Sentiment Classification of Drug Reviews: Traditional vs. Bio Clinical BERT Approaches

Project Description

This project explores how sentiment can be classified from patient-written drug reviews, with the goal of identifying whether a patient's experience was positive, neutral, or negative based on their descriptions of effectiveness and side effects. Examining these personal accounts provides meaningful insight into patient satisfaction, potential adverse reactions, and patterns that support real-world pharmacovigilance beyond traditional clinical data. The main objective was to compare classical machine learning models with a domain-specific transformer, Bio ClinicalBERT, to see which approach more accurately interprets medical language and contextual meaning within patient narratives.

Data Description

The dataset used in this project is the DrugLib Corpus, available through the UCI Machine Learning Repository. It contains patient-authored drug reviews originally collected from Drugs.com and DrugLib.com. Each record includes both structured information about the medication and free-text descriptions of the user's experience.

Each entry contains the following fields:

- urlDrugName – the name of the drug reviewed
- rating – the patient's numerical satisfaction score (ranging from 1 to 10)
- effectiveness, sideEffects, condition – structured metadata describing the reported outcome and use case
- benefitsReview, sideEffectsReview, commentsReview – narrative text fields capturing the patient's detailed account of their experience

All three narrative fields (benefitsReview, sideEffectsReview, and commentsReview) were merged into a single column called *text*, and the numerical ratings were converted into categorical sentiment labels:

- 1–3 → Negative
- 4–6 → Neutral
- 7–10 → Positive

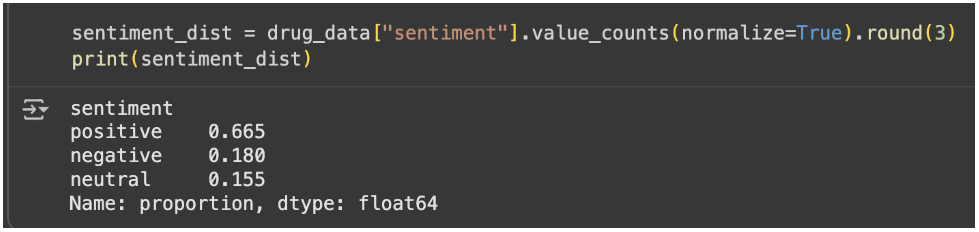
The cleaned and merged dataset served as the foundation for training and evaluating both the classical and transformer-based sentiment classification models.

Table 1. Sentiment Label Mapping

Rating Range	**Sentiment Label**
1 -3	Negative
4 - 6	Neutral
7 - 10	Positive

After cleaning, the dataset contained ≈ 4142 reviews.

Figure 1. *Sentiment Distribution after Preprocessing*



```
sentiment_dist = drug_data["sentiment"].value_counts(normalize=True).round(3)
print(sentiment_dist)
```

```
sentiment
positive    0.665
negative    0.180
neutral     0.155
Name: proportion, dtype: float64
```

Figure 1. *This is a bit of an imbalance as the positive sentiment far outweighs the others.*

Methods and Analysis

(i) Models Evaluated

Three models were used to classify sentiment in patient drug reviews. The first was a TFIDF-based Logistic Regression model, chosen for its simplicity and interpretability as a linear baseline. The second was a TFIDF-based Linear Support Vector Machine (SVM), a strong classical classifier known for handling high-dimensional text features effectively. The third model was Bio Clinical BERT, a transformer model fine-tuned on biomedical and clinical text. It represents a more advanced, context-aware approach designed to capture meaning beyond individual word frequencies.

(ii) Rationale

The classical classification models provided transparent benchmarks that depend on word frequency and direct lexical cues such as *"good," "bad,"* or *"effective."* They are efficient and easy to interpret, which makes them useful starting points for sentiment tasks. In contrast, ClinicalBERT brings contextual understanding to the table. Because it was pretrained on clinical notes and PubMed abstracts, it recognizes how medical terms and emotional tone interact in patient language. This allows it to interpret complex statements such as *"the medication worked but caused severe nausea,"* which traditional models would likely misclassify.

(iii) Variables and Features

The primary input feature was the combined review text, consolidated into a single column named *text.* The target variable was the sentiment label, categorized as *negative, neutral,* or *positive.* Numeric fields such as *rating* were not used directly in training but served as the basis for mapping the original reviews into these sentiment categories.

(iv) Model Training

Table 2. Models Training Architecture

Model	**Representation**	**Key Parameters**
Logistic Regression	TF-IDF (1–2 grams)	max_iter = 1000
Linear SVC	TF-IDF (1–2 grams)	C = 1 (default)
Bio Clinical BERT	Transformer embeddings	LR = 2e-5, epochs = 3, batch = 16

Table 2. Training used an 80/20 train-validation split with stratification across sentiment classes. Evaluation metrics included Accuracy, Macro Recall, and Macro F1 to capture performance.

Results and Visualizations

(a) Quantitative Performance

Table 3. Model Performance Metrics

Model	**Accuracy**	**Macro Recall**	**Macro F1**
Clinical BERT	0.768	0.590	0.574
TF-IDF + Linear SVC	0.734	0.512	0.520
TF-IDF + Logistic Regression	0.705	0.408	0.395

Table 3. *Performance Comparison across Models.*

(b) Confusion Matrices

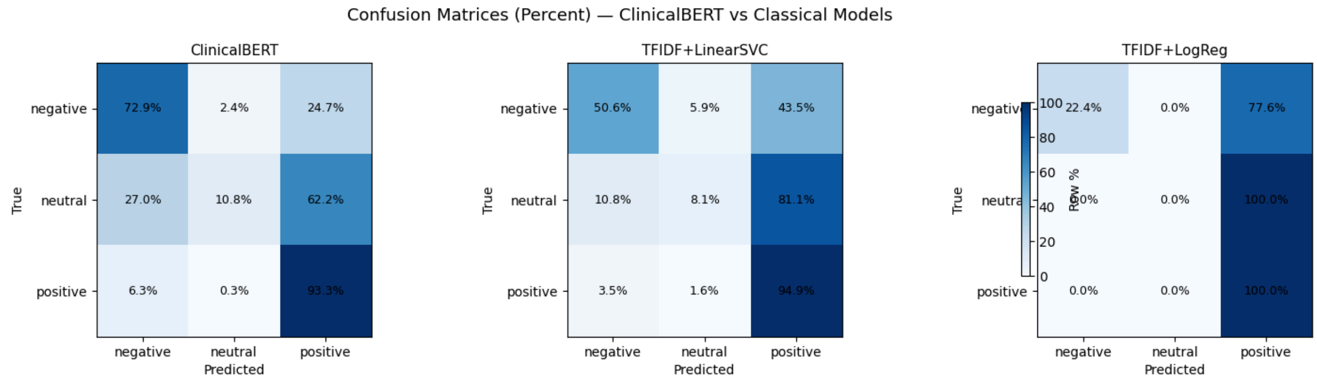


Figure 2. Confusion Matrices (Percentages)

Each matrix displays row-normalized percentages with TN/FP/FN/TP breakdown. Clinical BERT shows higher true-positive accuracy for both *positive* and *negative* classes, while *neutral* remains hardest to classify.

Interpretation

ClinicalBERT demonstrated the clearest and most balanced understanding of sentiment across all three categories. It correctly identified most positive and negative reviews while showing better control over mixed or contradictory phrasing, such as *"the medication worked but left me feeling drained."* Its strength came from reading language in full context rather than relying on isolated words, allowing it to weigh both the benefits and side effects described in a single statement.

The Linear SVM performed competitively for clearly polarized text, especially for strongly positive reviews, where it maintained high precision. However, it tended to misclassify more subtle or balanced phrasing, often shifting neutral or mildly negative reviews toward the positive category. This suggests that while it captures surface-level sentiment cues, it lacks the deeper contextual awareness needed for clinical text.

Logistic Regression showed the weakest differentiation. It overemphasized the positive class and failed to capture neutral or negative sentiment effectively. This likely occurred because the model's linear boundaries and reliance on frequent word patterns limited its ability to interpret sentiment that depends on tone or phrasing rather than individual terms.

Across all models, the neutral class remained the most difficult to classify. Many reviews contained mildly positive language even when overall sentiment was neutral, blurring the distinction between categories. This pattern suggests that additional fine-tuning or data balancing would help improve recall for more ambiguous expressions of patient experience.

Evaluation

ClinicalBERT achieved the highest overall performance, with a macro F1 score of 0.574. This result shows that it maintained a stronger balance across all three sentiment categories compared to the classical models. The TFIDF classifiers, while simpler and more computationally efficient, relied heavily on surface-level word cues and could not fully interpret how tone and meaning shift across sentences. They served as effective baselines for comparison but lacked the contextual understanding that ClinicalBERT demonstrated.

When reviewing the per-class results, the confusion matrices confirmed several consistent patterns. Positive reviews were the easiest to identify and made up the majority of the dataset, introducing a mild class imbalance. Neutral reviews were frequently misclassified as positive due to overlapping language—many patients used mildly positive wording even when their experiences were mixed. Negative reviews were occasionally labeled as neutral, especially when the tone was softened by phrases like *"helped a little"* or *"worked but not enough."* These trends highlight the need for better handling of ambiguous phrasing and underrepresented sentiment classes in future model refinements.

Discussion and Conclusion

Between the classification models and ClinicalBERT, the transformer model proved to be the most consistent and reliable. The TFIDF classifiers using Logistic Regression and Linear SVM provided a strong foundation and performed well on straightforward reviews, particularly when the sentiment was clearly positive or negative. However, they struggled when patients expressed mixed experiences, such as *"the medication worked but made me too tired to function."* Because these models rely on individual word counts rather than sentence meaning, they could not fully interpret how tone shifts within a review.

ClinicalBERT managed those complex cases much more effectively. Its transformer-based design allowed it to read the entire context of a review instead of weighing words in isolation. Since it was trained on biomedical and clinical language, it was better at connecting medical terms to sentiment, understanding that phrases like *"effective but caused nausea"* carry both positive and negative emotion. This contextual understanding helped it achieve stronger balance across all sentiment categories and handle subtle or contradictory phrasing more reliably.

Among the classification models, the Linear SVM performed the best, maintaining solid accuracy when the sentiment was clear. Logistic Regression remained useful for its interpretability and simplicity but showed less sensitivity to context. Together, these results illustrated a clear tradeoff between simplicity and depth. The TFIDF classifiers were easier to interpret, while ClinicalBERT demonstrated a deeper grasp of how patients actually communicate about their experiences.

The neutral sentiment class remained the most challenging across all models. Many reviews used slightly positive or negative wording even when the overall experience was neutral, which made it difficult for any model to separate subtle emotional cues. Improving this may require class-weighted training, longer sequence lengths, or additional fine-tuning to better capture longer, descriptive statements.

Looking forward, an important next step would be to apply aspect-based sentiment analysis, allowing the model to analyze specific parts of a review such as effectiveness, side effects, and satisfaction. This would make predictions more detailed and clinically useful. Overall, while the TFIDF classifiers provided a clear and interpretable benchmark, ClinicalBERT showed a meaningful improvement in understanding patient sentiment, capturing nuance, and transforming patient feedback into practical insights for healthcare and research.

References

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W. H., Jin, D., Naumann, T., & McDermott, M. (2019). *Publicly available clinical BERT embeddings. Proceedings of the 2nd Clinical Natural Language Processing Workshop (ClinicalNLP 2019)*, 72–78. https://doi.org/10.18653/v1/W19-1909

Dua, D., & Graff, C. (2019). *UCI Machine Learning Repository: Drug Review Dataset (Drugs.com and DrugLib.com).* University of California, Irvine, School of Information and Computer Sciences. https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+(Drugs.com)

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). *Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12*, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., … Rush, A. M. (2020). *Transformers: State-of-the-art natural language processing. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6