	0.602	0.653
MiniLM + LR	3-class	0.746	0.742	0.477	0.700
MiniLM + LR	10-class	0.394	0.389	0.162	0.308

TF-IDF clearly outperformed MiniLM on both 3-class and 10-class tasks, showing higher accuracy and better balance across classes.

5. Feature Analysis

TF-IDF model interpretability reveals sentiment-aligned n-grams:

- Negative (≤4): 'not recommend', 'never again', 'waste of', 'no relief'.
- Neutral (5–6): 'works okay', 'helps but', 'however it', 'some improvement'.
- Positive (≥7): 'love it', 'highly recommend', 'life saver', 'amazing', 'the best'.

6. Conclusions

- The TF-IDF + Logistic Regression approach achieved superior results in terms of accuracy, interpretability, and stability compared to transformer-based embeddings.

- As anticipated, the 3-class sentiment classification (negative / neutral / positive) performed significantly better than the 10-class rating task.
- The simpler 3-class formulation aligns well with how patients express general satisfaction levels, while predicting exact numeric ratings (1–10) introduces higher variability and ambiguity.
- Reviews with neutral ratings (5–6) presented the greatest challenge - they often contain mixed emotional tones, leading the model to misclassify them toward neighboring classes.
- This misclassification pattern reflects the linguistic overlap between slightly positive and slightly negative sentiments in natural language.
- Overall, the TF-IDF model effectively captured clear lexical sentiment markers (e.g., “love it”, “not recommend”) and provided a robust and interpretable baseline for healthcare-related text mining.
- While transformer embeddings (MiniLM) did not outperform TF-IDF in this setup, it is likely that embeddings could yield stronger results when combined with more complex classifiers **-** such as neural architectures (e.g., LSTM, CNN, or fine-tuned transformer heads) that can better leverage the semantic information encoded in the embeddings.
- The transformer-based MiniLM model didn’t perform as well because it was designed to capture general sentence meaning (semantic similarity) rather than emotional tone or sentiment. As a result, it struggled to accurately recognize neutral reviews (ratings 5–6), which often contain mixed or balanced opinions.
- In general, MiniLM tended to compress sentiment intensity toward the extremes, confirming that it was not optimized for fine-grained rating prediction tasks.

7. Future Work

- Experiment with more complex models such as neural networks (e.g., LSTM, CNN, or fine-tuned transformer-based classifiers) that can better capture contextual and emotional nuances in the text.
- Use a transformer model fine-tuned specifically for sentiment or healthcare-related reviews, to better interpret expressions of pain, frustration, and satisfaction.
- Explore alternative feature representations, for example by combining TF-IDF with contextual embeddings or using dimensionality reduction techniques (e.g., PCA, UMAP) to visualize sentiment space.
- Improve neutral class handling through targeted data augmentation or class-weight optimization to address class imbalance.
- Enhance interpretability with feature attribution tools (e.g., SHAP, LIME) to better understand which words or phrases influence predictions.