Enhancing URL Phishing Detection using GAN with LSTM and BERT Models

- By Adarsh Regulapati.
- First Advisor: Dr. Bimal Ghimire
- Second Advisor: Dr. Jeremy Blum
- Course: COMP 594 Master's Studies

PennState

Table of contents

1 Introduction

2 Literature Review

3 Data Set

4 Methodology

5 Results and Analysis

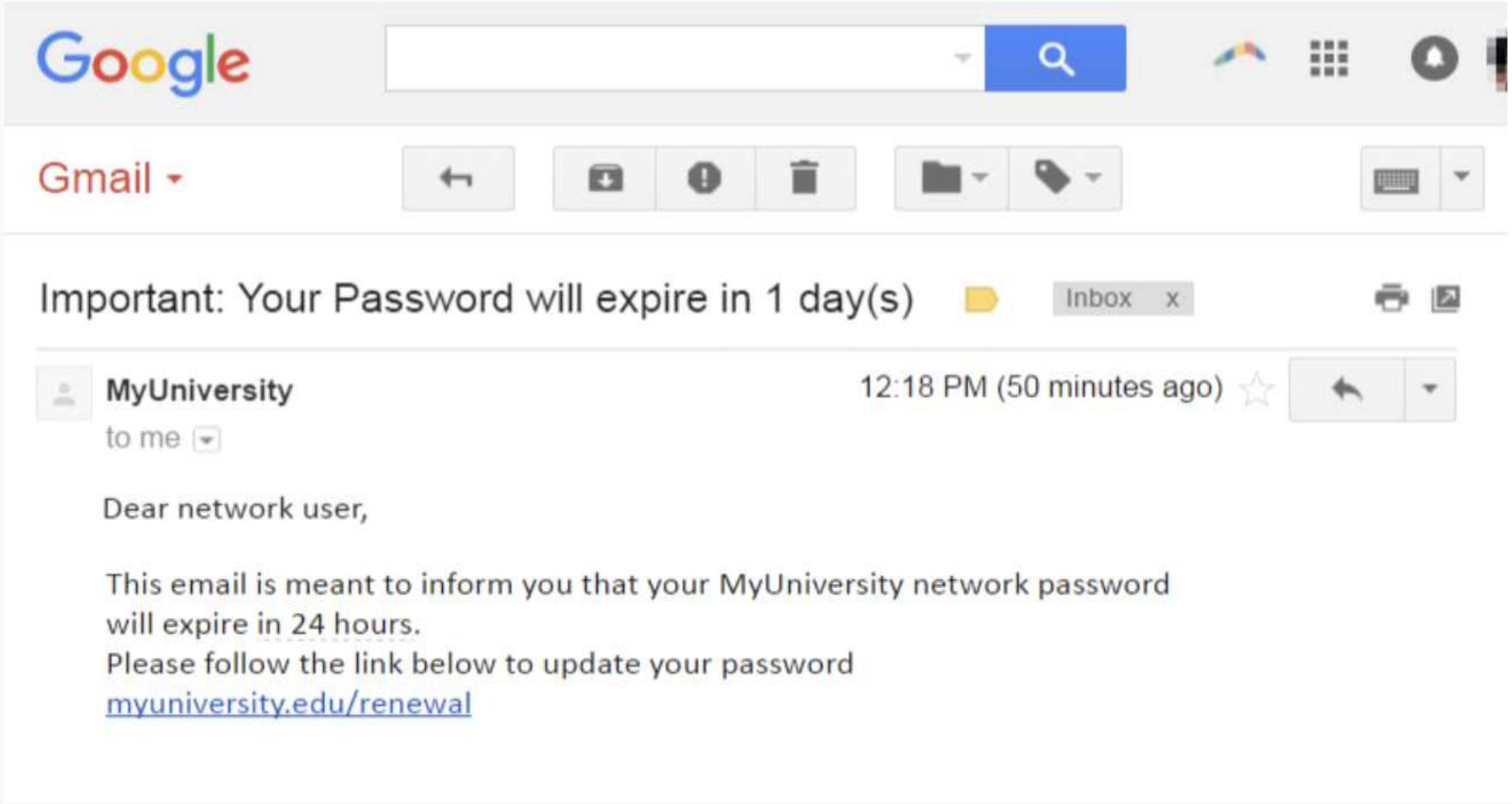
6 Conclusion

What is Phishing Attack

- Phishing is a type of cybersecurity attack during which malicious actors send messages pretending to be a trusted person or entity.
- Phishing attack manipulate a user, causing them to perform actions like installing a malicious file, clicking a malicious link, or divulging sensitive information such as access credentials.

What is URL Phishing

- URL phishing is a technique where a malicious actor sends a link that appears to be from a legitimate website but is actually a fake website.
- The goal of URL phishing is to trick the victim into providing sensitive information, such as login credentials, financial information, or banking details.


Google
Gmail
Important: Your Password will expire in 1 day(s)
Inbox x
MyUniversity
12:18 PM (50 minutes ago)
to me
Dear network user,
This email is meant to inform you that your MyUniversity network password will expire in 24 hours.
Please follow the link below to update your password
myuniversity.edu/renewal

Phishing Trend

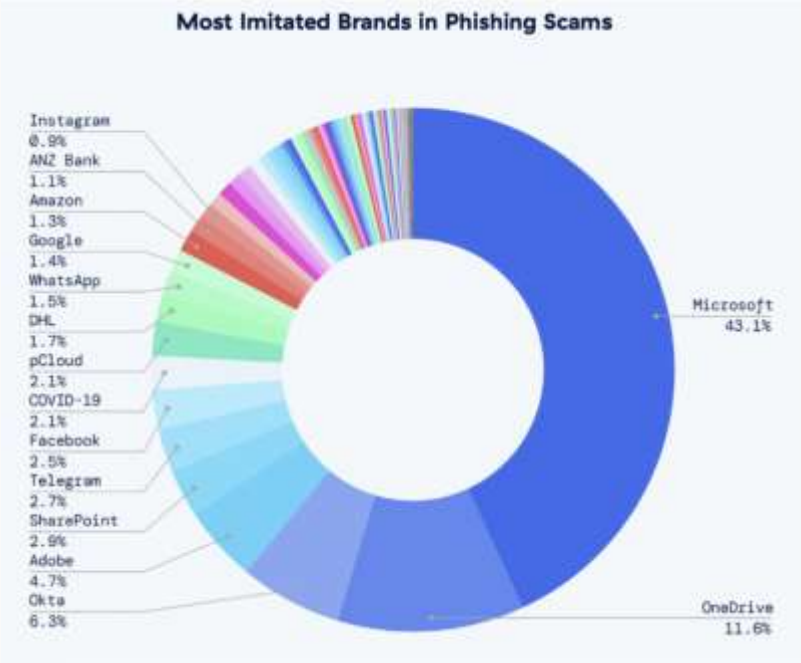
- Phishing attacks surged by 58.2% in 2023 compared to the previous year.[1]
- In Q3 of 2024, APWG observed 932,923 phishing attacks, up from 877,536 in the second quarter.[2]
- According to the report of Kaspersky, 85% of detected web threats are caused by malicious URLs. [3]
- 2024 Phishing Report reveals that HR and IT-related phishing emails claim a significant 48.6% share.[4]

Major challenges in detecting phishing URLs

- Balancing datasets for models can be challenging, as phishing URLs are typically much less frequent than legitimate ones. Overfitting on the majority class (legitimate URLs) can reduce the effectiveness of phishing detection.

- Dynamic Nature of URLs, URLs can change rapidly, and new phishing URLs emerge with new patterns. This makes it difficult to rely solely on blacklists, as they can quickly become outdated.

- Microsoft remains the most frequently imitated brand, with 43.1% of phishing attempts targeting it.



Problem Statement

Phishing detection remains a significant challenge due to rapidly evolving tactics and data imbalance; this project leverages the generative capabilities of GANs to create synthetic data, improving the accuracy and adaptability of phishing URL detection models.

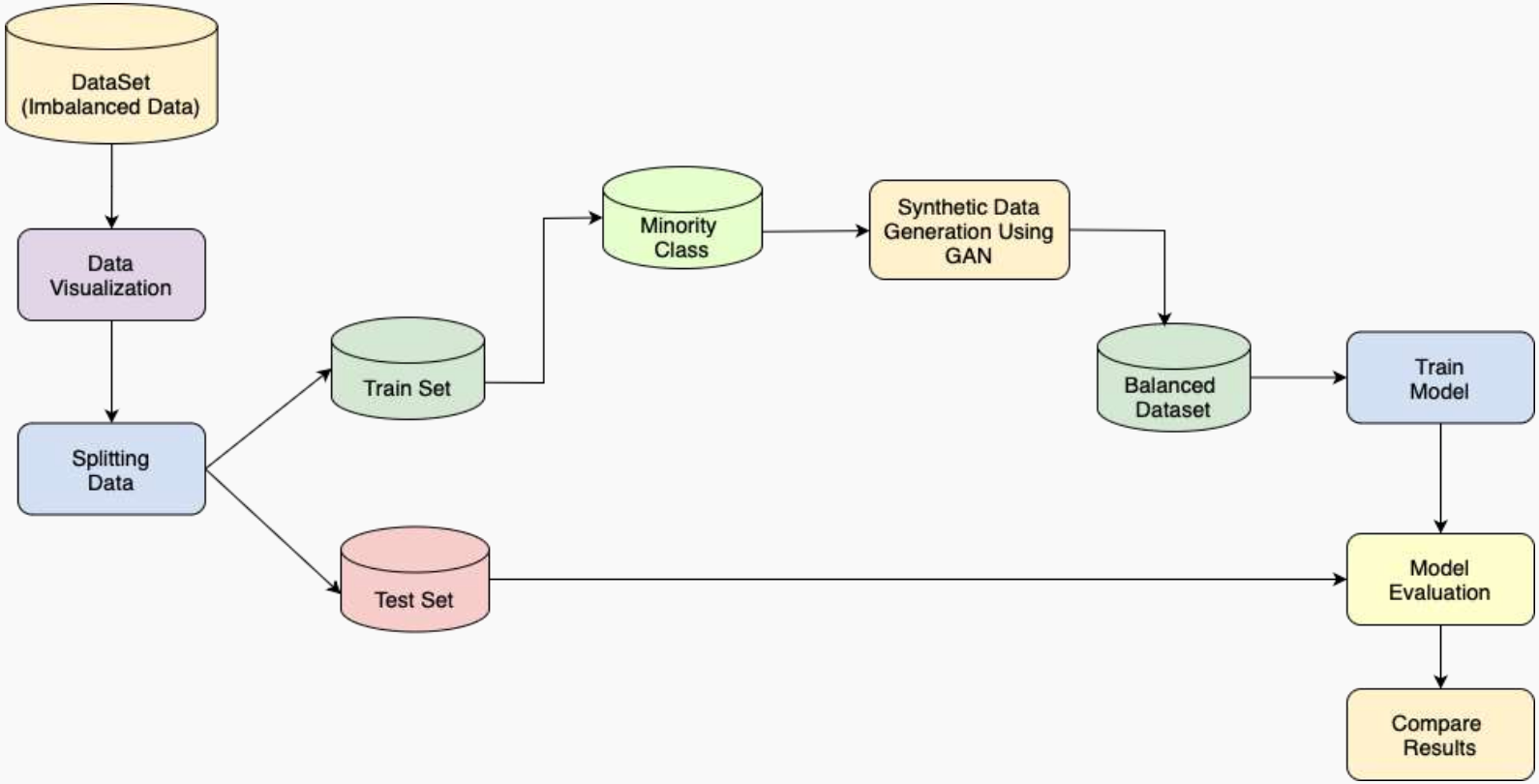
Literature Review

- BERT-Based Approaches to Identify Malicious URLs. [6]
- Bypassing Detection of URL-based Phishing Attacks Using Generative Adversarial Deep Neural Networks. [7]
- Generative adversarial network-based phishing URL detection with variational autoencoder and transformer. [9]

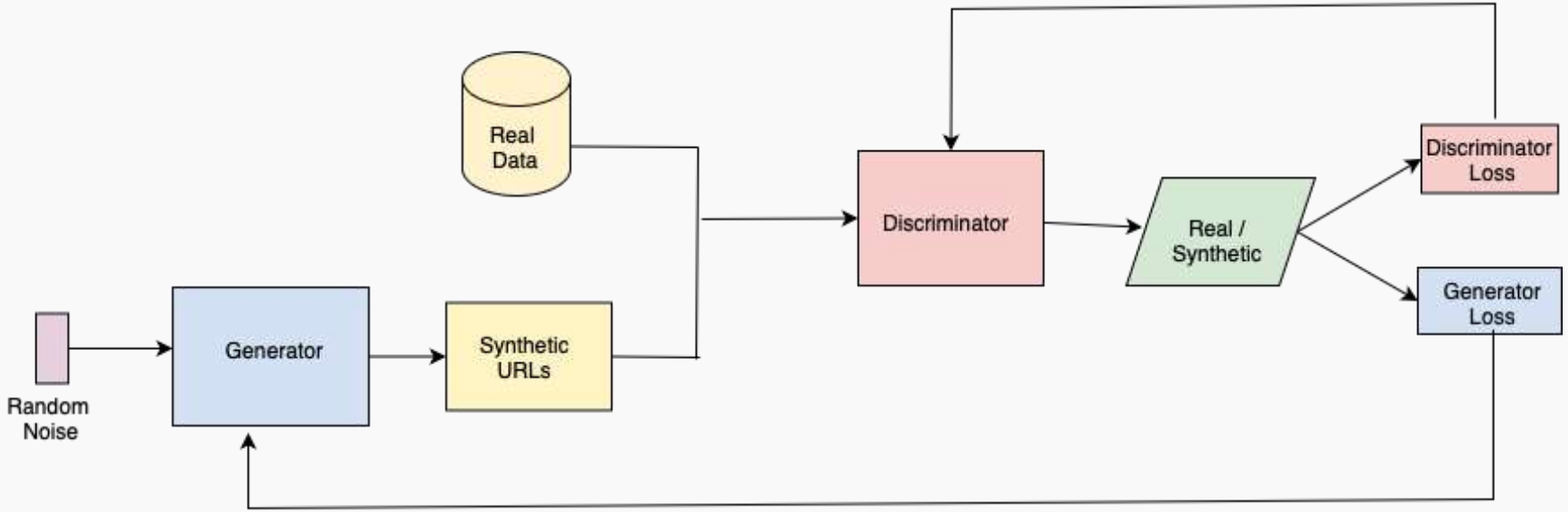
Data Collection

- I have taken Legitimate URL from Domcop and Common Crawl.
- I have taken top 100 thousand mostly used URLs.
- The phishing URLs are taken from different repositories like Phishtank, Openphish etc

Workflow of Model



GAN Architecture

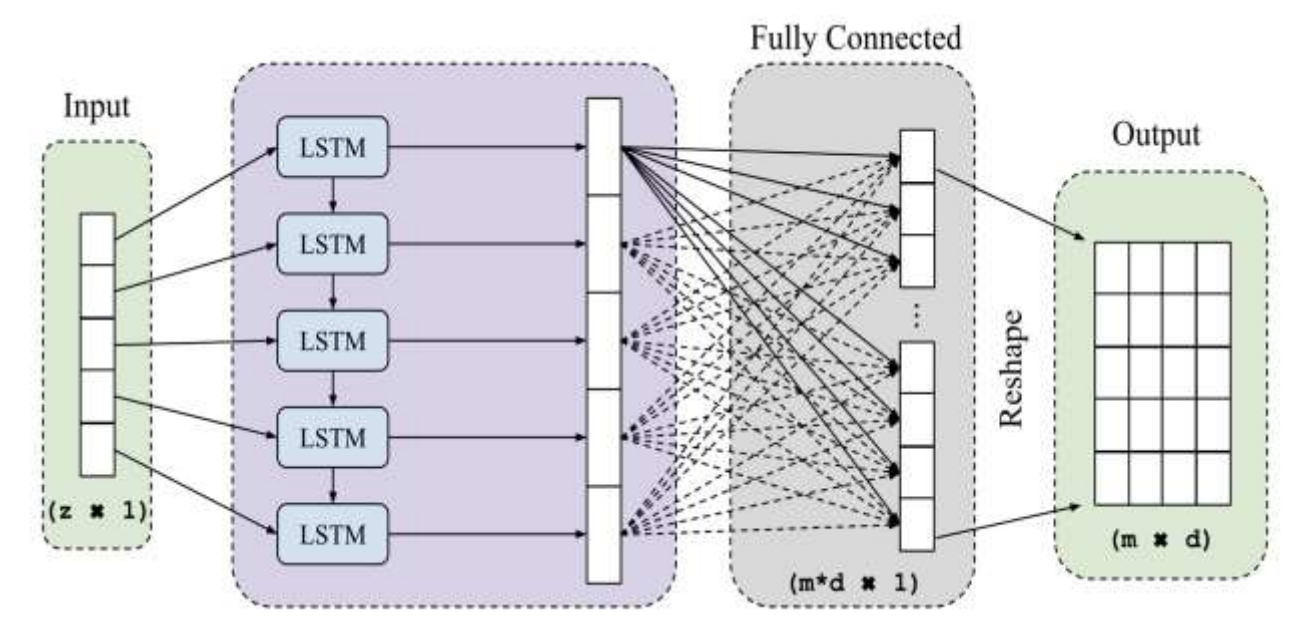


Generator

- The Generator is the core part of the GAN architecture responsible for generating synthetic URLs.
- Generators aim is to producing synthetic URL’s that closely resemble genuine ones.
- In this research we have used LSTM model for generator

LSTM Model as Generator

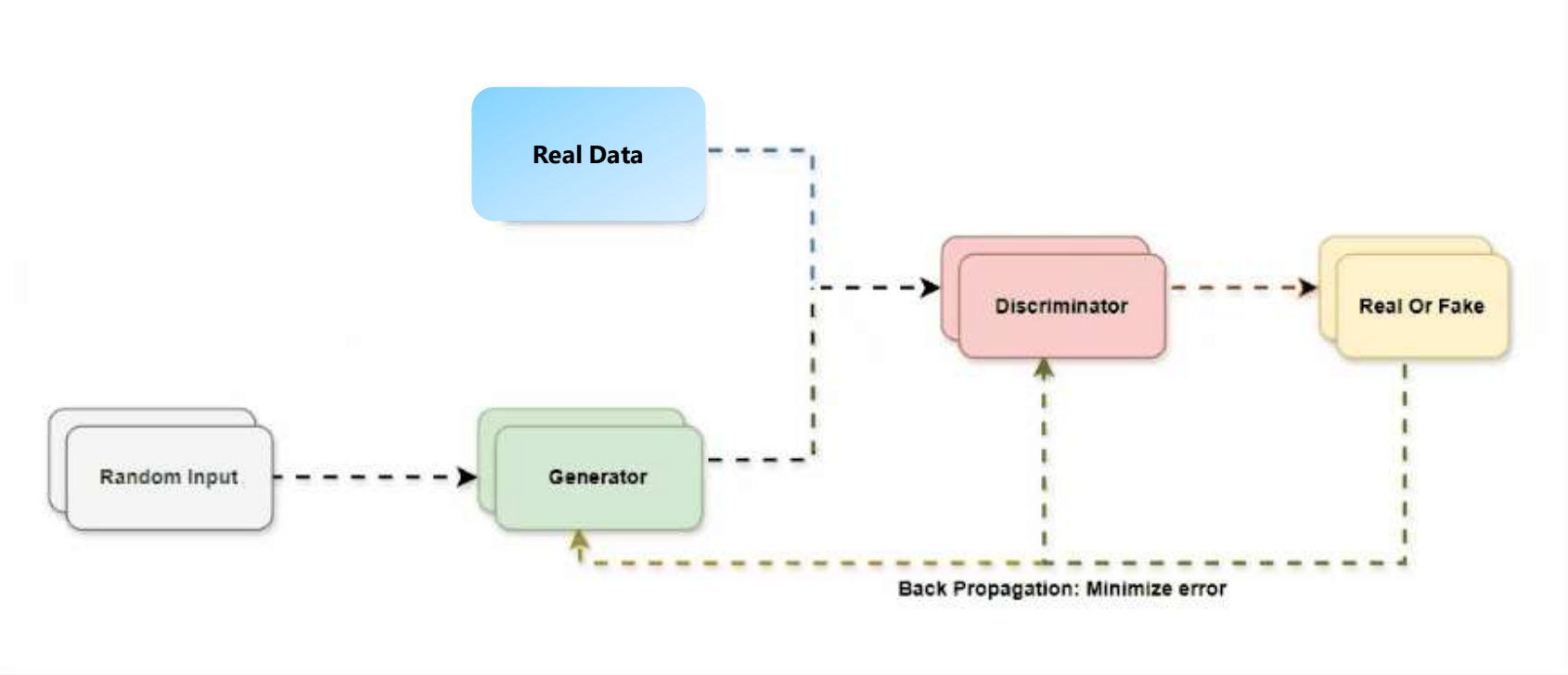
- Sequential Data
- Capturing Context
- Flexibility



Discriminator

- The Discriminator in the GAN is designed to distinguish between real and synthetic URLs.
- In this research we are using BERT model for Discriminator.
- The Discriminator's performance is evaluated using a loss function, which quantifies how well it differentiates real and synthetic URLs.
- The Generator and Discriminator adjusts its weights based on the feedback from the loss function to improve their performance.

Adversarial Learning

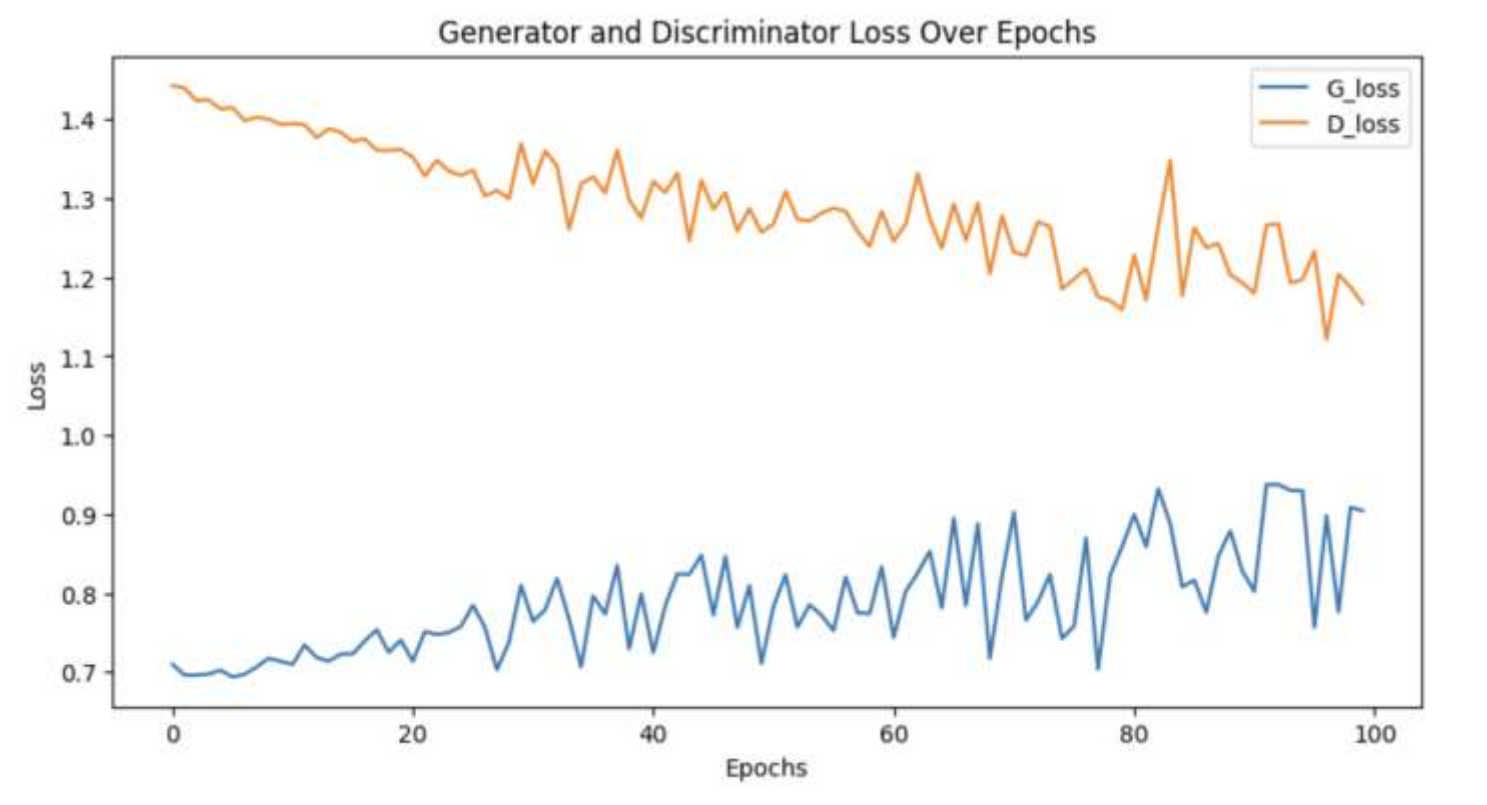


Adversarial Process

- The iterative feedback mechanism between the Generator and the Discriminator ensures that the Generator becomes better at creating realistic synthetic URLs while the Discriminator becomes better at distinguishing them from real URLs.

- This adversarial process continues until the Generator produces synthetic URLs that are indistinguishable from real URLs, achieving an optimal balance between the two models.

D-loss and G-loss Graphs



URL’s Generated by GAN Model

http://MTCY'IzWDnJbSFb55KokTX.stream/y/bltygwoodanuvyrbtacbbm
http://JDEsX.jD6omqKlPOHe03ue.top/dafxyrguxytndxmsbngxfult
https://LHpT2GdVmRukbeMULEaoap.space/ndamvkaxvvutwkocthrlxipe
ftp://I.bpdVoWDwwFw6bMMQMHec.men/ncmvfaktpnkefwimztazegpg
ftp://Q2rUxvBas5AjR++K6z1omE.space/wgbbzepp/sbbhpxsdsnilbfm
https://sT2KSpO6VR0gWEIFqRhine.win/zwndppsrfhawziyxiippxub
https://U5cKYIXRDIclO5Cz6BrA3P.xyz/fyykd/zwvbyskgfelsamkr
http://qA.9Pwh9EghjAz0.euqtz.men/kvlfalkredvewlyiyruuacgm
https://5XDAoKMvmrQTsu9ZNflI3s.racing/kdua/zaasdbynkeebsnwghx
ftp://uZjEgOqn2tyPiAgz22Tec.men/nf/dtloadwtomedzexlr/bd/
ftp://MNnE9'euwBfGS6wM'aLNJm.gq/xeedrorgtexeyiilpzlmmc/
ftp://Zb9Z1Xk+oRvqxMxVQ1NRYk.stream/wyvpmracpsvhmtwfmphcxbp
ftp://+aRYDBdA5oRgjOSrvV3qym.faith/i/ifpiphxfieecsvgixydizg
ftp://yXd6J0nDjtrBpUGjvU3sUp.top/yrx/cvohnp/ozkudc/xpead
ftp://rQz53AaWYjg1ZzruOHbmU.sci/x/tbcib/p/scxwfzmunpymz
ftp://Q.1O0XthqDeX0bZdo5yILa.stream/pffxilviyepdvxeixxhkcrcf
ftp://B2zg.GMCqPrFvXpyGjOWB.dcc/caznayxpereiasvlsrh/mlft
http://GhElL25BGrEdHtb3.oaRB.top/z/efopira/gpnvltdwegb
http://Ez+b+9RTCbUbZrUcQtU1.cric/bg/d/moakoce/ikgf/iuhxom

Balancing the Dataset with Synthetic URL’s

- The Generated URLs are added to the dataset with label phishing

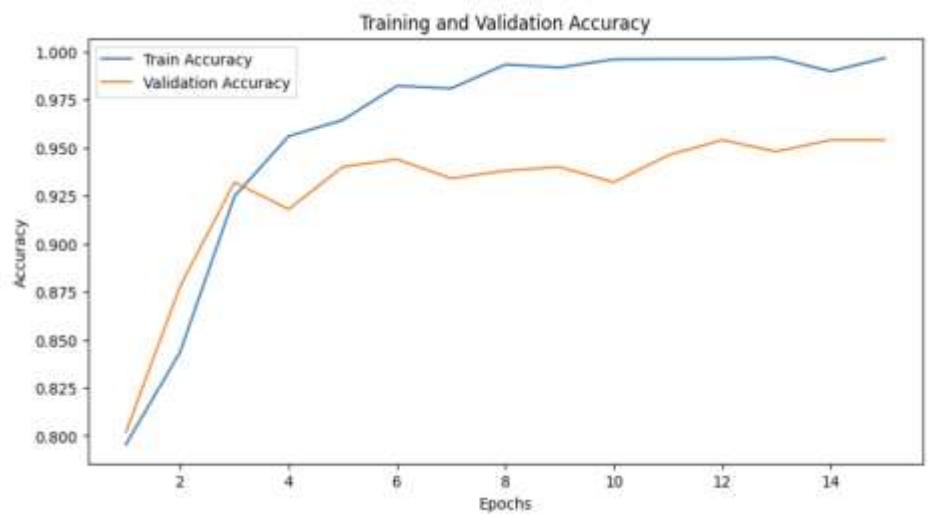
Using BERT Model to analyze the usefulness of Synthetic data

BERT Pre-Processing

- **Label Encoding**
  - The text labels (legitimate and phishing) are converted into numeric values using LabelEncoder
  - legitimate → 0, phishing → 1

- **Tokenization:**
  - The URLs are tokenized using the BERT tokenizer

Graphs of BERT



Metrics

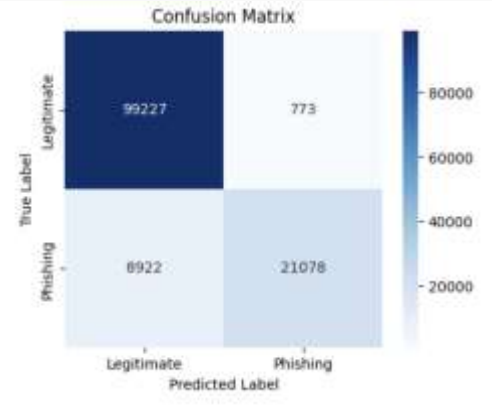
To evaluate the model performance, I used following metrics:

- Accuracy
- Precision
- Recall
- F1 score
- Confusion Matrix

BERT Results

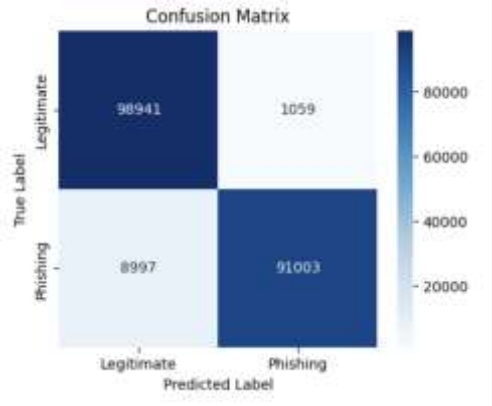
With Imbalance dataset:

- Testing Accuracy: 91.1%
- Precision: 0.97
- Recall: 0.84
- F1 score: 0.80

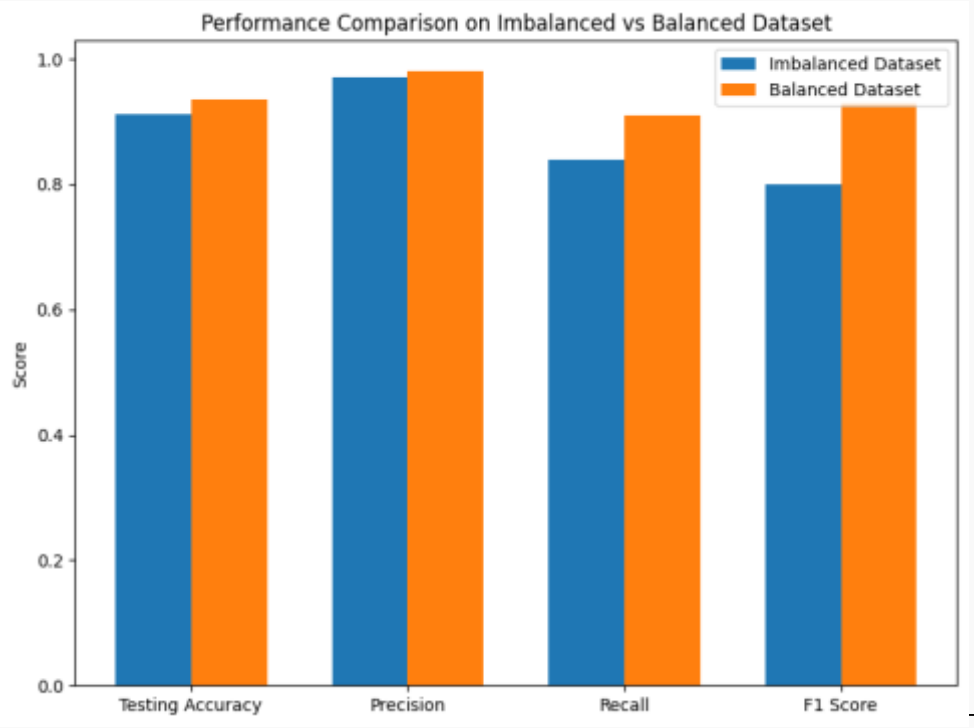


With balanced dataset

- Testing Accuracy: 94.62
- Precision: 0.98
- Recall: 0.91
- F1 score: 0.94



Comparison


Performance Comparison on Imbalanced vs Balanced Dataset
Imbalanced Dataset
Balanced Dataset
Score
1.0
0.8
0.6
0.4
0.2
0.0
Testing Accuracy
Precision
Recall
F1 Score

Conclusion

- The generated synthetic URLs are used to address the Imbalance dataset problem.
- The generator with LSTM model and discriminator with BERT model are performing well and giving real looking synthetic URLs.
- Using the recently discovered suspicious URLs we can generate synthetic data and keep the detection model updated.
- The balanced dataset is increasing the accuracy, precision, recall and F1 score.

Challenges and Future Work

- Achieving a balance where both models improve simultaneously can be challenging.
- The realistic nature of GAN-generated data can lead to unintended consequences, such as the potential misuse of synthetic data.
- Increase the efficiency of adversarial process.
- Using different GAN models for hostnames and path after hostnames.
- Develop a web-based application to generate synthetic data.

References

1. https://www.zscaler.com/blogs/security-research/phishing-attacks-rise-58-year-ai-threatlabz-2024-phishing-report
2. https://apwg.org/trendsreports/
3. https://www.kaspersky.co.uk/about/press-releases/malware-variety-grows-by-137-in-2019-due-to-web-skimmers
4. https://www.knowbe4.com/press/new-knowbe4-phishing-report-reveals-hr-and-it-related-emails-are-the-top-choices-for-phishing-scams
5. https://k21academy.com/ai-ml/gen-ai/generative-adversarial-networks/
6. https://www.mdpi.com/1424-8220/23/20/8499
7. https://dl.acm.org/doi/abs/10.1145/3375708.3380315?casa_token=qDA3A1ALiKwAAAAA:HwucSY3A7XUQBlV__d4-3bWH-ZFeDlRTd4eT_xrAqTub6puYfe9CQFSSot9gNRbfCmfXx0havxjj
8. https://ieeexplore.ieee.org/abstract/document/9585287?casa_token=NKMx0v9hyWYAAAAA:TwShwTc2mlknRtMbRpb23Qlhilrv5Lfc6rO45ip75m3SoA5VSyZF9iK3KYJ773db6uc3_oBB

References

9. https://www.researchgate.net/publication/380184574_Generative_adversarial_network-based_phishing_URL_detection_with_variational_autoencoder_and_transformer
10. https://www.astrill.com/blog/what-is-url-phishing/
11. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9965414

Thank You

- Adarsh Regulapati

PennState