# Model Performance Comparison Dashboard

## Overview

This dashboard presents a detailed comparison of multiple NLP models tested on the pre-operative radiology report dataset, evaluating their ability to predict surgical outcomes.

The comparison covers key classification metrics including Accuracy, F1 Score, Recall, Precision, AUROC, and confusion matrix components.

These visual insights help identify the best-performing configurations and highlight the strengths and limitations of each approach in handling class imbalance and improving predictive accuracy.

## Insights

1. RoBERTa with Data Augmentation delivered the highest performance, achieving ~55% Accuracy and F1 Score ~0.69, outperforming all other configurations.
2. GatorTron with Class Weighting achieved moderate F1 Scores (~0.52) but lower accuracy (~46%), suggesting challenges in balancing correct classification across both classes.
3. Baseline RoBERTa (No Augmentation) underperformed, with Accuracy ~44% and F1 Score ~0.52, indicating scope for improvement through targeted data strategies.
4. Data Augmentation significantly boosted recall for RoBERTa (~98%), improving sensitivity to the "Operated" class, but at the cost of more false positives due to moderate precision (~54%).
5. GatorTron variants exhibited high recall (86–88%) but low precision (~37–38%), leading to frequent false alarms compared to RoBERTa.
6. Across all models, Precision remained lower than Recall, indicating a consistent trade-off toward sensitivity over specificity in surgical outcome prediction.

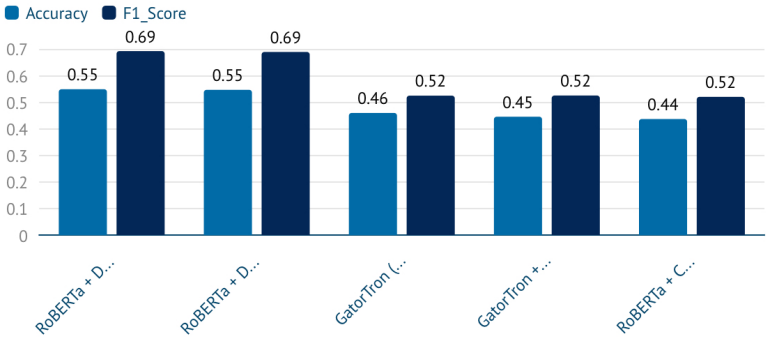
🏆 **Best Accuracy**
**0.55**

⭐ **Best F1 Score**
**0.69**

🔍 **Best Recall**
**0.98**

🎯 **Best Precision**
**0.54**

## Model Accuracy and F1 Score Comparison



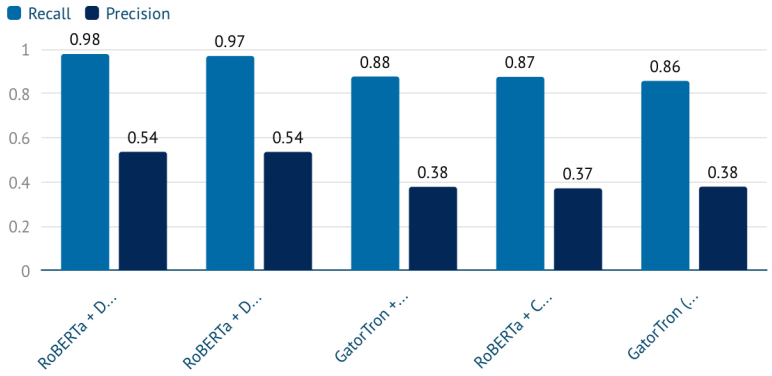

## Key Insights

- RoBERTa models with data augmentation lead in both accuracy (~0.55) and F1 score (~0.69), outperforming other variants
- GatorTron with class weighting achieves moderate F1 (~0.52) but lower accuracy (~0.46), indicating imbalance in predictions
- Baseline RoBERTa shows the lowest metrics (Accuracy ~0.44, F1 ~0.52), suggesting room for improvement without augmentation

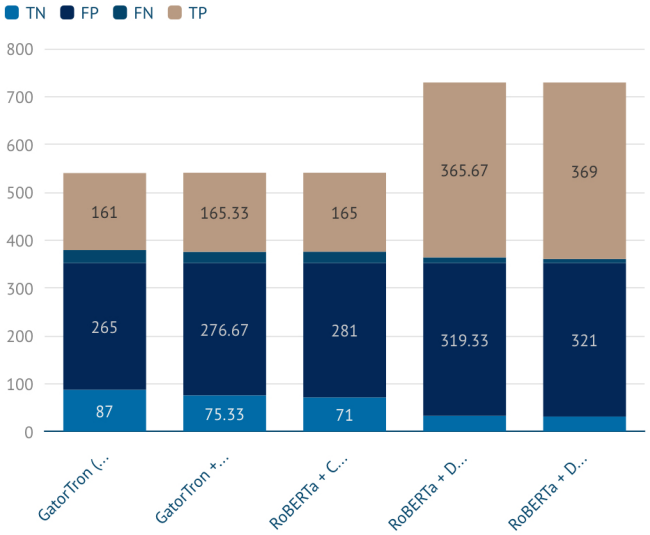
## Key Insights

- RoBERTa + Data Augmentation achieves very high recall (~0.98) but moderate precision (~0.54), indicating strong sensitivity at the expense of false positives
- GatorTron variants exhibit both lower recall (~0.86–0.88) and precision (~0.37–0.38), reflecting weaker overall classification balance
- Across all models, precision lags significantly behind recall, highlighting a consistent trade-off toward sensitivity over specificity

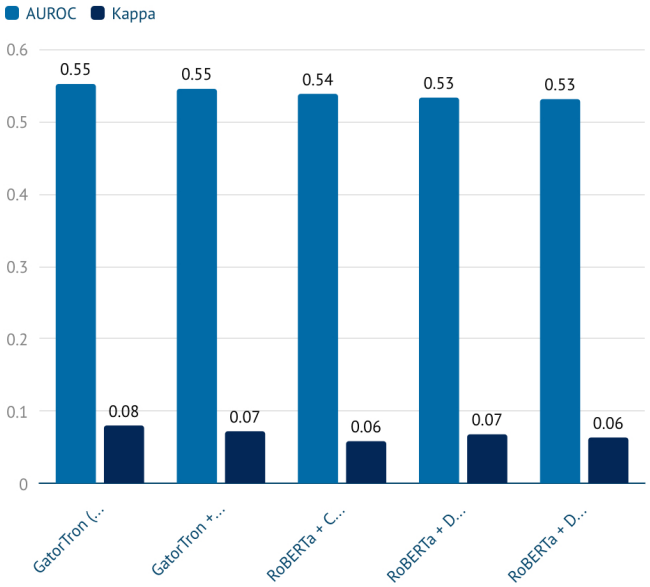
## Recall vs Precision by Model




## Confusion Matrix Components per Model




## AUROC and Kappa by Model




## Top Metrics Summary

| Model | F1_Score | Accuracy | Recall | Precision |
|---|---|---|---|---|
| RoBERTa + Data Augmentation | 0.6917 | 0.5485 | 0.9779 | 0.5355 |
| RoBERTa + Data Augmentation + Weight Decay | 0.6886 | 0.5462 | 0.9691 | 0.5351 |
| GatorTron + Class Weighting | 0.5249 | 0.4452 | 0.8762 | 0.3771 |