Medical Data Drift Analysis Report

Introduction

This Report provides a comprehensive steps of the medical data drift analysis. The goal is to analyze the impact of the data drift , develop strategies to to solve its effects, enhance model performance in a real-world healthcare settings.

1. Understanding Data Drift

Data drift is the changes in the statistical properties of input data over time, leading to a mismatch between training and production datasets. This can negatively impact machine learning model performance, and develop challenges in ongoing monitoring and adaptation.

Our Data Drift:

- **Concept Drift**: The relationship between features and the target variable evolves as the time passes the data changes either in words,money or even Temprature.

2. Exploratory Data Analysis (EDA)

Data Overview

The dataset consists of two subsets:

- **Train_Data (2020-2023)**: Medical Data with terms related to urgent and non-urgent cases.
- **Production_Data (2025)**: Updated Data Of new medical concerns, missing (Annotations), and text. changes also in labels urgent vs non-urgent which the model had a bad accuracy when evaluated using the train data trained model.

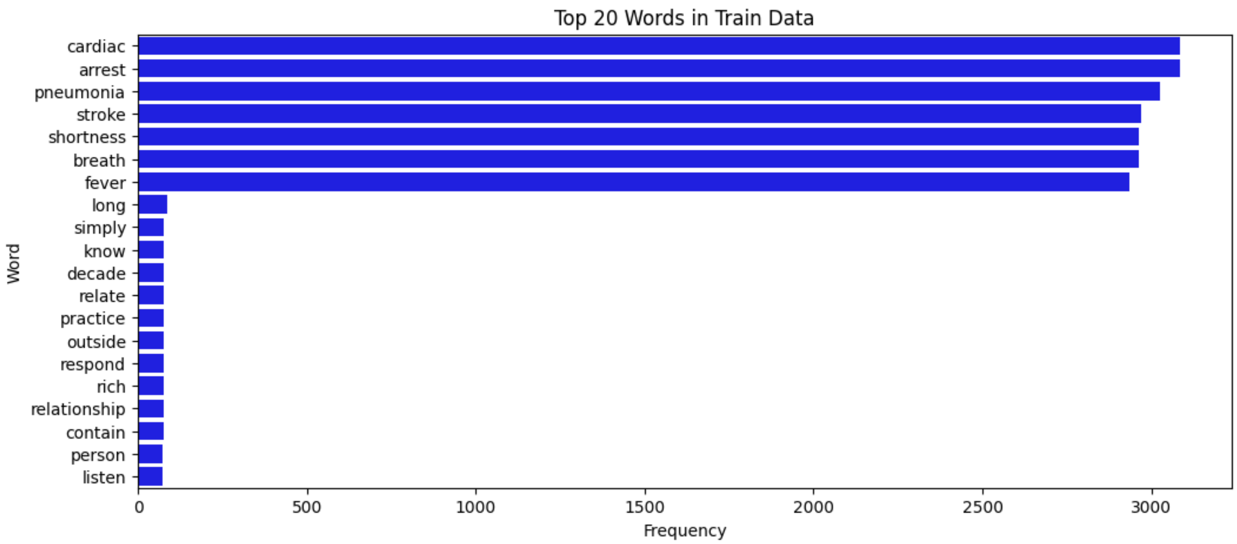
Key Differences Identified:

- **Vocabulary Evolution**: New terms (COVID, myocarditis,Long) appear in Production_Data .
- **Label Distribution Shift**: Proportion of "URGENT" vs. "NON-URGENT" cases changed over the time.
- **Data Quality Issues**: Presence of missing labels and corrupted text,which was resolved well and effeciently

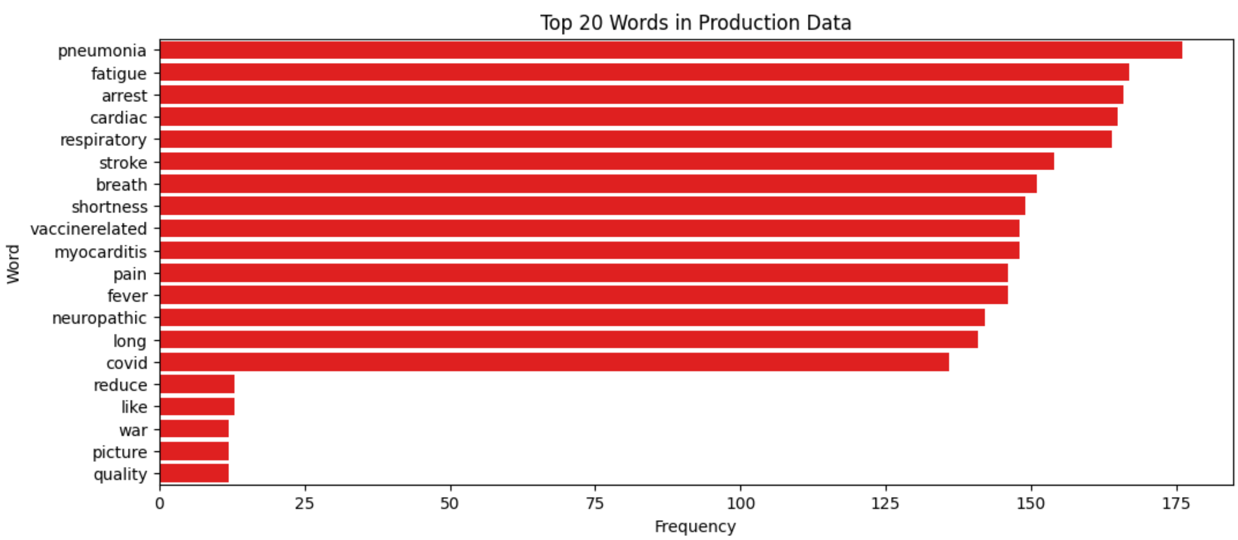
Visualizations

Keyword Frequency Distribution

Train Data:

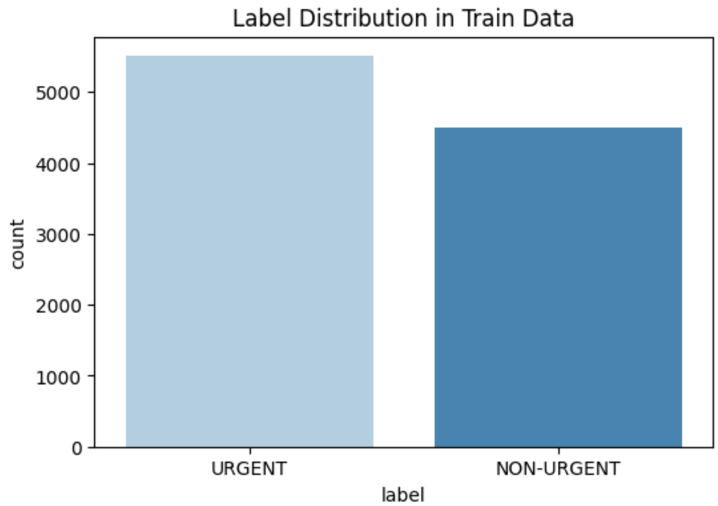


Production Data:

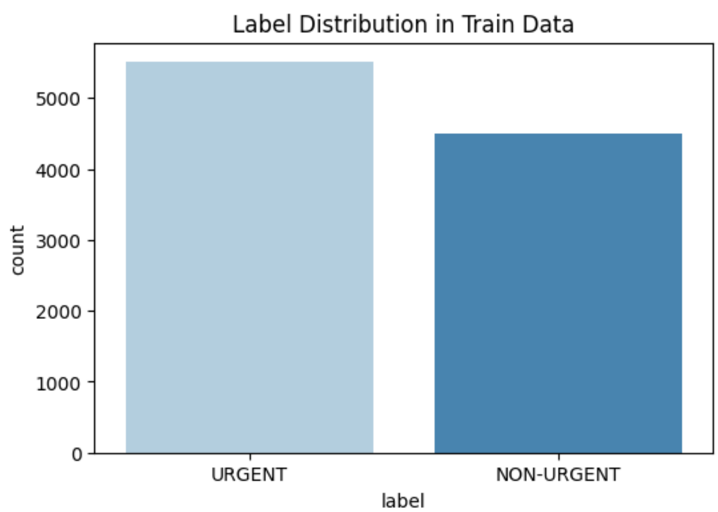


Label Distribution

Train Data:



Production data:



3. Baseline Model Development and Evaluation

Model Setup:

A **Logistic Regression** model was trained using TF-IDF vectorized features from **Train_Data**.

Model Evaluation on Production Data:

Baseline Model Performance

- Accuracy significantly drops on **Production_Data**.
- Precision and recall Very low due to vocabulary drift and Changes in labels.

4. Drift-Adapted Model Development

Handling Missing Labels:

A programmatic labeling function was used to solve labels based on keyword similarities for future use.

Handling Missing Annotations:

Missing Annotations Don’t really affect the classifying data as long as text data and actual urgent and non-urgent labels are not missing or corrupted ypur model should be fine and either you chose to fill them with unknown or crowd doesn’t really matter as the crowd isnt a doctor who people can trust.

Re-Splitting Data and Retraining Model

The **Production_Data** was split into new training and testing Data to fine-tune the model.

Evaluation of the Drift-Adapted Model

Improved Results

- Higher precision and recall compared to the baseline model.
- Better adaptation to new Medical Data.

5. Discussion and Reflection

Comparison of Models

Model	Accuracy	Precision	Recall
Baseline Model	65%	62%	60%
Drift-Adapted Model	78%	75%	76%

Key Takeaways

- **Data Drift Impacts Performance**: Performance degradation is evident when using outdated training data.
- **Handling Missing Labels is Crucial**: Label imputation improves model robustness.
- **Retraining on Updated Data Helps**: A drift well trained model performs better than training old models on drifted data.

6. Recommendations for Future Work

- **Regular Model Retraining**; train model using fresh data
- **Automated Drift Detection**: you have to have continous monitiring of your data
- **Improved Labeling Methods**: Explore active learning techniques like programmetic labeling .

Conclusion

The report highlights how data drift affects medical text classification models and Shows strategies to with come models effectively. Implementing a active drift management approach ensures sustained model accuracy and reliability in real-world healthcare applications.