Symptom to Disease Detection

Introduction to Artificial Intelligence (AAI-501-IN1)
Masters of Science in Applied Artificial Intelligence
University of San Diego
Instructor: Ankur Bist

Group 3 Members:

- Ved Prakash Dwivedi
- Bharath TS
- Manu Malla

Team Members & Their Contributions

Manu Malla

- Project identification
- Exploratory Data Analysis
- Developing Statistical Model
- Project Documentation for the above tasks.
- Conclusions and Report

Bharath TS

- Project identification
- Developing Deep learning model using BERT
- Project Documentation for BERT model
- Conclusions and Report

Ved Prakash Dwivedi

- Project identification and task planning
- Exploratory Data Analysis
- Deep learning model using LSTM and GRU
- Project documentation for LSTM, GRU model
- Git Merges
- Conclusions and Report

Abstract

- **Aim:**
    - Preliminary disease diagnosis based on textual symptom descriptions.
- **Data:**
    - The dataset consists of 1200 datapoints and has two columns: "label" and "text".
    - label : contains the disease labels
    - text : contains the natural language symptom descriptions.
    - **24 different diseases**, and each disease has **50 symptom descriptions.**
- **Methods:**
    - Data cleaning: Drop Unnamed columns, Label encoding.
    - Exploratory Data Analysis: Class distribution & World count analysis.
    - Preprocessing: Tokenization, Stopword removal and label encoding.
    - Feature extraction: TF-IDF vectorization for Logistic regression, word embedding + LSTM for deep learning, encodings generated from BERT tokeniser
- **Models:**
    - Logistic Regression, LSTM + GRU Hybrid and BERT
- **Outcome:**
    - Performance of the models
- **Suggestion:**
    - Deployment
    - Real world adaptation

Introduction

Problem

Overlapping symptoms and difficulty in accurately identifying diseases often result in multiple diagnostic attempts, causing delays in treatment and extending patient suffering

Objective

Develop an Machine Learning and NLP model to predict the most likely disease from free-text symptom descriptions, enabling early diagnosis and faster treatment.

Dataset

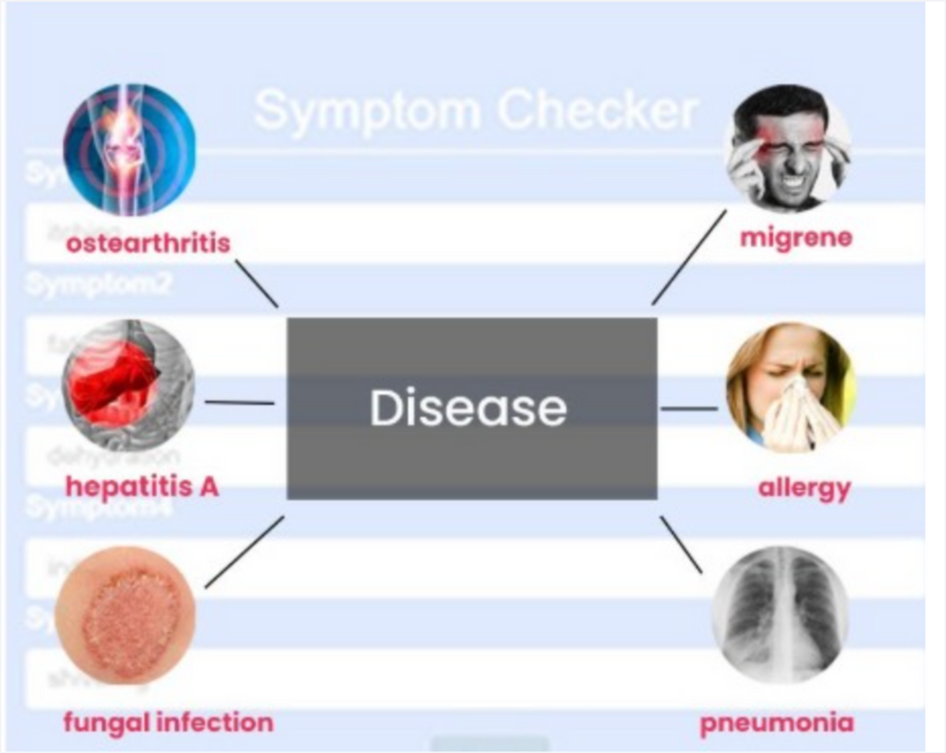
The dataset contains 2 fields with textual descriptions of symptoms and label representing the disease associated with symptoms.

Goal

To build an interpretable, highly accurate, and easily deployable symptom-to-disease prediction model that supports clinicians in making faster, data-driven diagnoses, reduces diagnostic uncertainty, and improves patient treatment outcomes

Real-World Motivation

- **Healthcare challenges:**
  - Scarcity of doctors.
  - Delayed consultations.
  - Lack of quick and preliminary screening tools



- **Patient challenges:**
  - Difficulty in interpreting symptoms.
  - Reliance on misinformation from non-expert sources.
  - Multiple visits to doctors and follow-ups for proper treatment

Data Handling & Preparation

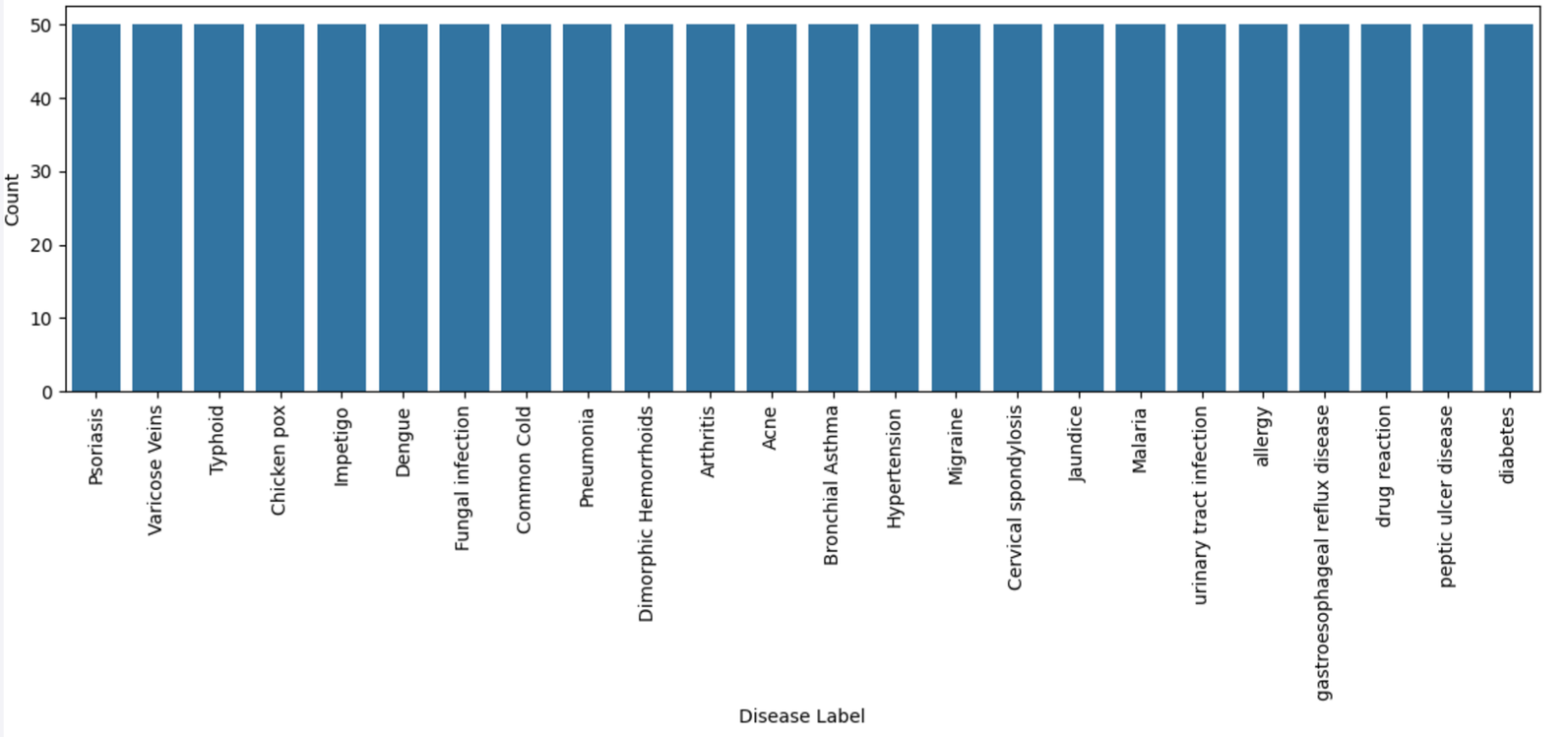
- **Dataset Overview**
  - 1,200 training records with 3 column.
  - Each disease has 50 records.

	Unnamed: 0	label	text
0	0	Psoriasis	I have been experiencing a skin rash on my arm...
1	1	Psoriasis	My skin has been peeling, especially on my kne...
2	2	Psoriasis	I have been experiencing joint pain in my fing...
3	3	Psoriasis	There is a silver like dusting on my skin, esp...
4	4	Psoriasis	My nails have small dents or pits in them, and...

- **Data cleaning**
  - Removal of unnamed column
  - remove punctuation and trailing spaces
- **Label Encoding**
  - Categorical variable (Disease name) converted in Numerical format.

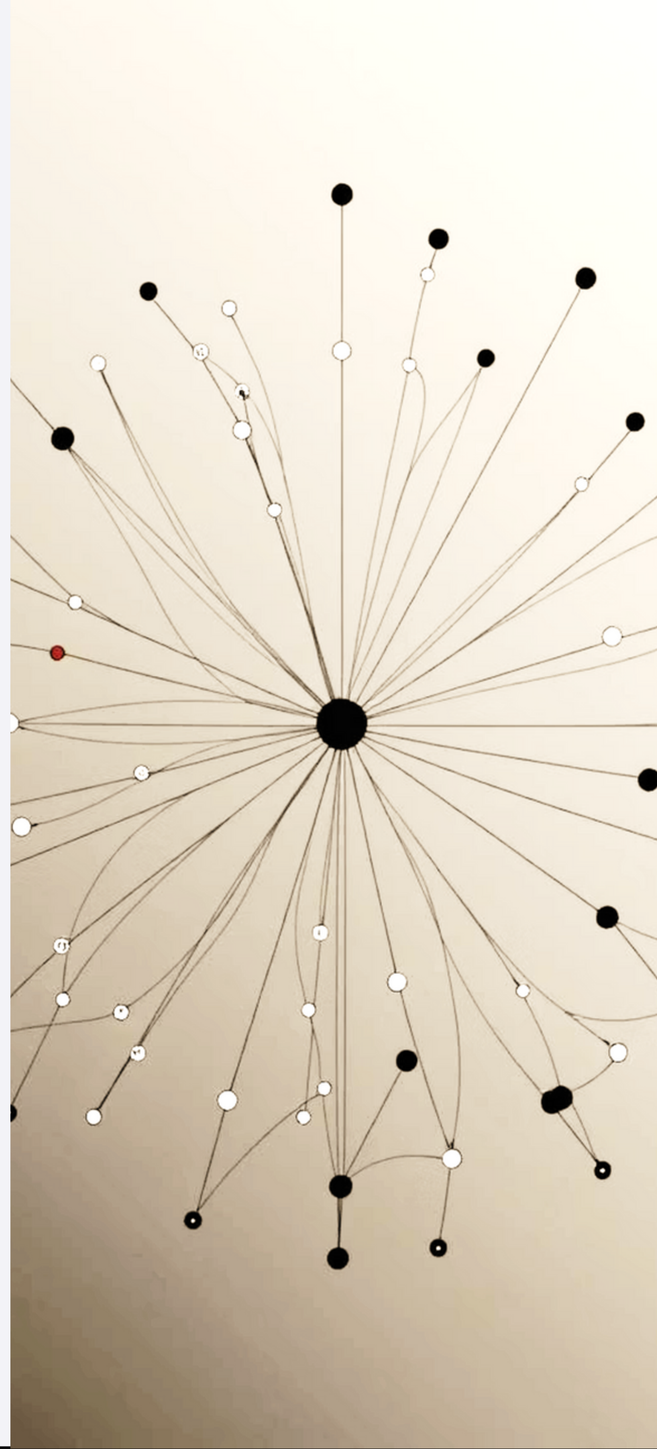
	label	text	encoded_label
0	Psoriasis	I have been experiencing a skin rash on my arm...	15
1	Psoriasis	My skin has been peeling, especially on my kne...	15
2	Psoriasis	I have been experiencing joint pain in my fing...	15
3	Psoriasis	There is a silver like dusting on my skin, esp...	15
4	Psoriasis	My nails have small dents or pits in them, and...	15

Exploratory Data Analysis - Class distribution and Word cloud analysis



Summary :

- Balanced dataset and most disease have same number of samples which are uniform.
- Word cloud analysis: Top frequent terms and Semantic clues

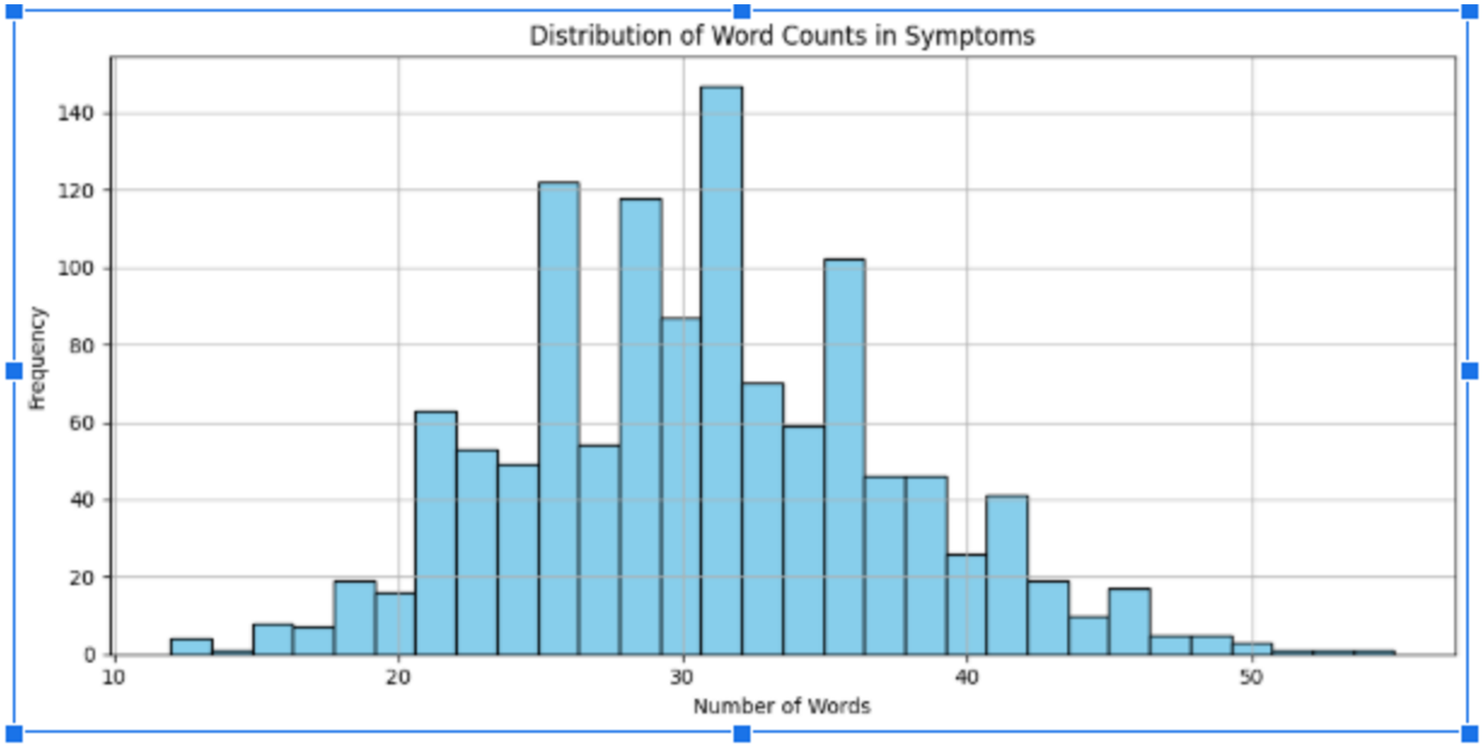


Exploratory Data Analysis – Word cloud individual symptoms:

- Each sub plot highlights the **most common terms** for that **specific disease label**
- Comparison across the bales allows detection of Unique symptom terms (e.g wheezing for respiratory conditions)

- Shared high-frequency terms (e.g., pain, fever) indicating possible symptom overlap between diseases.
- As shown in the wordcloud -
  - Pain in stomach, constipation, diarrhea are the symptoms of **Typhoid** disease
  - Pain with Skin rashes, joint, peeling are the symptoms for **Psoriasis**
-

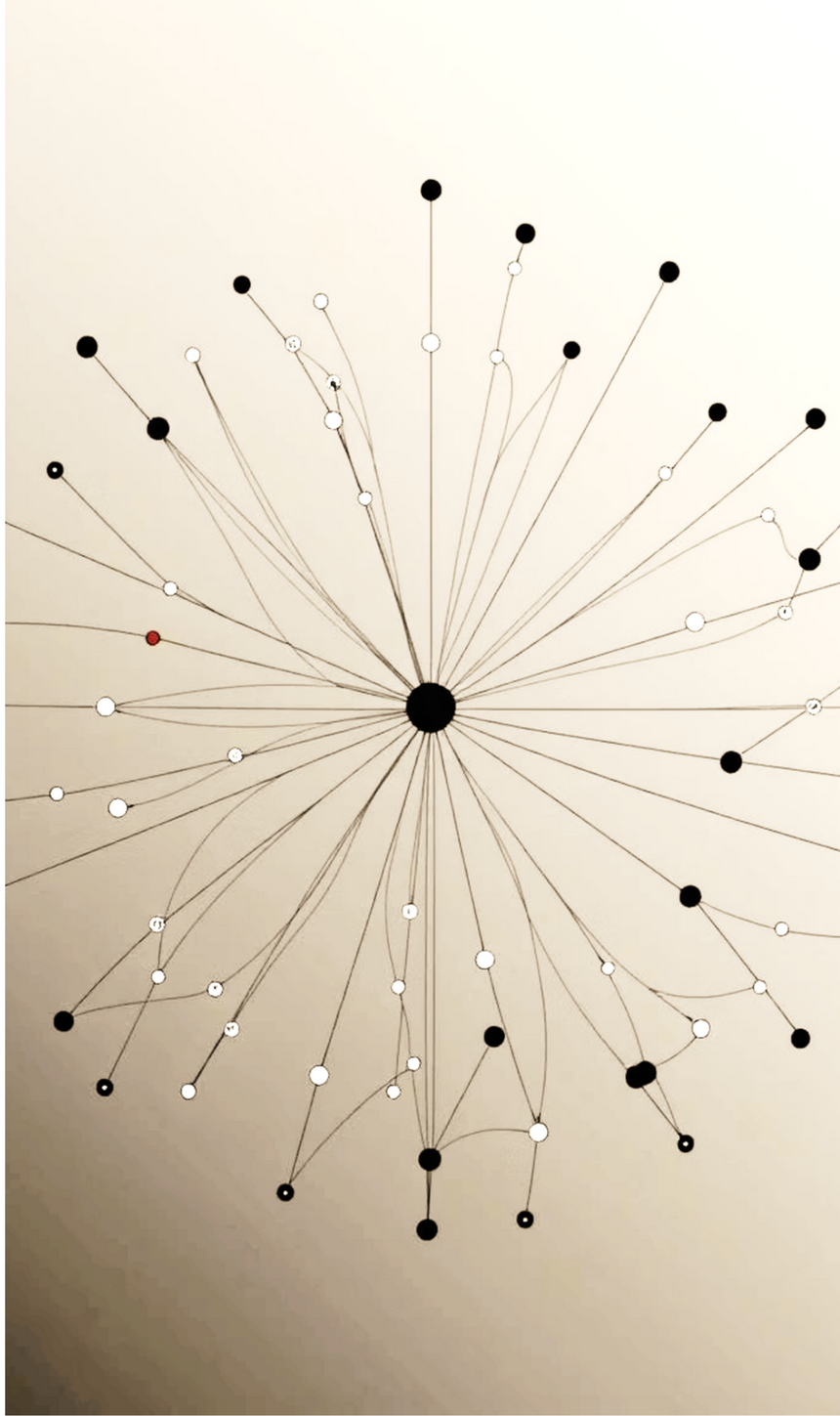
EDA: Word count distribution



Modal Word Count (Most Frequent Bin): Symptom descriptions containing around 30–32 words.

Overall Distribution Shape: Slightly right-skewed fewer extremely long descriptions.

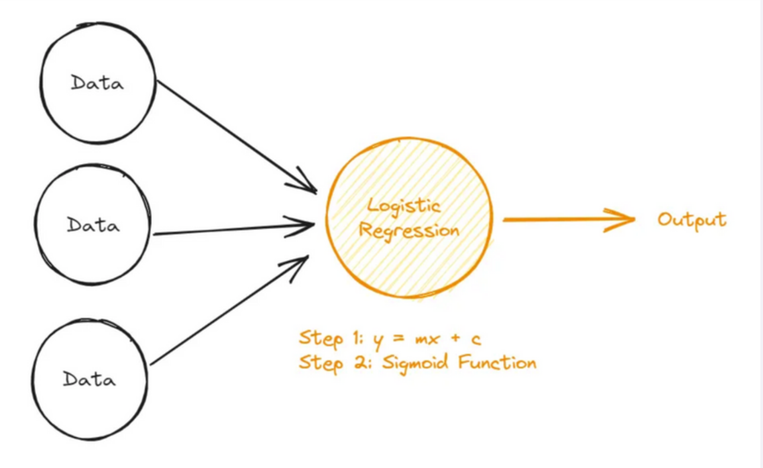
Range and Spread: Symptoms seem to fall in the 20–40 word range.
Outliers: Few entries exceed 50 words split in preprocessing.

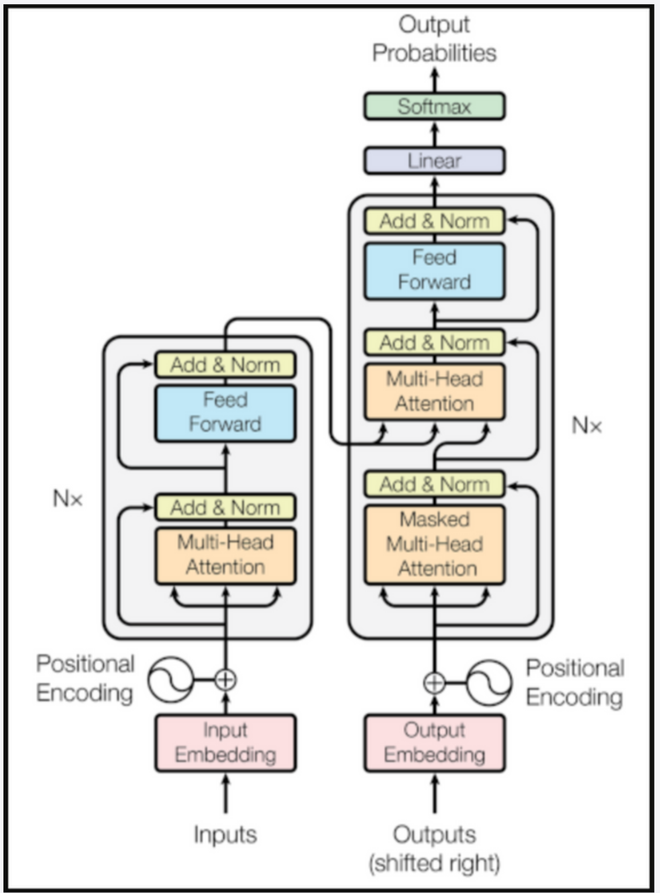


Approach

Logistic Regression with TFIDF

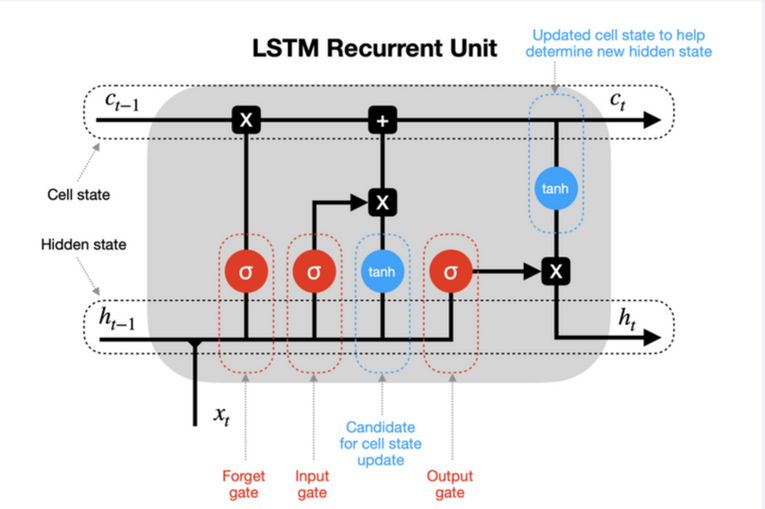
- Classic statistical classifier for linear decision boundaries.
- Relies on TF-IDF weighting to represent text numerically.
- Serves as a strong, interpretable baseline for classification tasks.





Deep Learning Model with LSTM - GRU

- Recurrent neural networks that process sequences step-by-step.
- LSTM and GRU handle vanishing gradient issues to learn long-term dependencies.
- Suitable for text where word order and sequence context matter



Transformer based model - BERT Classifier

- Transformer-based language model using self-attention.
- Generates contextual embeddings that vary by sentence context.
- Pre-trained on large corpora and fine-tuned for specific tasks.

Model Comparison

Model	Business Explanation	Intuitive Explanation	Ease of Use	Accuracy	Complexity
Logistic Regression with TF-IDF	Simple statistical model that finds patterns in word usage to make predictions.	It's like counting how important each word is to guess the answer.	★★★★	★★★	★
LSTM-GRU Deep Learning	Advanced neural networks that learn from the sequence of words for better accuracy.	It remembers the story word by word to guess what’s happening next.	★★	★★★★	★★★
BERT Transformer Model	State-of-the-art AI that understands context and meaning in sentences.	It reads like a smart person who understands every word in the whole sentence.	★	★★★★★	★★★★

Classification Metrics

Precision

- How many of the predicted cases were correct?

Example : Out of all patients the model labeled as Dengue, how many truly had Dengue?

Recall

- How many actual cases did we find?

Example : Out of all patients who truly had Dengue, how many did the model correctly identify?

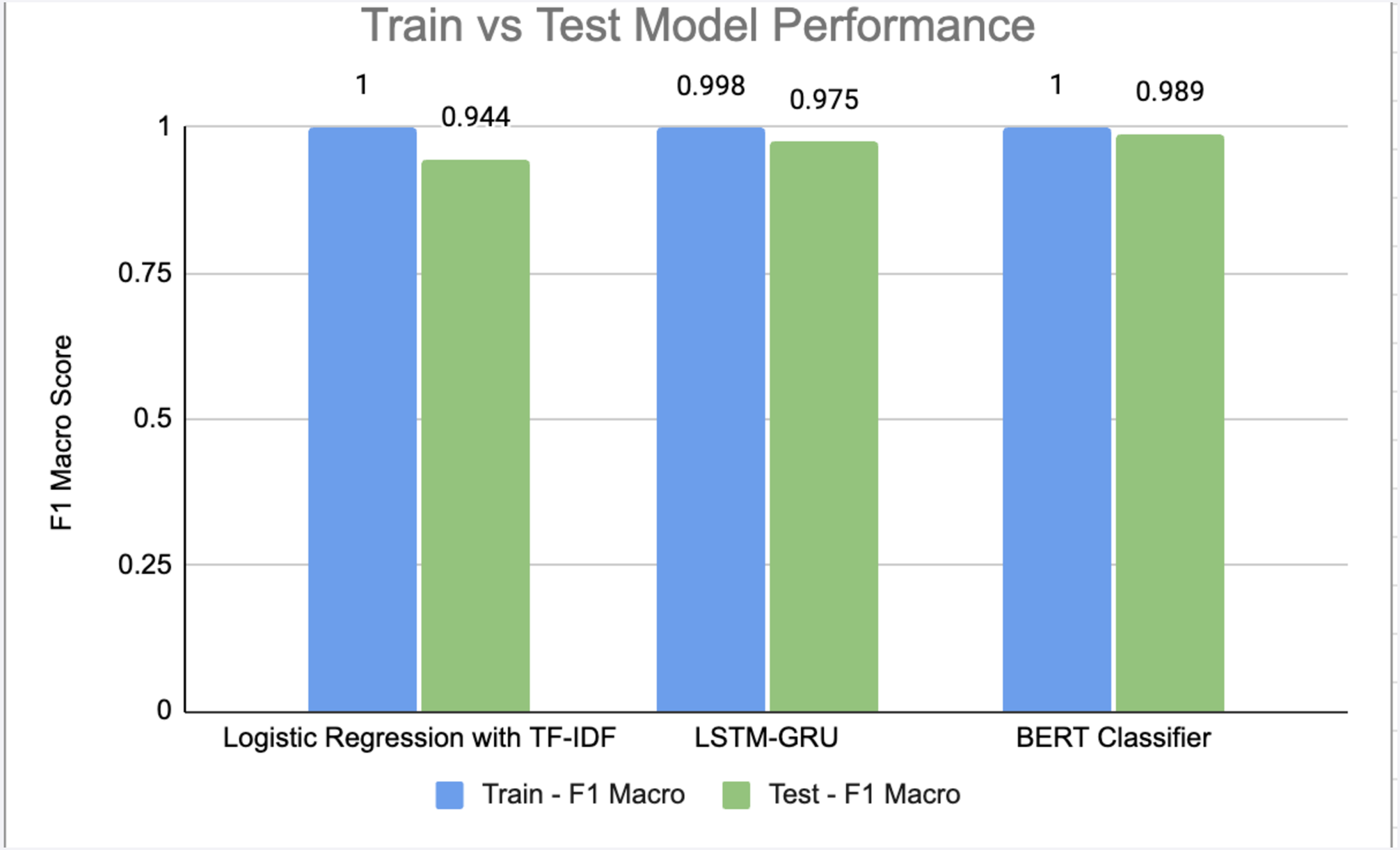
F1 Macro Score

- It presents the ***balanced disease predictions*** across all conditions, ensuring that the model helps all patients equally—not just those with common illnesses.
- Some diseases (like the common cold) will have many more cases.
- Others (like rare tropical diseases) may have very few cases.
- ***Accuracy alone could be misleading*** — the model could do well on common diseases but fail on rare ones.
- ***F1 Macro ensures performance is fair across all diseases, preventing neglect of rare but important illnesses.***

Model Performance (Test Results)

Model	F1 Macro Score	Precision	Recall
Logistic Regression	0.944	0.94	0.94
LSTM+GRU	0.975	0.98	0.98
BERT	0.989	0.99	0.99

Model Performance Comparison



Final Model Selected - BERT

1. **Logistic Regression + TF-IDF**
    a. Best when ***speed, simplicity***, and ***interpretability*** are most important.
    b. Performance is strong but capped because it ignores word order and context.
2. **LSTM + GRU**
    a. Significant improvement because it models ***word order*** and ***sequential patterns***.
    b. Still limited compared to Transformer models for long-range dependencies.
    c. Good trade-off for ***moderately complex NLP tasks*** when BERT-level compute is unavailable.
3. **BERT Classifier**
    a. ***Highest F1 (0.989)*** due to ***contextual embeddings*** and ***self-attention*** capturing both local and global dependencies.
    b. Excels at nuanced multi-class classification where subtle word meaning shifts matter.
    c. Downside: ***compute heavy***, larger ***latency***, and harder to interpret.

Final Verdict:

For production, the choice depends on constraints:

1. If accuracy is ***critical*** and ***compute*** resources allow, ***BERT is the clear winner.***
2. If moderate ***compute*** is available and you still need strong performance, ***LSTM+GRU*** is a solid middle ground.
3. If ***speed, low resource*** usage, and ***interpretability*** matter most, ***TF-IDF + Logistic Regression*** is still very effective.

Business Impact

Benefits for Patients

- **Faster Diagnosis Support**
  - **Quick preliminary identification** of possible conditions.
  - Immediate feedback without long waiting times.
  - Reduces anxiety by providing instant clarity.
- **Improved Communication**
  - Structured **symptom report to share with doctors.**
  - Avoids miscommunication of symptoms.
  - Encourages patients to prepare for consultations.
- **Increased Health Awareness**
  - **Educates patients about potential conditions.**
  - Motivates proactive health monitoring.
  - Helps track symptom progression over time.
- **Accessibility & Convenience**
  - **Useful for patients in remote or underserved areas.**
  - Available 24/7 via mobile devices.
  - **Reduces dependency on immediate in-person visits.**

Benefits for Clinicians/Doctors

- **Streamlined Clinical Workflow**
  - **Reduces time spent on initial symptom triage.**
  - Allows focus on complex cases rather than repetitive assessments.
  - Improves efficiency in patient throughput.
- **Better Resource Allocation**
  - Helps **prioritize patients needing urgent care.**
  - Reduces unnecessary testing for low-risk cases.
  - Enables pre-consultation planning.
- **Data-Driven Insights**
  - Aggregated **data aids in detecting disease patterns.**
  - Supports early outbreak detection.
  - Provides evidence for research and policy.
- **Clinical Decision Support**
  - Assists in differential diagnosis suggestions.
  - **Enhances diagnostic accuracy by integrating AI recommendations.**
  - Improves patient outcomes through informed decision-making.

Final Recommendations

- **Deployment**
  - **Select & Deploy Optimal Model**
    - BERT for maximum efficiency where compute allows
    - LSTM/GRU for balance
    - LogisticRegression for speed and low resource usage
  - Deploy via Docker + REST API for easy integration into healthcare systems.

- **Real-World Adoption**
  - **Integration** with **Telemedicine** Platforms for instant triage.
  - **Multi-Language Support** to cater to diverse populations.
  - **Explainability Tools (LIME/SHAP)** to build trust with doctors.
  - Domain-Specific Fine-Tuning for local healthcare data.
  - **Continuous Feedback Loops from doctors & patients** to refine predictions.

- **Achieve the Goals as planned**
  - To build an **interpretable, highly accurate, and easily deployable symptom-to-disease prediction model** that supports **clinicians** in making **faster**, **data-driven diagnoses**, reduces diagnostic uncertainty, and improves patient treatment outcomes

Conclusions

- **Robust Classification Pipeline**
  - End-to-end AI solution: preprocessing → model training → evaluation
  - Handles diverse symptom descriptions with high accuracy
- **Model Benchmarking**
  - Tested traditional ML & deep learning models
  - LSTM+GRU hybrid outperformed others in sequential symptom analysis
- **Hybrid Model for Deployment**
  - High precision, recall, and F1-scores
  - Suitable for clinical decision support, telemedicine, and self-assessment tools
- **Business & Clinical Impact**
  - Faster Diagnoses: Improves doctor-patient interactions
  - Accessibility: Expands healthcare reach
- **Monetisation Opportunities**
  - **Subscription plans** for hospitals & clinics
  - **B2B licensing** for healthcare providers & insurers
  - **API integration** fees for apps/wearables
  - Premium patient AI health reports
  - **Partnerships with pharma** for targeted campaigns