CMPE 452 - Neural &
Genetic Computing

URL PHISHING DETECTION

Presented by: Kieran Cosgrove, Mile Stosic, and Lucas Coster

Meet the Team Members

Kieran Cosgrove

B.Sc. (Eng.) Computer Engineering

Mile Stosic

B.Sc. (Eng.) Computer Engineering

Lucas Coster

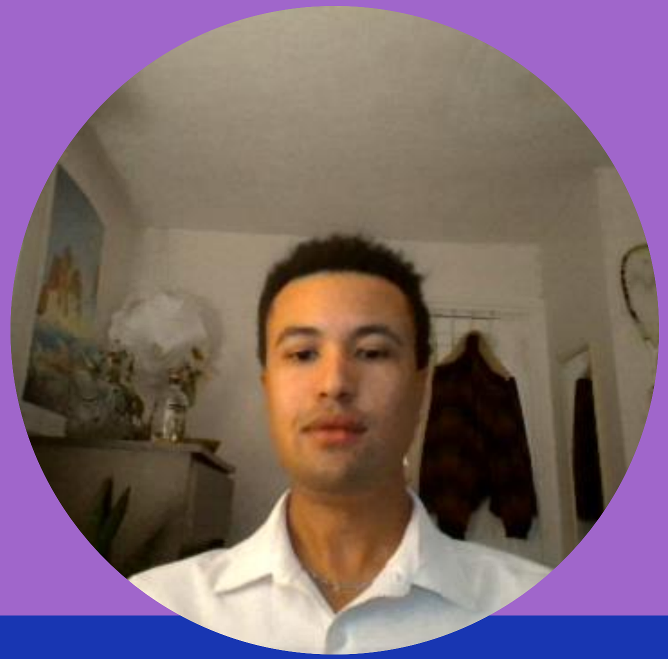
B.Sc. (Eng.) Computer Engineering

Group 4

Agenda

Overview of the Presentation

- Motivation
- Problem Description
- Existing Work
- Dataset Description
- LSTM Introduction
- Design Changes
- Training/Testing
- Results
- Conclusion

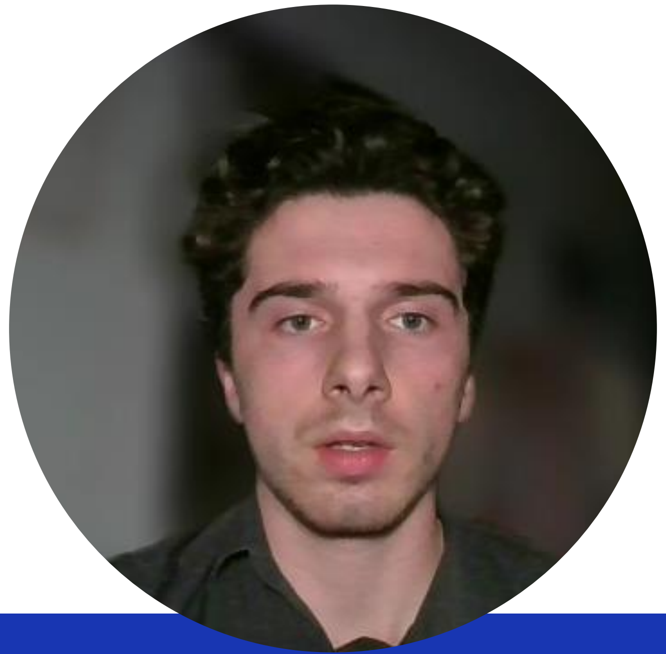


Motivation

Unmasking the Hidden Dangers of Phishing Attacks

Key Points

- **Prevalence of Phishing Attacks:** Highlighting the widespread issue of phishing attacks as a major online security threat.
- **Impact on Individuals and Organizations:** Discussing the significant financial and data losses resulting from phishing attacks
- **Sophistication and Adaptability:** Emphasizing the evolving nature of phishing attacks that can bypass traditional security measures.



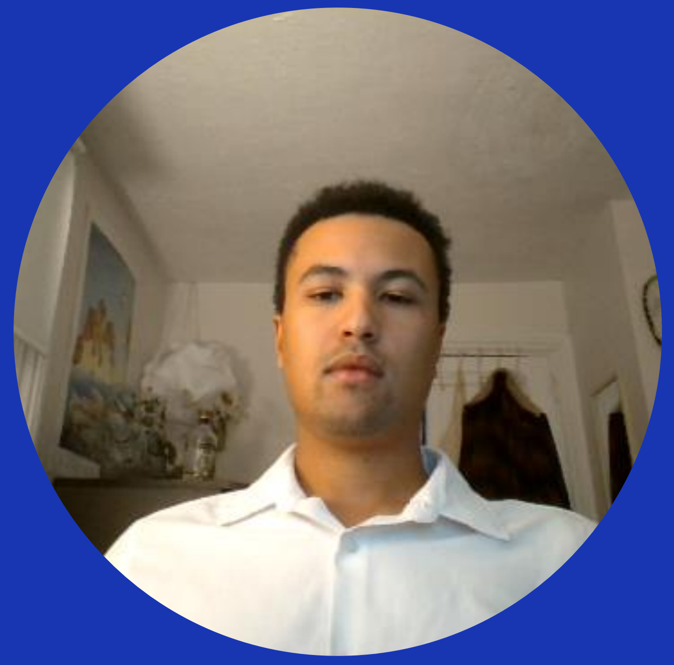
Problem Description

01 Challenges in Detection

02 Limitations of Conventional Approaches

03 Need for Sophisticated Solutions

04 Project Objective



Existing Work

Broad Paper

Do we need hundreds of classifiers to solve real-world classification problems?

Model Related Paper

Segmentation from Natural Language Expressions

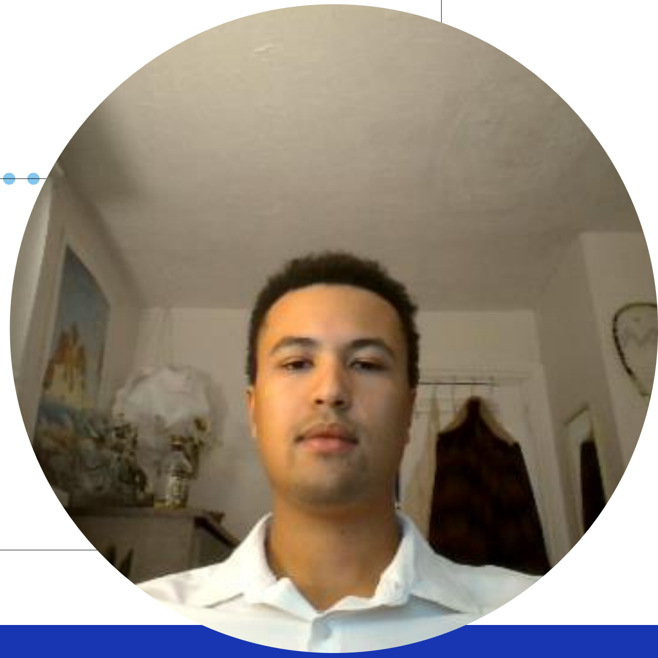
Project Specific Paper

A Deep Learning-Based Phishing Detection System Using CNN, LSTM, and LSTM-CNN

Broad Paper

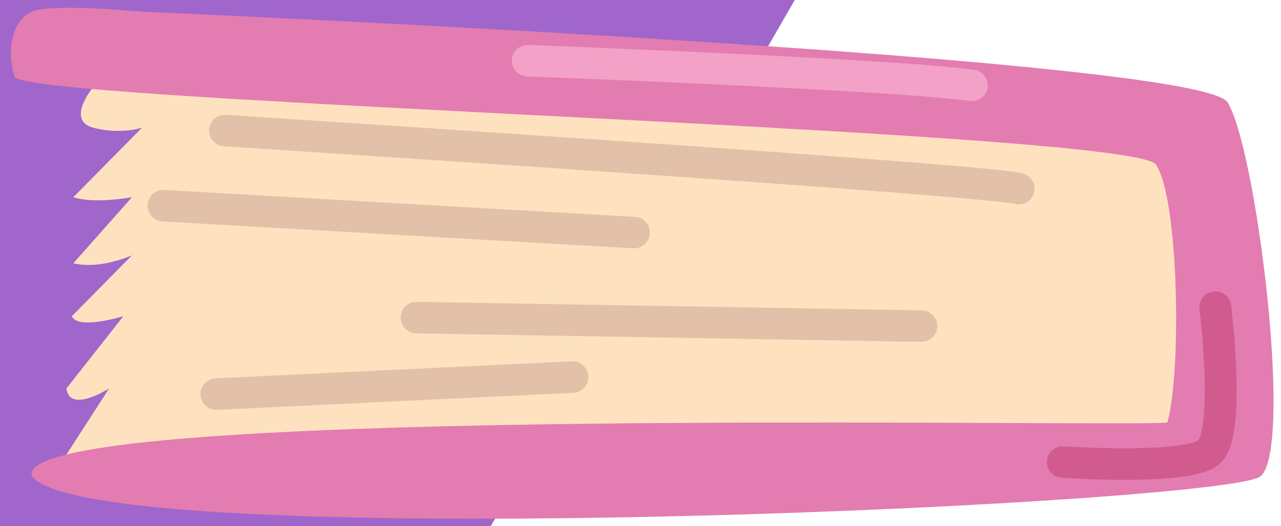
Classifier Selection

■	01	Classifier Efficiency Comparison
■	02	179 Classifiers & 121 Datasets
■	03	LSTM's Advantage
■	04	Parameter Optimization vs. Ensemble Model
■	05	Relevance

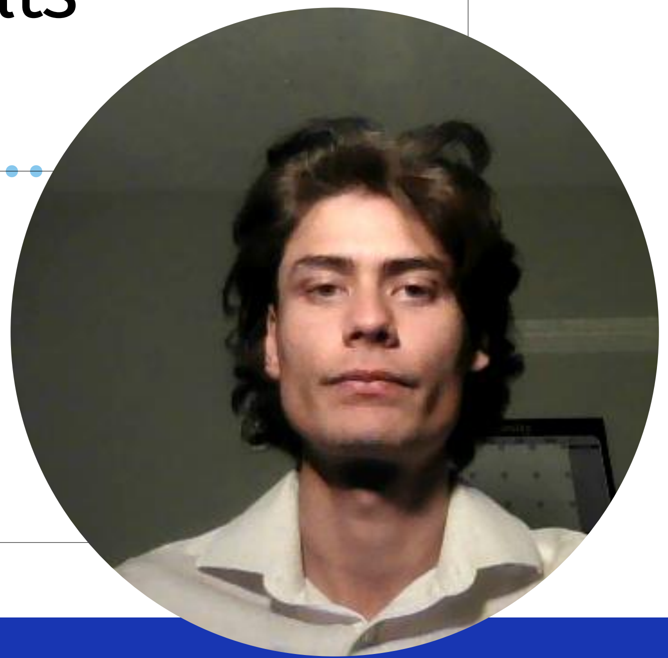


Model Paper

LSTM Model



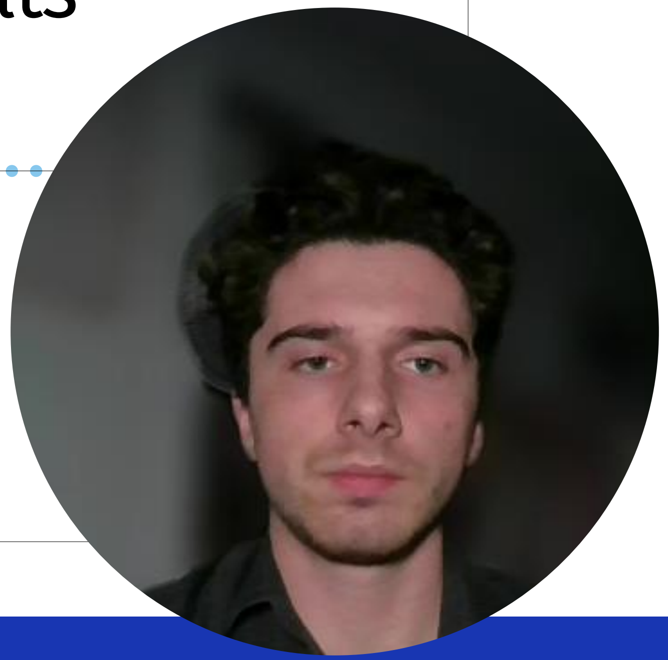
- 01 Image Segmentation Problem
- 02 ReferIt Dataset
- 03 Encoding Language Expression
- 04 Testing And Results
- 05 Relevance



Project Specific Paper

LSTM Model and CNN

- 01 Phishing Attacks
- 02 ISCX-URL2016 Dataset
- 03 CNN-LSTM Model
- 04 Testing And Results
- 05 Relevance

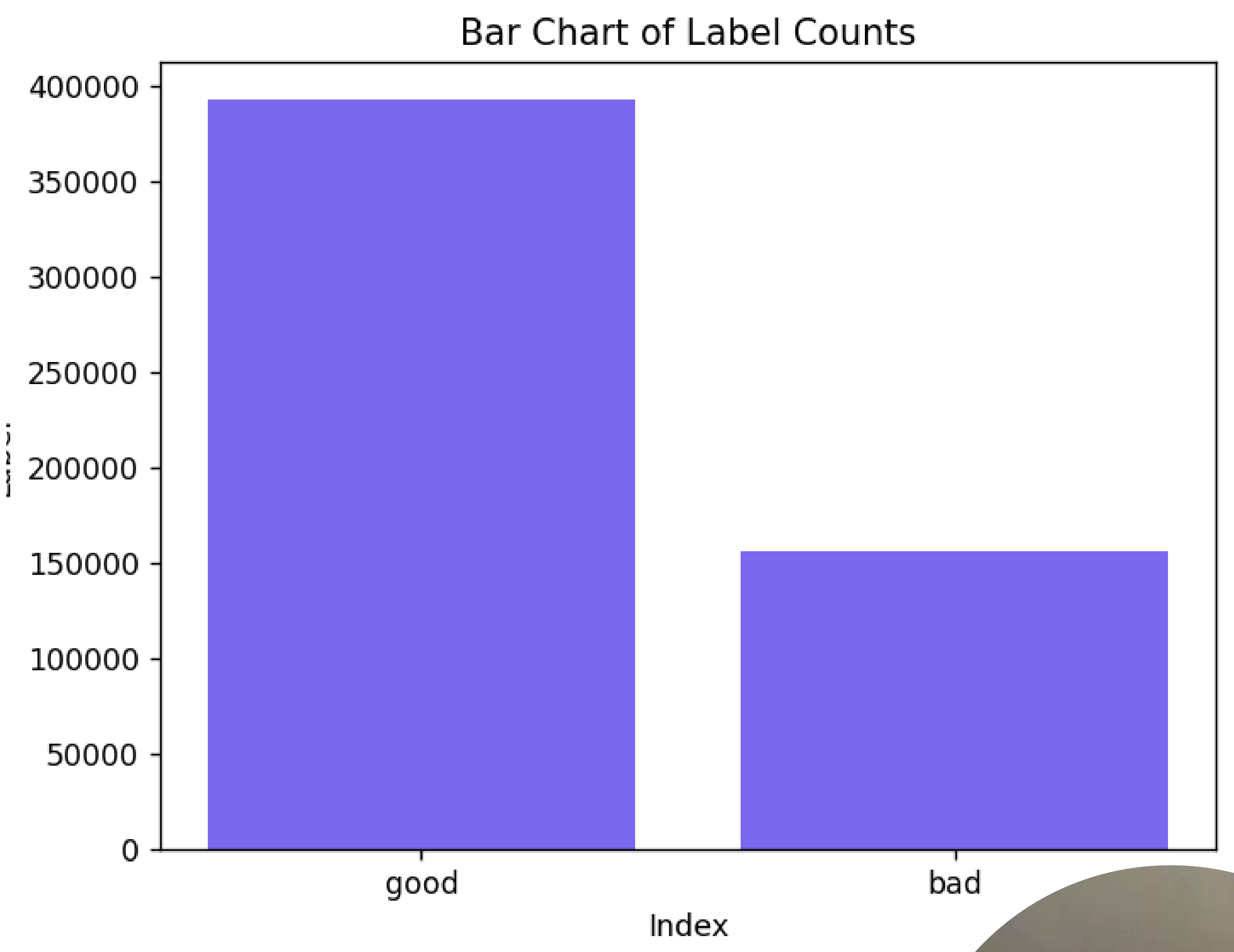


Dataset Description

Research Data vs. Selected Data

01	Literature Datasets
02	Factors
03	Pre-Processing
04	Features

Kaggle: Phishing URLs



Introduction to LSTM

LSTM Models

- A subtype of RNNs
- Ideal for sequential data like language, time series, and URL patterns
- Learns structural and compositional patterns

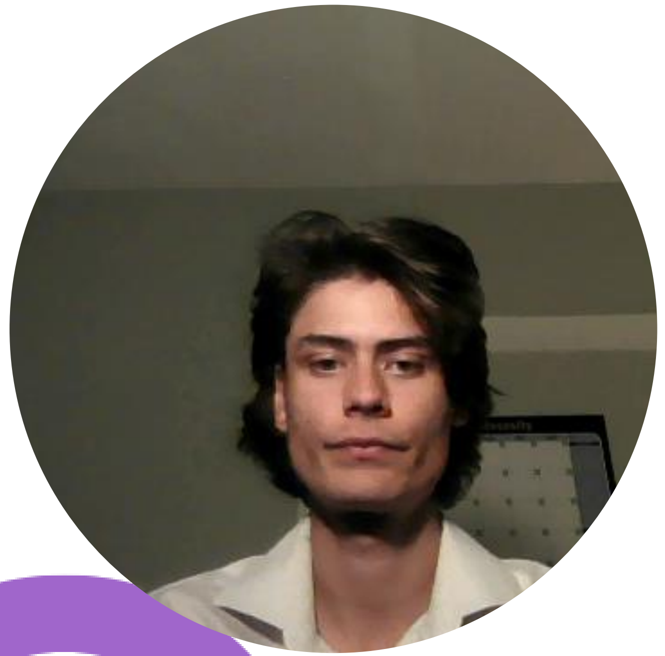
Advantages of LSTM

- Memory of Context
- Adaptability
- Sequential Data Handling

Challenges with LSTM

- Reliant on the quality and variety of training data
- Computationally intensive, requiring significant resources
- Potential for overfitting to training data

Design Changes



SGD Optimzer & MSE loss

Mile

Starting the model training with the implementation of SGD & MSE loss

Dropout & LSTM Units

Lucas

Included Dropout layers in the model and increased the LSTM units

Regularization & Embedding

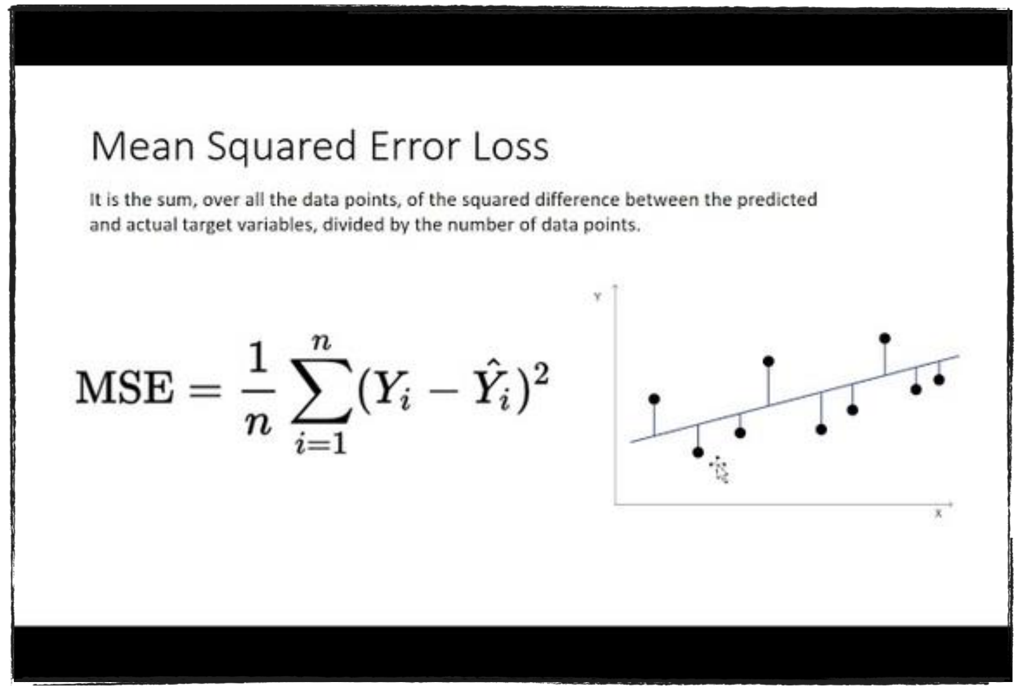
Kieran

Implemented L2 Regularization and a pre-trained Embedded Layer from GloVE

Model 1

Implementing High Level Changes

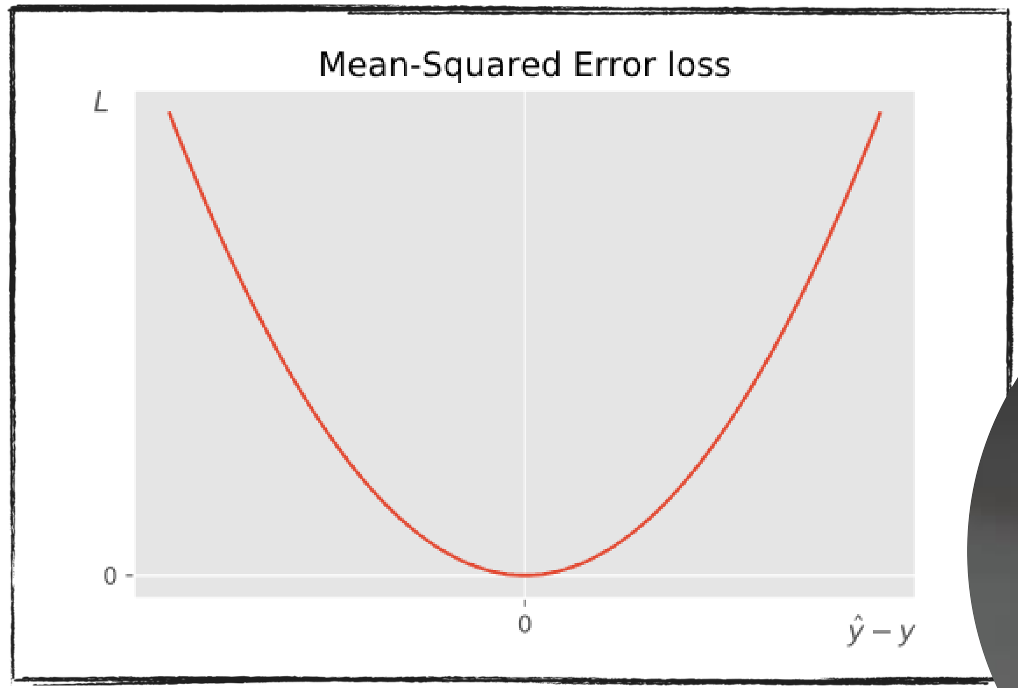
MSE Loss

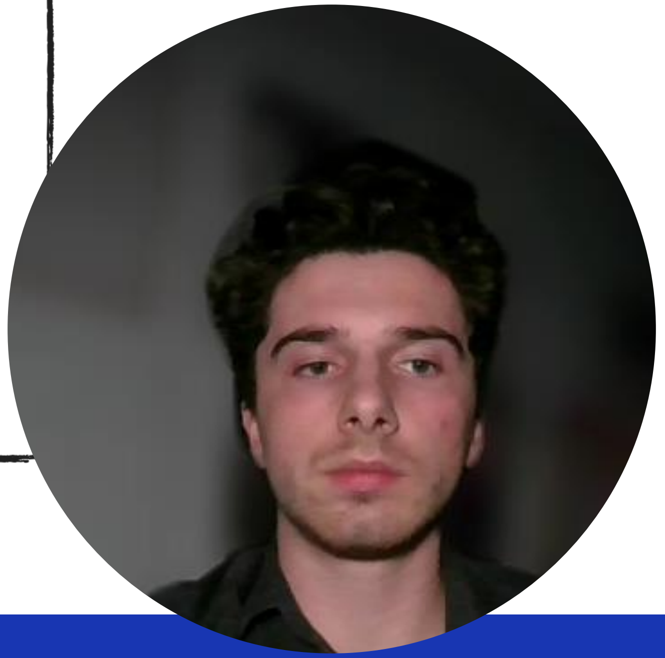


Averaging the squared differences between predicted and actual values.

More sensitive to outliers than other loss functions

A loss function that measures the average of the squares of the errors or deviations





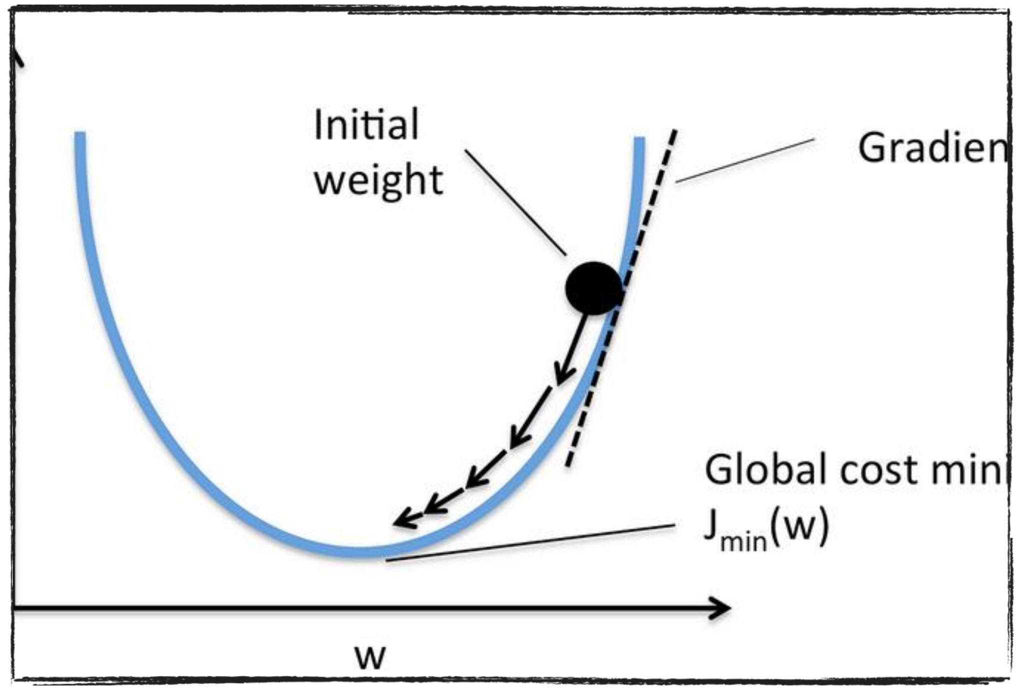
SGD Optimzier

An iterative method for optimizing an objective function with suitable properties

Particularly useful when dealing with large datasets

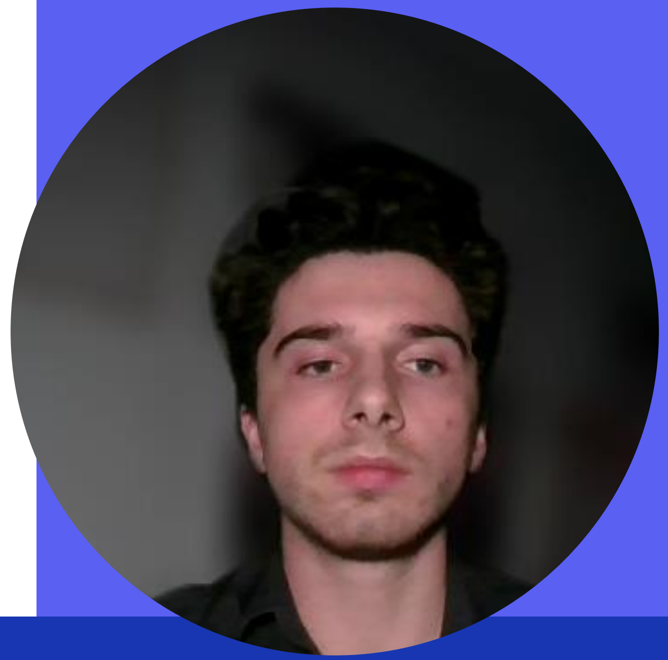
```
# Model with GloVe Embedding Layer
model = Sequential()
model.add(Embedding(len(tokenizer.word_index) + 1
                    embedding_dim,
                    weights=[embedding_matrix],
                    input_length=max_sequence_len
                    trainable=False))
model.add(LSTM(100))
model.add(Dropout(0.5))
model.add(Dense(1, activation='sigmoid', kernel_r

# Compile the model
model.compile(optimizer='SGD', loss=loss_function
```



Introduces randomness during optimization

Includes a learning rate and momentum to stabilize convergence

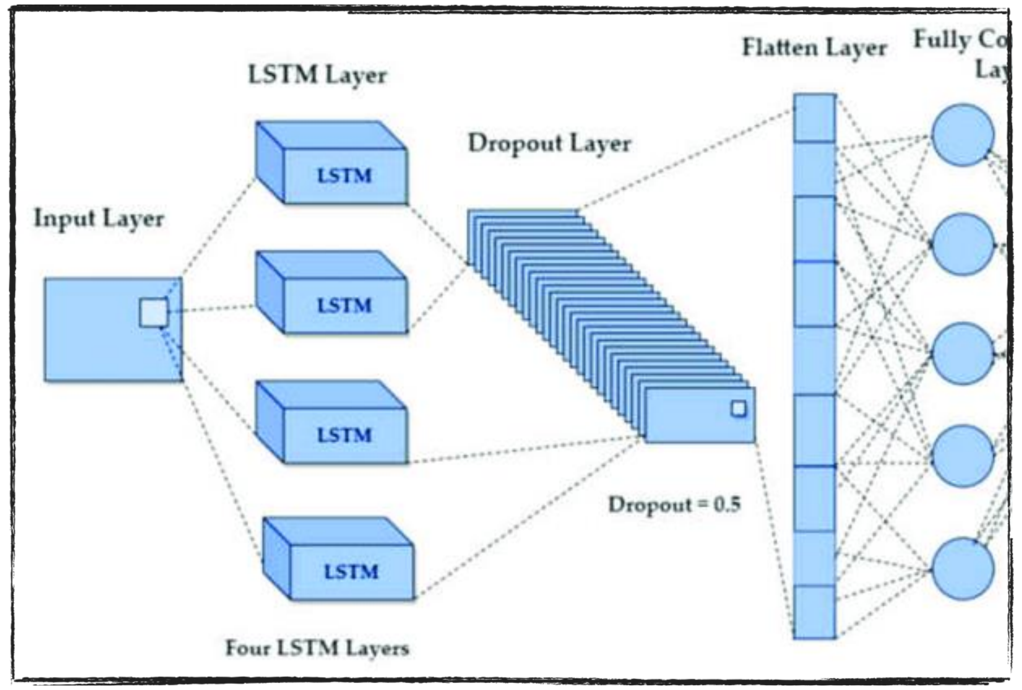


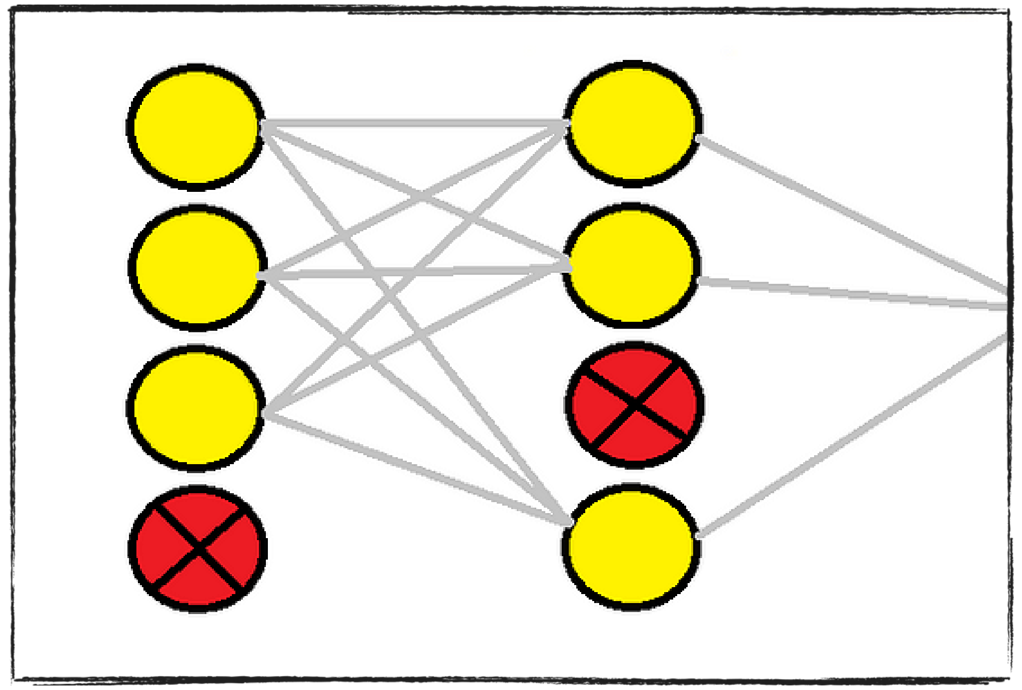
Model 2

Expanding On Base Model

Dropout Layers

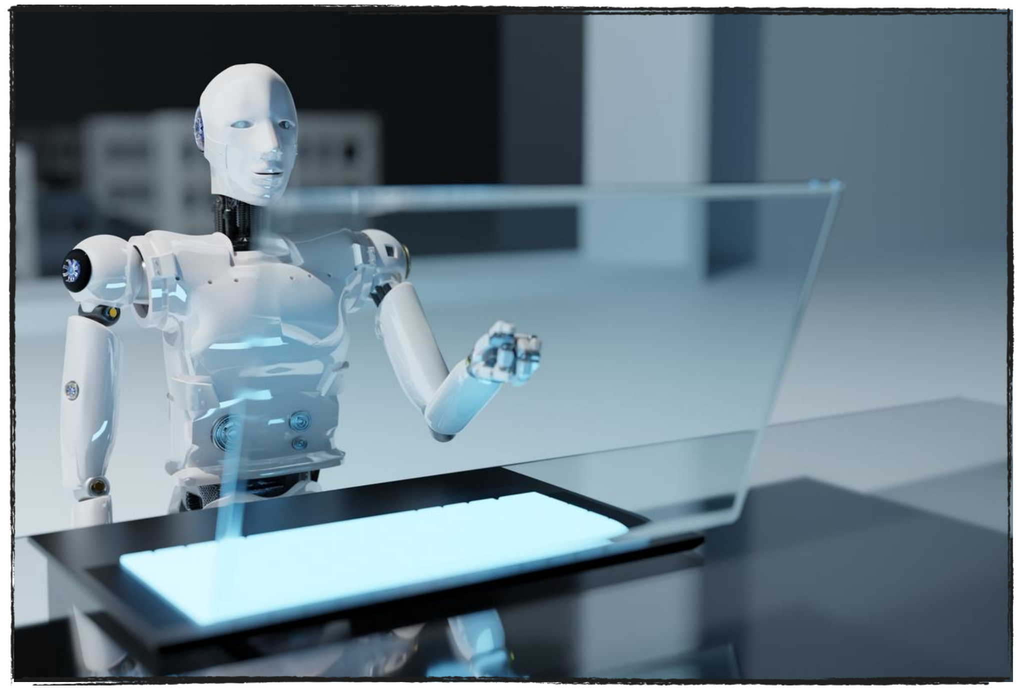
Reduces overfitting by dropping neurons to reduce reliance on specific connections





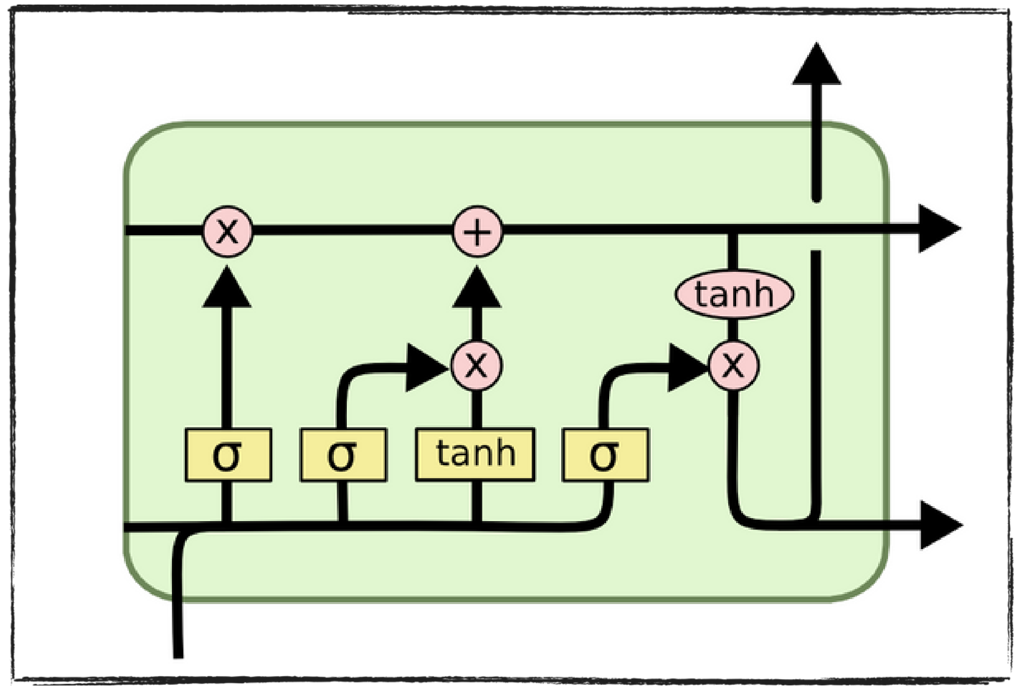
Promotes the learning of diverse features by preventing the co-adaption of hidden units

LSTM Units



LSTM unit is each individual memory cell in the LSTM structure

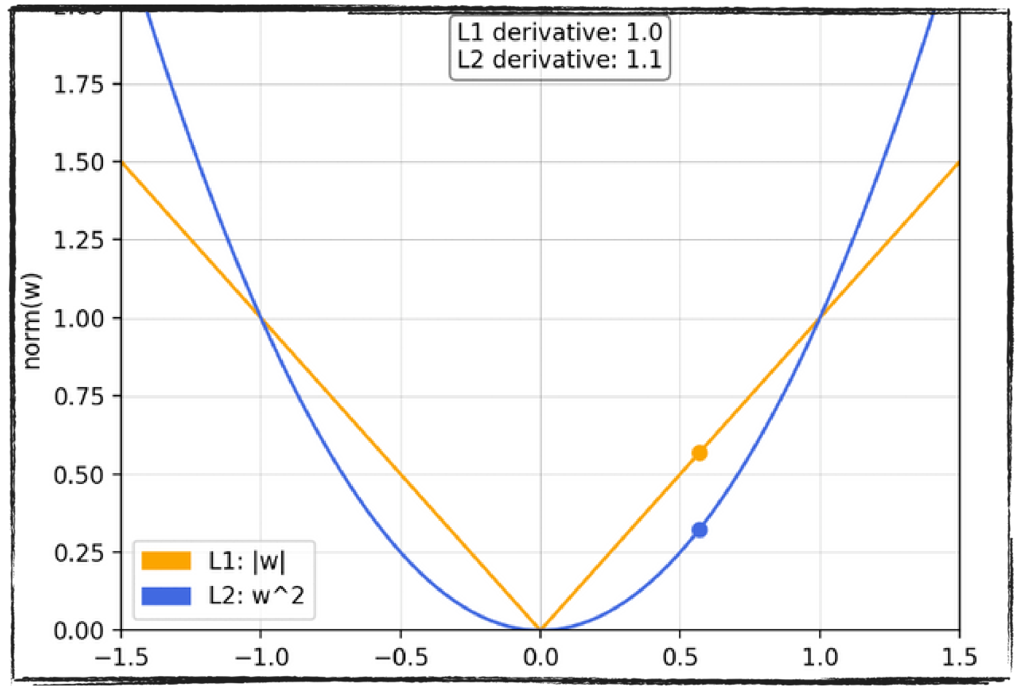
Increasing units allows the model to handle larger datasets and improve generalization



Model 3

Deeper Level Changes

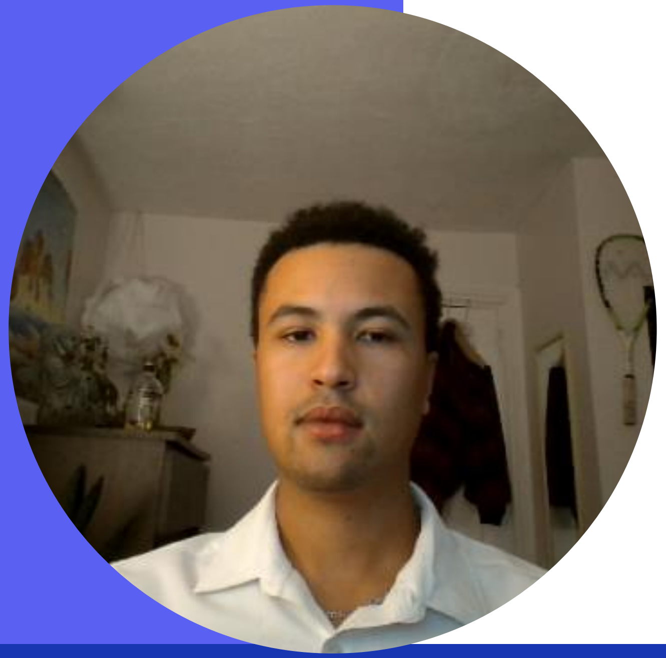
L2 Regularization Implementation



Prevents overfitting by penalizing large weights, enhancing model generalization.

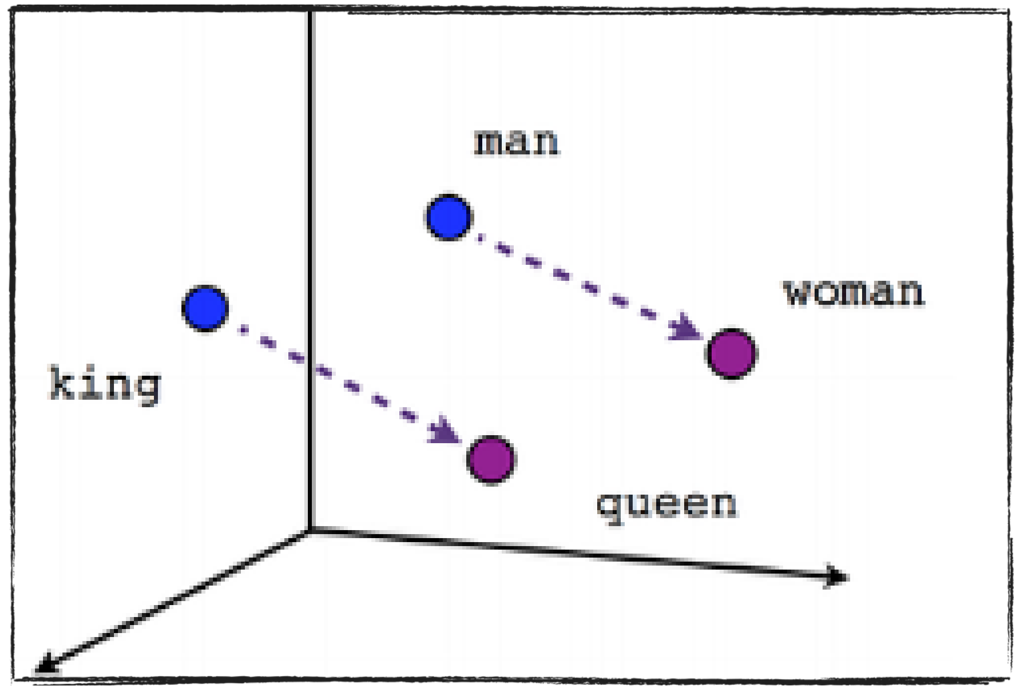
Encourages simpler, more generalizable model without compromising learning capacity.

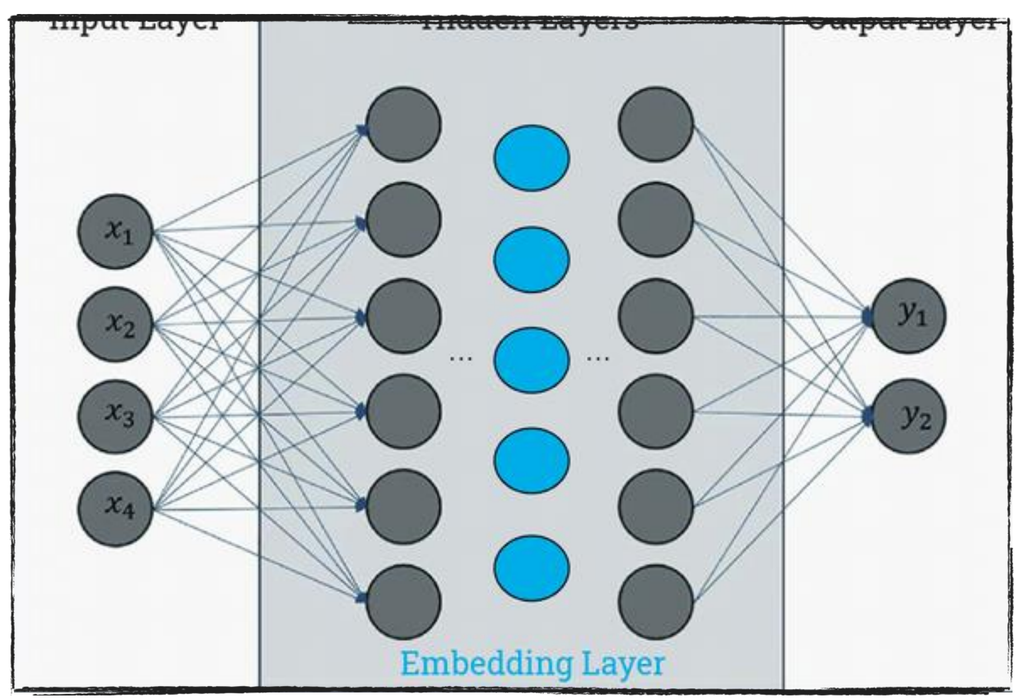
$$\ldots x_j)\Big)^2 + \boxed{\lambda \sum_{i=1}^{k} w_i^2}$$



GloVe Embedding Layer

Represents words as vectors in a high-dimensional space based on their co-occurrence probabilities.





Provides rich, pre-trained word representations, improving the LSTM model's ability to detect phishing URLs

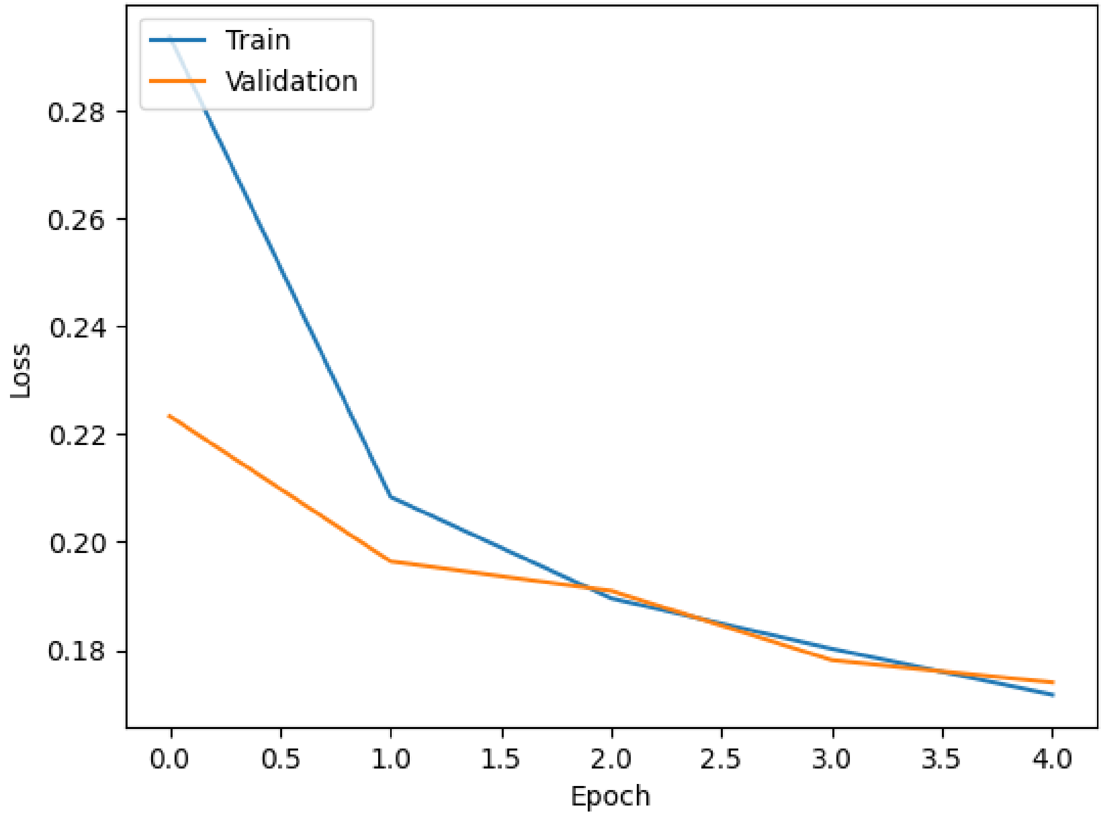
Training/Testing

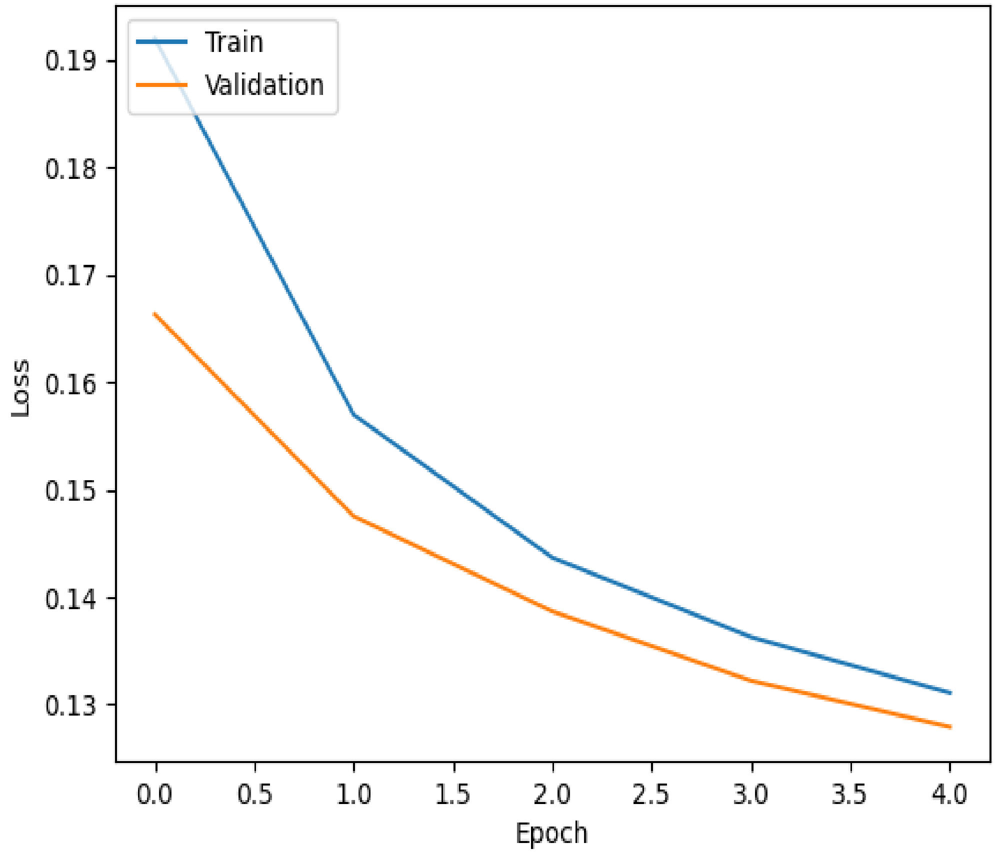
Training

- Hyper-Parameters
- Loss Plot

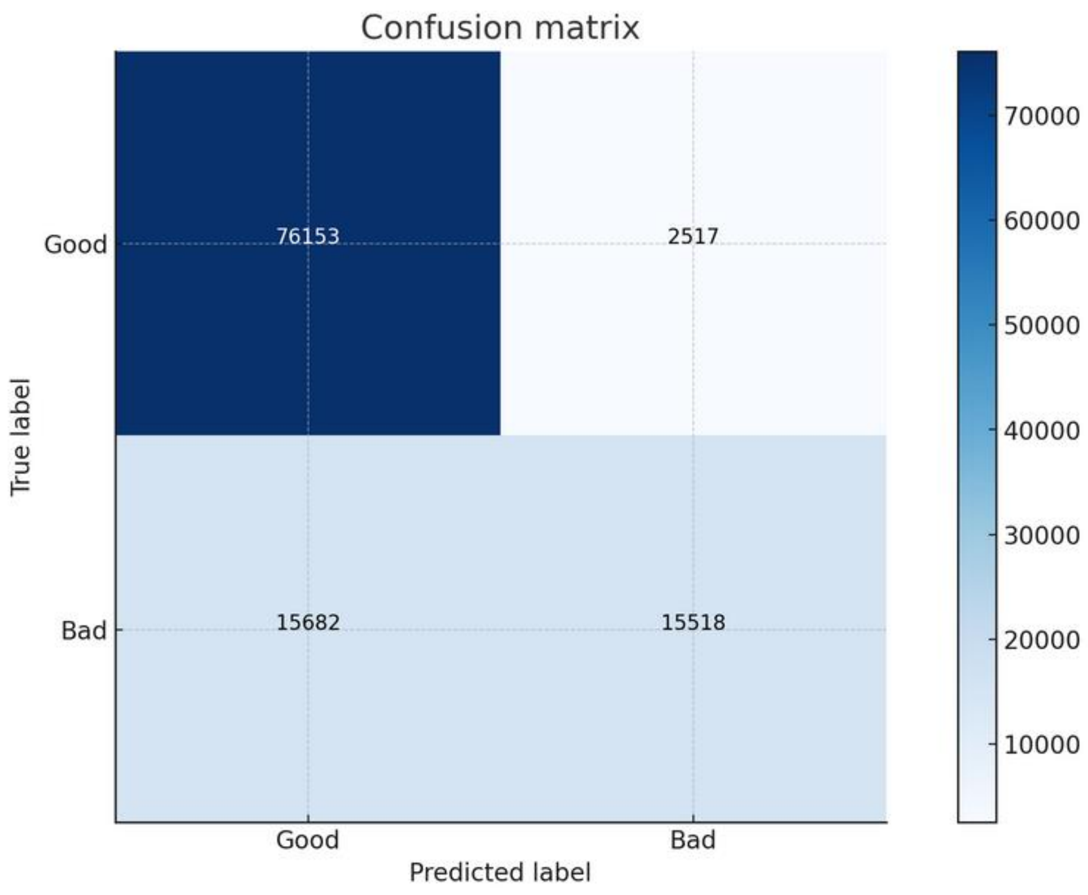
Testing

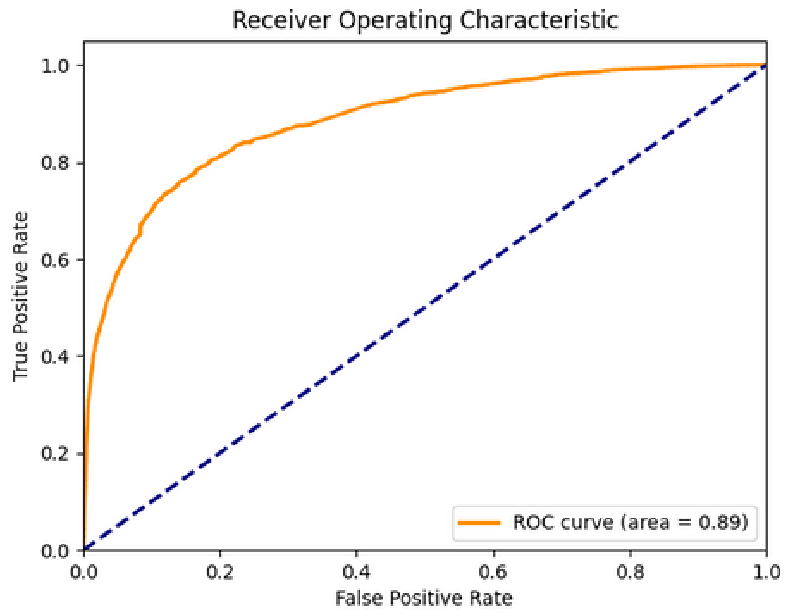
- Train/Test Split
- URL Examples

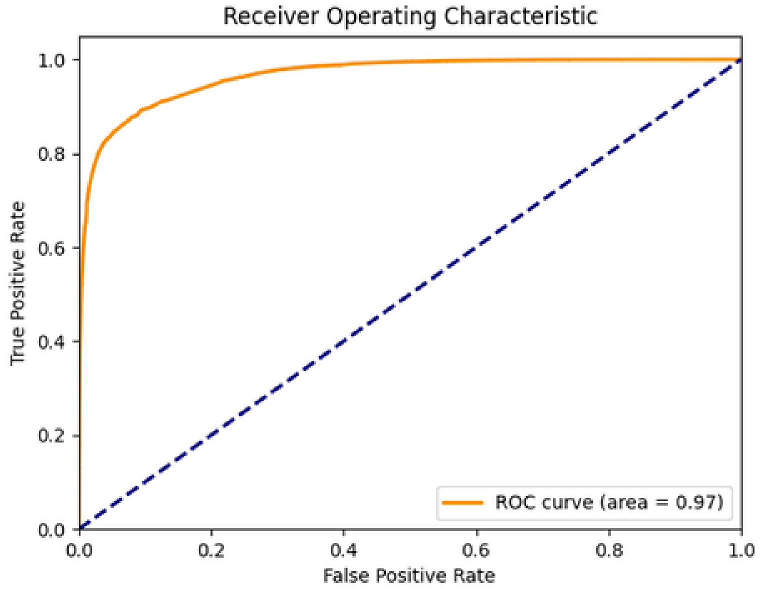


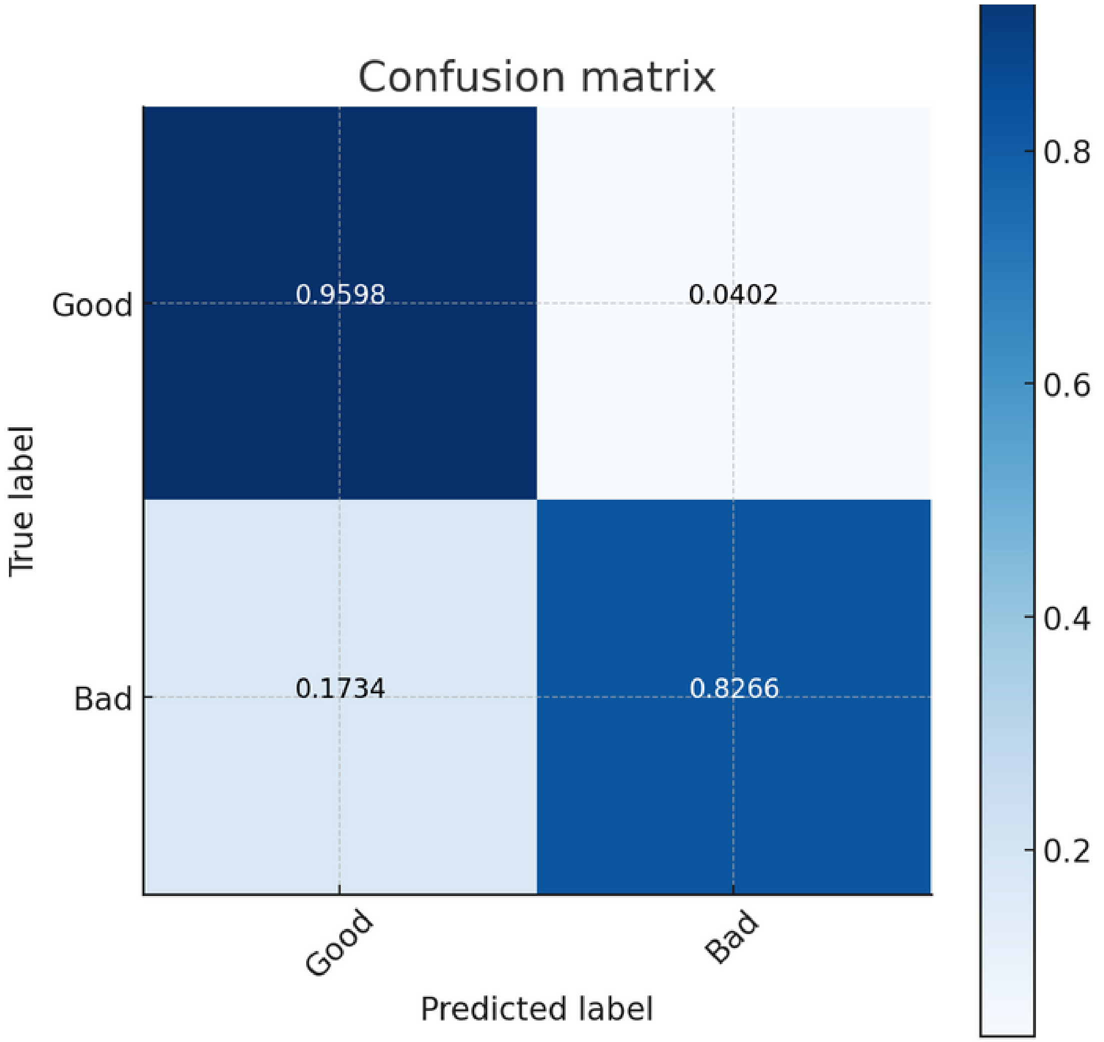


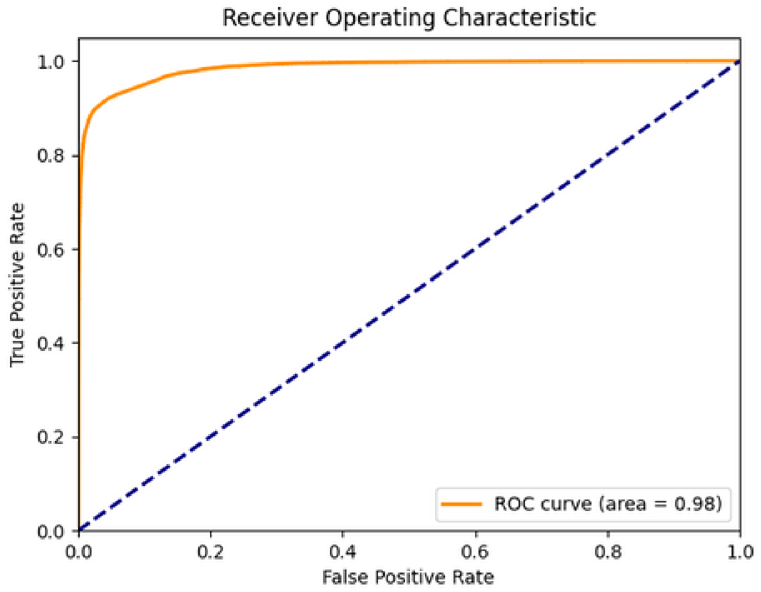
Project Results



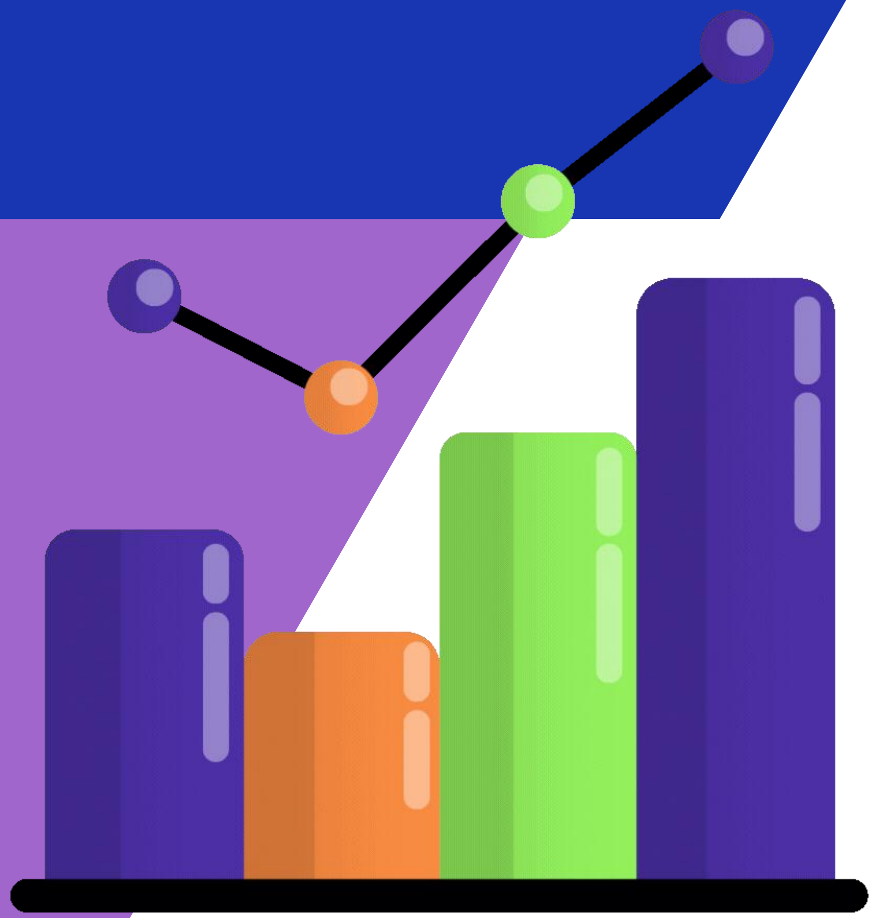








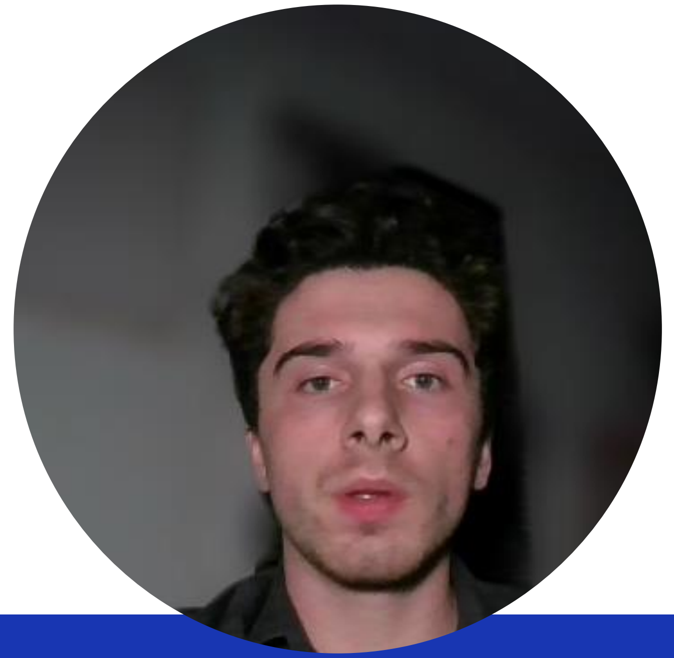
Project Results Continued



```
AUC-ROC: 0.9848680116227149
Confusion Matrix:
[[77314  1356]
 [ 3779 27421]]
Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.98      0.97     78670
           1       0.95      0.88      0.91     31200

   macro avg       0.95      0.93      0.94    109870
weighted avg       0.95      0.95      0.95    109870
```



Conclusion

Model Summary

- Model Architecture
- Best Accuracy Result
  - Model 2: 97.5% Accuracy
  - Base Model: 96% Accuracy

Problems Address

- Overcame limitations of traditional detection methods with advanced sequential data analysis.
- Improved Previous Analysis

Future Work

- Continue to Improve Model
- Train on New Datasets
- Real-Time Application

Thank you!