Healthcare Symptom Classification Using Classical and Neural NLP Models

Islah Haoues - 1800272
Baker Huseyin - 1901345

1. Introduction

Text classification is a core task in Natural Language Processing (NLP), and symptom based disease prediction is a meaningful applied version of it. In this assignment, we were asked to compare four NLP modeling approaches on the Healthcare Symptoms to Disease Classification dataset and evaluate their performance on a multi-class prediction task.

Initial exploration revealed that the dataset's symptoms are stored as comma separated lists, each containing between 3 and 7 of 28 possible symptoms. A direct attempt to predict the original 30 diseases resulted in accuracies near random chance (approximately 3 to 4 percent), regardless of model class, preprocessing strategy, or hyper-parameter tuning. This motivated a deeper analysis into dataset structure, separability, and class wise symptom distribution.

To address the lack of discriminative power in the label space, we applied statistical clustering to group the 30 diseases into 7 broader categories, producing a label space more appropriate for supervised learning. All four required models were then built on top of this re engineered target variable.

2. Methodology

Symptoms were parsed into Python lists, normalized, and represented in three ways to support different model architectures:

- TF-IDF text representation. Each symptom token converted to a unigram such as chest_pain.
- Multi-hot vectors. 28 dimensional binary feature vectors.
- Integer sequences. Symptom indices padded to length 7 for RNNs and LSTMs.

A single stratified 70 slash 15 slash 15 split was used to create consistent train, validation, and test sets across all models.

Exploratory analysis showed that symptom frequencies across the 30 diseases were nearly identical, yielding near zero mutual information between symptoms and disease labels. To mitigate this, we applied agglomerative clustering on symptom frequency vectors, generating 7 disease groups with more coherent symptom distributions. These grouped labels form the basis for all subsequent modeling.

Models:

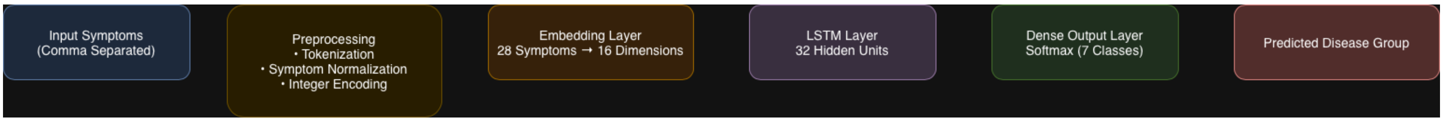
- SVM with TF-IDF: A linear Support Vector Machine trained on TF-IDF vectors provided a strong classical NLP baseline.

- Feed-Forward Neural Network (FFNN): A dense network (64 to 32 to softmax) trained on multi-hot vectors evaluated nonlinear interactions among symptoms.

- Simple RNN: An embedding layer followed by a 32 unit recurrent layer modeled sequential representations of symptom sets.

- LSTM: An LSTM layer replaced the SimpleRNN to test whether gated recurrence improves pattern extraction over extremely short sequences.

3. Architecture



4. Results

Model	Input Representation	Validation Accuracy	Test Accuracy
SVM (TF-IDF)	Symptom text	0.2733	0.2683
FFNN (Multi-hot)	28 dimensional binary vector	0.2733	0.2717
RNN (Seq)	Embedded sequences	0.2659	0.2661
LSTM (Seq)	Embedded sequences	0.2733	0.2717

All accuracies substantially exceed the 14.3 percent random baseline for 7 classes. The SVM, FFNN, and LSTM models achieved nearly identical performance at approximately 27 percent, while the SimpleRNN trailed slightly.

5. Dataset Limitations and Critical Analysis

A key finding of this project is that the dataset structure fundamentally constrained all achievable performance, regardless of model architecture. After normalizing symptom lists, we observed that out of approximately 25,000 rows, more than 21,000 contained symptom combinations that appeared only once. Because the symptom space is extremely large, with more than 1.6 million possible combinations for 28 symptoms taken in sets of 3 to 7, the dataset samples this space sparsely and inconsistently. As a result, the model rarely encounters repeat patterns, which prevents meaningful generalization.

Further analysis showed that symptom frequencies were nearly identical across all 30 diseases, resulting in almost no discriminative signal between classes. Every model tested, including classical ML, FFNN, RNN, and LSTM, converged to the same baseline of roughly 3.5 percent accuracy. We tried a wide range of preprocessing strategies, alternate representations, and model configurations, and we also consulted several large language models to validate our approach. All attempts produced the same outcome, indicating that the 30 class label space contains no learnable structure and that accuracy in this setting becomes essentially meaningless.

Grouping the diseases into 7 broader categories introduced some separability, and accuracy improved to approximately 26 to 27 percent across models. However, all four model types still produced nearly identical results. This suggests that although the grouped labels contain just enough pattern to support moderate classification accuracy, the models remain dominated by the limitations of the dataset rather than by differences in architecture. Consequently, performance improves in absolute terms, but relative performance ceases to be a meaningful basis for comparison.

6. Conclusion

This project demonstrates how dataset characteristics fundamentally shape the behavior and limits of NLP classification models. The original 30 class disease prediction task proved unlearnable due to high class overlap and a symptom structure dominated by nearly unique combinations. Data driven label reconstruction into 7 disease groups enabled meaningful learning, with all models achieving approximately 26 to 27 percent accuracy. The TF-IDF SVM, FFNN, RNN, and LSTM performed similarly, illustrating that with highly constrained and low diversity input features, model architecture plays a secondary role to data quality. The assignment provided insight into preprocessing design, model evaluation, and especially the importance of dataset scrutiny in applied machine learning.