LSTHMM and HMM Random Forests for Malware Classification

Ritik Mehta, Advisors: Dr. Genya Ishigaki, Dr. Mark Stamp

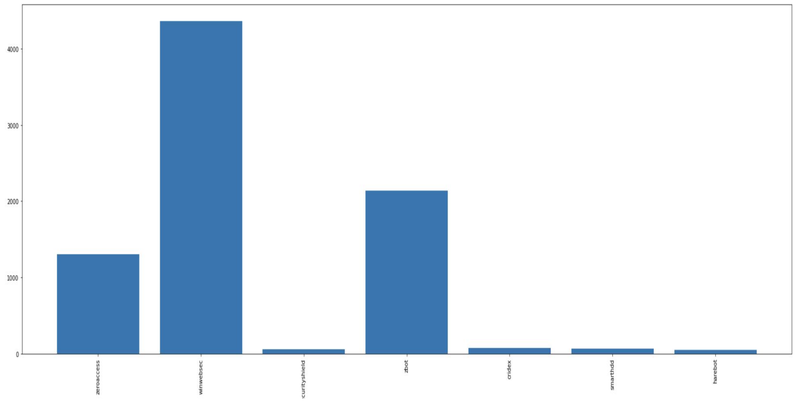
Department of Computer Science, San José State University

ABSTRACT

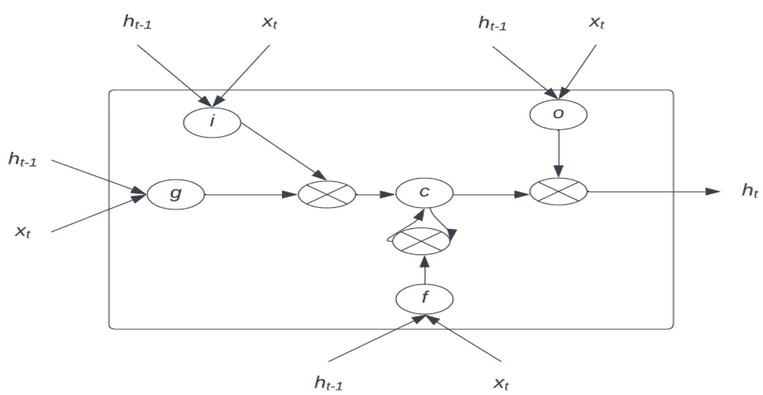
Malware are softwares or programmes that are designed to infect a computer, server, computer network, or leak sensitive data. Classifying malware into different categories help in determining their behaviour and the extent of damage they can do to a computer system. Traditionally, techniques such as LSTM, HMMs, and Random Forests have been used for malware classification. In this research, we aim to combine these techniques techniques and compare the results with the traditional approaches.

BACKGROUND AND DATASET

The dataset consisted of 48 classes. For the purpose of classification, we removed the classes which had less than 50 samples. We were left with 7 classes. The data distribution of the 7 classes are as follows:

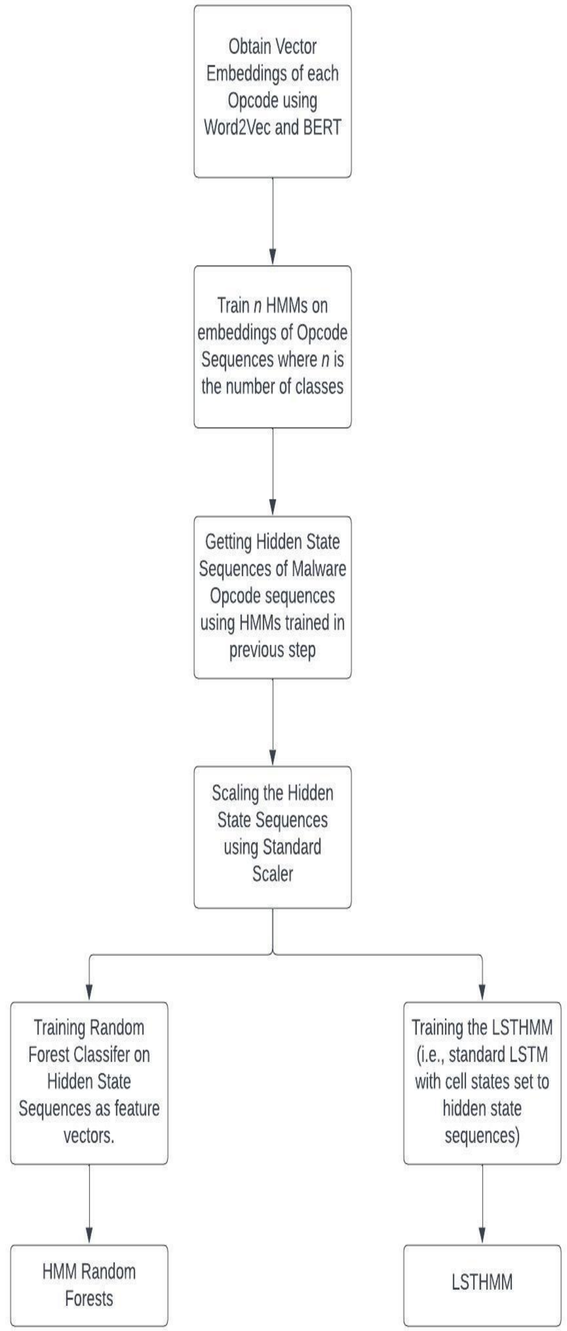


Distribution of Dataset

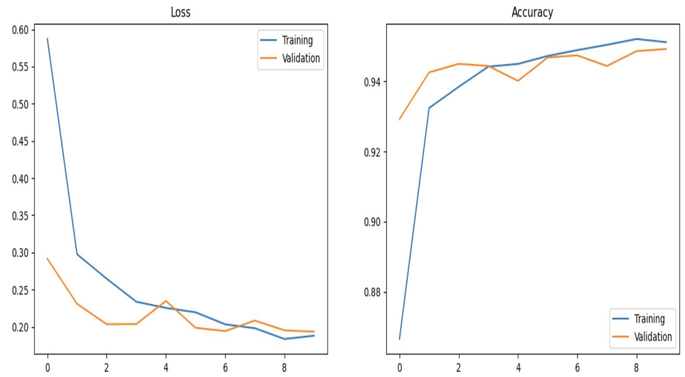


A Simple LSTM cell. For LSTHMM, we replace the cell state with hidden state sequences obtained from HMM.

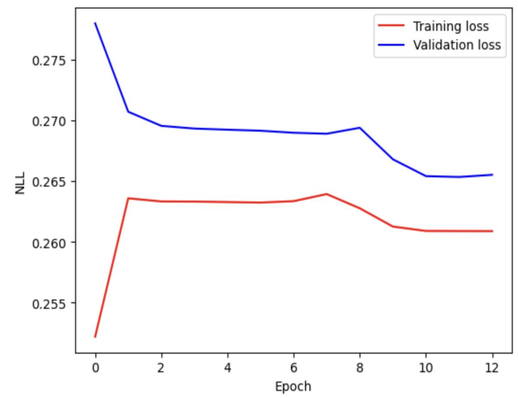
METHODOLOGY



RESULTS



LSTM using Word2Vec Embeddings



LSTHMM

Accuracy of training Hidden State sequences generated by HMM using Random Forests: 99.75%. The best parameters for the Random Forest were as follows:

- criterion: gini
- max_features: None
- n_estimators: 100

FUTURE WORK

1. 1. Test HMM Random Forests on more standard datasets.
2. 2. Get Results of LSTHMM and HMM Random Forests using BERT embeddings.