LangTriage

Classification of Patient Free-Text Symptom Descriptions by Urgency

Final Presentation

Team members:

Nofar Kedmi

Diana Akoshvili

Project Description

CHALLENGE:

Hospitals, especially emergency rooms, are overwhelmed with patients presenting a wide variety of medical complaints, often using informal language and vague phrasing.

Triage decisions must be made quickly to prioritize care and ensure that urgent cases are treated first — but this manual process is time-consuming and can lead to delays.

OUR SOLUTION:

Developing an NLP model based on a Large Language Model (LLM) to classify patient symptom descriptions into:

0 — **Not urgent**: Can wait or be monitored
1 — **Urgent**: Requires immediate medical attention

GOAL:

Support faster and more accurate triage decisions.

Project Objectives

Exploring Different Aspects and Solutions to Our Problem

- Assess the impact of different text representations methods - classical vectorization (TF-IDF) versus contextual embedding (transformers) - on classification accuracy.
- Fine-tune transformer-based models like **DistilBERT** and **T5** alongside a traditional **TF-IDF + Logistic Regression** baseline, and compare their performance.
- Analyze the models' ability to handle imbalanced datasets and noisy clinical narratives.

Task Specification

- **Input:** Natural text generated from structured clinical measurements.
- **Output**: Classification labels indicating urgency level (0 or 1).
- **Metrics:** Classification Accuracy, F1-Score, Recall, Precision.

From Clinical Indicators to Urgency Prediction - Subtasks:

- **Data Generation/Labeling**: Generate a labeled dataset by converting structured clinical measurements into patient natural language free-text descriptions using GPT-4.
  These descriptions will serve as inputs for the models.
- **Training**: Fine-tune transformer-based models such as DistilBERT and generative T5 on the generated dataset, and train a traditional baseline model using TF-IDF with Logistic Regression to classify patient urgency levels accurately.
- **Evaluation**: Evaluate and compare the performance of the fine-tuned models against the baseline.

Prior Art

Source/Title	Machine learning–based triage to identify low-severity patients with a short discharge length of stay in emergency department, 2022	Development of an Anticipatory Triage-Ranking Algorithm Using Dynamic Simulation of Patients With Trauma	Improving ED Emergency Severity Index Acuity Assignment Using Machine Learning and Clinical Natural Language Processing, 2021
Task solved	Early identification of low-acuity emergency department (ED) patients likely to have short discharge times, in order to optimize ED workflow.	Instead of static triage labels, this study ranks trauma patients based on simulated deterioration over time to prioritize treatment more accurately.	Improve the accuracy of Emergency Severity Index (ESI) level assignments using EHR data and clinical text processing.
Approach/Model	CatBoost, Logistic Regression, Random Forest	LIFE Priority algorithm(A scoring system for trauma patients to estimate severity and survival probability)	KATE triage model (combines: Machine Learning classifiers, Clinical NLP pipelines to process free-text symptoms and clinical notes)
Data	Triage records collected from two hospitals in Taiwan	An artificial database that included 82,277 patients with trauma injuries.	Electronic medical records from two US hospitals
Metrics	AUC, Precision	Comparison with START algorithm (A widely used triage method in mass casualty situations)	Accuracy
Results	AUC = 0.755, Precision = 88.7%	LIFE improves the identification of clinical urgency compared to existing algorithms.	75.7% Accuracy

Methodology - Data Preparation

Source dataset description:

Patient Priority Classification dataset, sourced from Kaggle.

The dataset contains clinical measurements and triage labels for patients arriving the ER.

For example: BMI, Blood pressure, Insulin, Glucose, Cholesterol, Heart disease, Exercise angina, Chest pain.

Data generation:

Structured Prompt was designed for the GPT-4 model to generate natural, first-person, free-text descriptions without medical terms, simulating how a patient might describe their symptoms upon arrival at the emergency room. These were based on the patient's clinical measurements.

New Fields:

'text_input' column: Contains synthesized patient complaint texts generated by GPT-4, used as model input.

'label': binary urgency labels (0 = Not Urgent, 1 = Urgent).

Methodology - Data Preparation

Data properties:

- 6962 patient records with 17 structured clinical features.
- 'label' distribution after SMOTE: 4,861 samples for Not Urgent (0) and 4,861 samples for Urgent (1).
- Average input length (patient description): 35 words.

Patient case examples:

“I'm a 54-year-old male. I feel a tight feeling in my chest and kind of dizzy, it's weird. I can't keep up with simple tasks around the house, it's like my energy is gone".

"I’m 34. Since this morning my heart has been beating really fast. It gets worse when I move, but I still feel it even when I sit. I’m tired all the time, even when I’m not doing anything, and it scares me."

Models & processing pipeline

Model	Data Split	Epochs	Learning Rate	Batch Size (Train/Eval)	CPU/GPU	Input Max Length
TF-IDF + LR	80/20 stratified	No epoch-based training	Not applicable (solver)	None	CPU	Not applicable
DistilBERT (fine-tuned)	80/20 stratified	3	0.00002	16 / 32	GPU (Tesla T4)	128
T5 (fine-tuned)	80/20 stratified	3	0.00005	16 / 16	GPU (Tesla T4)	128

*All models were executed in the Google Colab environment.

Pipeline

Structured Patient Data

Demographics, Vital Signs, Lab Values, Pain Scale

↓

Data Cleansing & Map Labels to binary format

Treat missing/invalid data. (Urgent = 1, Not Urgent = 0)

↓

Prompt Creation

"You are patient arriving to the ER. Describe how you feel, use simple, natural language without medical terms"

↓

Tokenization & Vectorization

↓

Urgency Prediction

Models Output Binary classification: (0 or 1)

Metrics

How the metrics are computed ?

During Training (Validation Set) 🔬

- Loss Function
- Accuracy

During Evaluation (Test Set) 📊

- Accuracy
- Precision
- Recall
- F1-score
- AUROC

Code Organization

[Link to GitHub]

Data files: (Format: CSV)

- **patient_priority -** Raw dataset as provided, includes triage info and clinical indicators
- **df_clean_original** - Cleaned version with mapped binary labels
- **df_clean_with_text** - Final dataset including structured input and free-text complaint

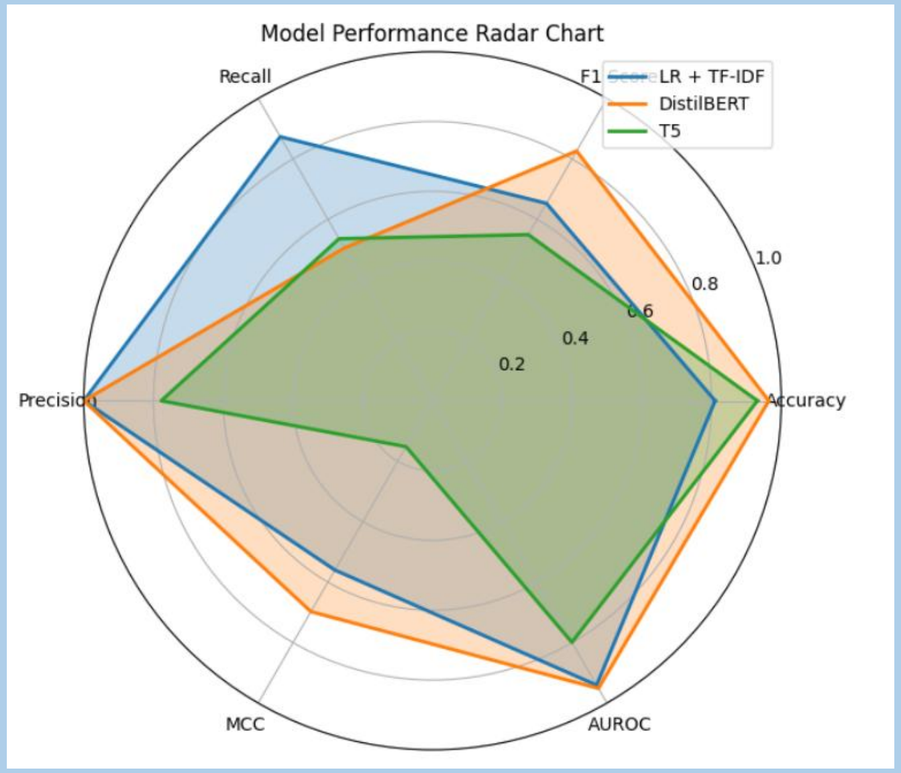
Code Notebooks:

- **EDA:** NLP_Project_EDA.ipynb
- **Free-Text Generation:** Generate_free_text_complaints.ipynb
- **Training:** Train_and_Evaluate_models.ipynb
- **Model Comparison:** Comparison_of_Model_Performance.ipynb

Results

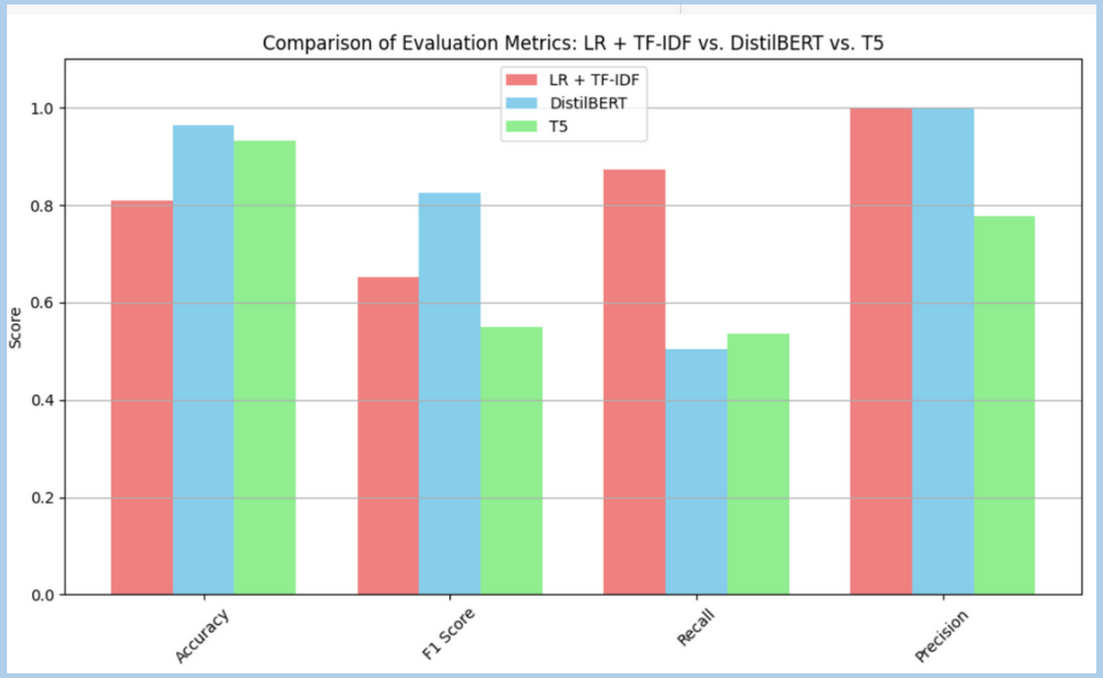
Comparison of model performance on the test set:

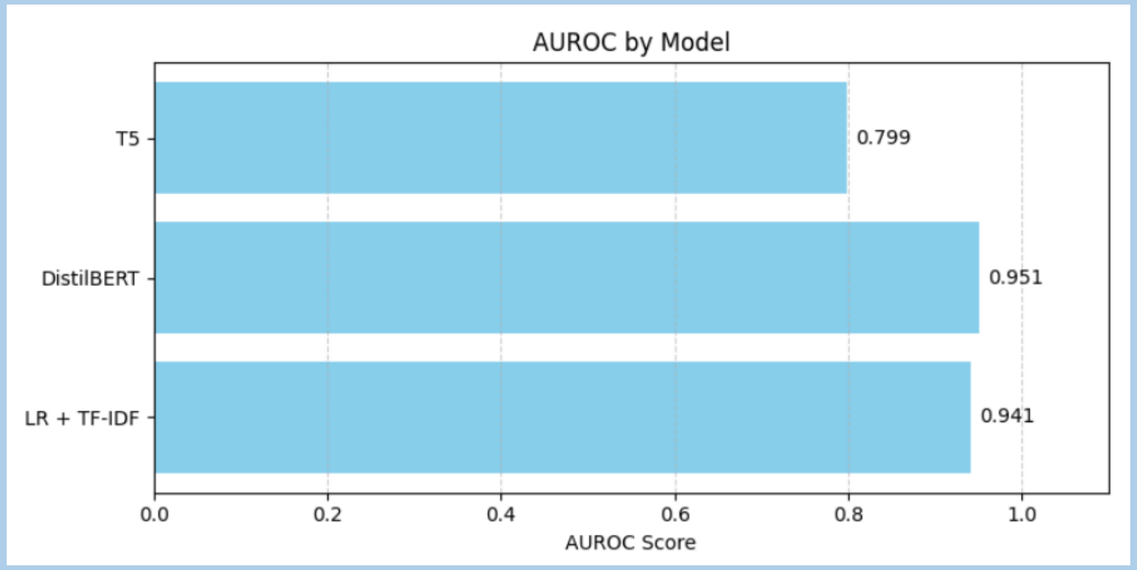
Metric/Model	TF-IDF + LR	DistilBERT	T5
Accuracy	0.8101	0.9641	0.9314
F1-Score	0.6530	0.8262	0.5494
Recall	0.8734	0.5939	0.5360
Precision	0.9999	0.9999	0.7778



Results

- DistilBERT demonstrated **the most balanced and superior performance across all metrics**, achieving the highest AUROC and F1-score.
- TF-IDF + Logistic Regression showed strong recall, indicating high sensitivity to urgent cases, but with lower precision and overall accuracy.
- T5 achieved good overall accuracy, but demonstrated weaker consistency in classification metrics such as F1-score and recall.





Conclusions & Recommendations for Future Work

Improving Recall in transformer models:
Recall was relatively low. Future work can explore threshold tuning or smarter sampling to improve urgent case detection.

Enhancing synthetic text generation: Texts can be made more natural and diverse using larger models or clinical fine-tuning.

Real-world clinical validation: Test the model in actual ER settings and gather staff feedback on its usefulness in triage decisions.

Visual Abstract

LangTriage

Classification of Patient Free-Text Symptom Descriptions by Urgency

Background

- Emergency rooms overloaded with patients presenting a wide variety of medical complaints.
- Often using informal language and vague phrasing.
- Prioritizing care and making decisions takes time.

Goal: Support faster & accurate triage.

Models Compared

- TF-IDF + Logistic regression
- DistilBERT
- T5

⇩

Tokenization & Vectorization
Train & Test

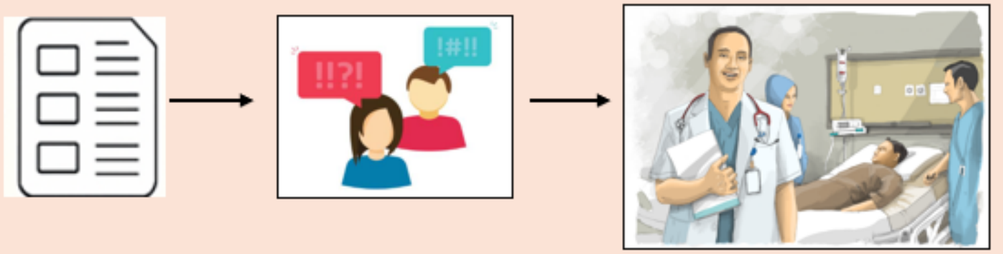
⇩

Results

- DistilBERT achieved the highest and most balanced performance.
- **TF-IDF + LR** had high recall but lower precision and accuracy.
- **T5** reached good accuracy but lacked consistency in recall and F1-score.

Pipeline

Structured data (Clinical indicators ⇨ **Free-Text patient generation by GPT-4** ⇨ **Urgency Classification (Not urgent), 1 (urgent)**



Comparison of Evaluation Metrics