Düzce Üniversitesi
Siber Güvenlik ve Dijital Ekonomi

Düzce University
Journal of Cybersecurity and Digital Economics
https://dergipark.org.tr/tr/pub/cdej

Araştırma Makalesi/Research Article

Sosyal Mühendislik Saldırılarının Modellemesi GAN Tabanlı E-Posta Üretimi ve Tespiti

Kubilay ATALAY[a,*]

[a]Karadeniz Teknik Üniversitesi, Trabzon, Türkiye.
[*]**Sorumlu Yazar/Corresponding author:** katalay@ktu.edu.tr

Makale Bilgileri/Article Information:
Geliş/Received: 26/05/2025, Kabul/Accepted: 16/06/2025.

ÖZET

Bu çalışma, phishing (oltalama) e-postalarının gerçekçiliğini artırmak ve savunma mekanizmalarını test etmek amacıyla, Generative Adversarial Network (GAN) tabanlı bir simülasyon yaklaşımı sunmaktadır. LSTM tabanlı bir generator ve BERT tabanlı bir discriminator kullanılarak, insan benzeri phishing e-postaları üretilmiş ve bu içeriklerin tespit araçları ve insan gözlemciler tarafından ayırt edilme başarısı analiz edilmiştir. 1000 dengeli e-posta içeren veri seti ile eğitilen model, %95'in üzerinde doğruluk oranıyla gerçek ve sahte içerikleri ayırt edebilmiştir. Sonuçlar, GAN mimarisinin sosyal mühendislik saldırılarını anlamada ve siber savunma sistemlerini test etmede güçlü bir araç olabileceğini göstermektedir.

Anahtar Kelimeler: *Phishing, Generative Adversarial Networks (GAN), LSTM, Sosyal Mühendislik*

Modeling of Social Engineering Attacks GAN Based Email Generation and Detection

ABSTRACT

This study presents a simulation approach based on Generative Adversarial Network (GAN) to increase the realism of phishing emails and test their defense mechanisms. Human-like phishing emails were generated using an LSTM-based generator and a BERT-based discriminator, and the success of distinguishing these contents by detection tools and human observers was analyzed. The model, trained on a dataset of 1000 balanced emails, was able to distinguish real and fake content with over 95% accuracy. The results show that GAN architecture can be a powerful tool in understanding social engineering attacks and testing cyber defense systems.

Keywords: *Phishing, Generative Adversarial Networks (GAN), LSTM, Social Engineering*

I. GİRİŞ

Dijitalleşmenin hızla yayılmasıyla birlikte, bireylerin ve kurumların internet üzerindeki faaliyetleri her geçen gün artmakta, bu da siber saldırıların etkisini ve çeşitliliğini önemli ölçüde artırmaktadır. Bu saldırılar arasında phishing, yani oltalama saldırıları, en yaygın ve zarar verici tehdit türlerinden biri olarak öne çıkmaktadır. Phishing, genellikle kullanıcıları sahte bir kimlikle kandırarak hassas bilgilerini (şifreler, kredi kartı bilgileri, kişisel veriler vb.) ele geçirmeye çalışan sosyal mühendislik tabanlı bir saldırı türüdür. Bu tür saldırılar, yalnızca bireyleri değil, aynı zamanda finansal kurumlar, sağlık hizmetleri, kamu kuruluşları ve çok uluslu şirketler gibi kritik sistemleri de hedef alarak büyük ölçekli güvenlik açıklarına neden olabilmektedir (Kavya et. all., 2025). Özellikle e-posta, phishing saldırılarının birincil aracı konumundadır. Siber saldırganlar, gerçekmiş gibi görünen e-postalar aracılığıyla kullanıcıları yönlendirmekte, zararlı yazılım içeren bağlantılara tıklamaya veya hassas bilgilerini paylaşmaya teşvik etmektedir. Bu nedenle phishing saldırılarını daha iyi anlamak ve bu tür

Atıf için: Atalay K. (2025). Sosyal Mühendislik Saldırılarının Modellemesi GAN Tabanlı E-Posta Üretimi ve Tespiti. *Düzce Üniversitesi, Siber Güvenlik ve Dijital Ekonomi*, 1(1): 10-21.

tehditlerle başa çıkmak için yeni nesil analiz ve simülasyon tekniklerine ihtiyaç duyulmaktadır (Albahadili et. all., 2024).

Phishing saldırılarına karşı geliştirilen savunma sistemleri, genellikle kurallara dayalı filtreleme mekanizmaları, anahtar kelime analizleri ve kara liste algoritmaları gibi geleneksel yöntemlere dayanmaktadır. Ancak saldırganlar, bu yöntemleri aşmak için gittikçe daha sofistike teknikler geliştirmektedir. Günümüzde sahte e-postalar yalnızca yazım tarzı bakımından değil, aynı zamanda görsel tasarım ve dilbilgisel yapı açısından da gerçek e-postalara oldukça benzeyebilmektedir (Gan et. all., 2024). Ayrıca saldırganlar, hedefe özel içerikler kullanarak (spear phishing) savunma sistemlerini daha da zor durumda bırakmaktadır. Bu bağlamda, klasik makine öğrenmesi modelleri de belirli örüntülere dayanarak karar verdiği için yeterince esnek ve dinamik bir çözüm sunamamaktadır. Öte yandan, savunma sistemlerinin eğitildiği veri setlerinin çoğu, gerçek dünya saldırılarını temsil etmekten uzaktır ve dolayısıyla sahadaki tehditleri algılamada yetersiz kalmaktadır. Bu durum, savunma sistemlerinin test edilmesi ve güçlendirilmesi için gerçekçi, insan benzeri ve tehdit içeriği taşıyan e-posta simülasyonlarına duyulan ihtiyacı artırmaktadır (Elberri et. all., 2024).

Generative Adversarial Network (GAN) mimarileri, son yıllarda özellikle yapay veri üretimi alanında önemli başarılar elde etmiş derin öğrenme teknikleri arasında yer almaktadır. GAN'lar, bir üretici (generator) ve bir ayırt edici (discriminator) modelden oluşur; üretici model gerçek veriye benzer örnekler üretmeye çalışırken, ayırt edici model bu örneklerin gerçek mi yoksa sahte mi olduğunu tahmin etmeye çalışır. Bu iki modelin rekabetçi eğitimi sonucunda, üretici giderek daha gerçekçi veriler üretme yeteneği kazanır. Özellikle metin üretimi, görsel oluşturma ve sentetik veri üretimi gibi alanlarda etkili olan GAN mimarileri, siber güvenlik alanında da son derece işlevsel bir potansiyele sahiptir (Nanda ve Goel, 2024). Metin tabanlı phishing e-postalarının üretimi için kullanılan GAN yapıları, sosyal mühendislik saldırılarını simüle edebilmekte ve güvenlik sistemlerinin zayıf yönlerini belirlemeye yardımcı olmaktadır. Bu tür modeller sayesinde, yalnızca saldırganların kullandığı yöntemler anlaşılmakla kalmaz, aynı zamanda savunma sistemlerinin eğitimi için daha çeşitli ve gerçekçi veri kümeleri oluşturulabilir. GAN mimarilerinin adaptif ve üretken yapısı, özellikle dinamik tehdit ortamlarında savunma stratejilerinin proaktif olarak geliştirilmesine olanak tanımaktadır (Kritika et. all., 2024).

Bu çalışma, yukarıda belirtilen zorluklara çözüm üretmeyi amaçlayan bir deneysel yaklaşım sunmaktadır. Çalışmanın temel amacı, GAN mimarisi kullanılarak insan benzeri phishing e-postalarının üretilmesi ve bu içeriklerin hem insanlar hem de tespit sistemleri tarafından ne kadar başarıyla ayırt edilebildiğinin değerlendirilmesidir. Bu bağlamda, LSTM tabanlı bir üretici model ile BERT tabanlı bir ayırt edici model birlikte eğitilmiş ve gerçek/sahte e-posta ayrımı üzerine yapılandırılmıştır. Çalışma, yalnızca modelin doğruluk ve başarı oranlarını rapor etmekle kalmamakta, aynı zamanda elde edilen çıktılar üzerinden sosyal mühendislik saldırılarının evrimsel doğasına dair nitel gözlemler sunmaktadır. Ayrıca çalışmada kullanılan veri setlerinin açık kaynaklı olması, yöntemin tekrarlanabilirliğini ve geliştirilebilirliğini artırmaktadır. Bu yönüyle çalışma, siber güvenlik eğitimi, savunma sistemlerinin test edilmesi ve gerçekçi tehdit simülasyonlarının oluşturulması açısından değerli bir kaynak olma potansiyeline sahiptir.

II. LİTERATÜR TARAMASI

Generative Adversarial Network (GAN) mimarileri, son yıllarda phishing saldırılarının simülasyonu ve tespiti alanında önemli bir araştırma konusu haline gelmiştir. GAN'lar, sahte phishing örnekleri üreterek tespit sistemlerinin eğitiminde kullanılan veri setlerini zenginleştirmekte ve böylece tespit performansını artırmaktadır. Örneğin, Kamran, Sengupta ve Tavakkoli (2021), yarı denetimli bir Conditional GAN modeli geliştirerek hem phishing URL'lerinin üretilmesini hem de tespit edilmesini sağlamışlardır. Bu model, saldırgan ve savunmacı arasındaki etkileşimi oyun teorisi perspektifinden ele alarak, gerçek zamanlı phishing URL tespiti için etkili bir yaklaşım sunmuştur. Benzer şekilde, Gayathri ve arkadaşları (2023), GAN tarafından üretilen sahte phishing örneklerini Random Forest Classifier (RFC) ile birleştirerek, tespit doğruluğunu artıran bir sistem önermişlerdir. Bu yaklaşım, GAN'ın veri üretme yeteneği ile geleneksel makine öğrenimi algoritmalarının sınıflandırma gücünü birleştirerek, daha etkili bir phishing tespiti sağlamıştır. Ayrıca, Pham ve arkadaşları

(2023), GAN tabanlı modellerin phishing URL sınıflandırıcıları üzerindeki performansını değerlendirmiş ve bu modellerin tespit doğruluğunu artırmada önemli katkılar sağladığını belirtmişlerdir.

Doğal dil işleme (NLP) alanında, Long Short-Term Memory (LSTM) ve Bidirectional Encoder Representations from Transformers (BERT) modelleri, metin tabanlı phishing tespiti çalışmalarında yaygın olarak kullanılmaktadır. LSTM, özellikle sıralı verilerde uzun vadeli bağımlılıkları öğrenme yeteneği sayesinde, phishing e-postalarının dilsel özelliklerini analiz etmekte etkili olmuştur. Örneğin, Fang ve arkadaşları (2019), Bi-LSTM tabanlı bir model olan THEMIS'i geliştirerek, dengesiz veri setlerinde %99.84 doğruluk oranı elde etmişlerdir. Bu model, phishing e-postalarının tespitinde yüksek performans sergilemiştir. Öte yandan, BERT modeli, çift yönlü bağlam anlayışı sayesinde, kelimelerin anlamını daha derinlemesine analiz edebilmekte ve bu da phishing tespitinde önemli avantajlar sağlamaktadır. Thapa ve arkadaşları (2023), BERT tabanlı modellerin phishing e-postalarının tespitinde yüksek doğruluk oranlarına ulaştığını ve geleneksel yöntemlere kıyasla daha üstün performans sergilediğini belirtmişlerdir. Bu çalışmalar, LSTM ve BERT'in metin tabanlı phishing tespitinde önemli araçlar olduğunu göstermektedir.

Phishing tespiti üzerine yapılan çalışmalar, zamanla çeşitli yöntemlerin geliştirilmesine yol açmıştır. Geleneksel olarak, kara liste tabanlı ve kural tabanlı sistemler kullanılmış olsa da, bu yöntemler yeni ve bilinmeyen saldırıları tespit etmede yetersiz kalmaktadır. Bu nedenle, makine öğrenimi ve derin öğrenme tabanlı yaklaşımlar ön plana çıkmıştır. Abdolrazzagh-Nezhad ve Langarib (2021), phishing tespiti tekniklerini anti-phishing araçlar, sezgisel yaklaşımlar, makine öğrenimi tabanlı teknikler ve meta-sezgisel algoritmalar olarak sınıflandırmış ve her birinin etkinliğini analiz etmişlerdir. Ayrıca, Castaño ve arkadaşları (2021), içerik tabanlı ve hibrit phishing tespiti yöntemlerini incelemiş ve bu yöntemlerin phishing sitelerinin tespitinde önemli katkılar sağladığını belirtmişlerdir. Bu çalışmalar, phishing tespiti alanında çeşitli yöntemlerin geliştirildiğini ve her birinin belirli avantajlar sunduğunu göstermektedir.

III. MATERYAL METOT

A. Veri Seti Hazırlığı

Phishing e-posta üretimi ve sınıflandırılması için kullanılacak modelin başarısı, büyük ölçüde eğitildiği veri setinin kalitesi ve çeşitliliğine bağlıdır. Bu çalışmada, dengeli ve zengin içerikli bir veri seti oluşturmak amacıyla iki farklı açık kaynak platformundan veri derlenmiştir: Hugging Face ve Kaggle. Spesifik olarak, ealvaradob/phishing-dataset veri kümesi Hugging Face üzerinden, naserabdullahalam/phishing-email-dataset veri kümesi ise Kaggle platformu üzerinden temin edilmiştir. Bu veri kümeleri, phishing saldırılarını temsil eden e-posta örnekleri ile meşru (legit) e-posta örneklerini dengeli bir biçimde içermektedir. Oluşturulan nihai veri seti, toplamda 1000 adet e-posta içermektedir. Bu e-postaların 500'ü phishing saldırılarına aitken, diğer 500'ü yasal ve güvenli kurumsal iletişim örneklerinden oluşmaktadır. Her bir e-posta, aşağıdaki yapısal bileşenleri içerecek şekilde düzenlenmiştir:

- E-posta başlığı (subject line)
- İçerik metni (body)
- Gönderen bilgileri (sender info)
- Ek içerikler ve bağlantılar (attachments, URLs)

Bu yapı, hem metinsel hem de bağlamsal bilgilerin modellenmesine olanak sağlayarak GAN mimarisinin daha gerçekçi e-posta üretmesini mümkün kılmaktadır. Ham veri kümesi doğrudan modellemeye uygun olmadığından, çeşitli doğal dil işleme (NLP) teknikleri kullanılarak kapsamlı bir ön işleme süreci uygulanmıştır. Bu sürecin temel amacı, metin verisini daha temiz, düzenli ve anlamlı hâle getirmektir (Aden et.all., 2024). Uygulanan adımlar şu şekildedir:

- ***HTML ve Özel Karakter Temizliği*** E-postaların içeriğinde sıklıkla karşılaşılan <html>,
,   gibi işaretler ve özel karakterler kaldırılmıştır. Bu işlem, gereksiz formatlama bilgilerinin modele etki etmesini önlemek amacıyla gerçekleştirilmiştir.
- ***Küçük Harfe Dönüştürme*** Büyük/küçük harf duyarlılığından kaynaklı kelime tekrarlarını engellemek için tüm metinler küçük harfe dönüştürülmüştür.
- ***Noktalama İşaretlerinin Kaldırılması*** Nokta, virgül, ünlem gibi noktalama işaretleri, dil modellemesi açısından anlam taşımayan türden olduğunda kaldırılmıştır. Bu işlem, kelime frekanslarının daha net analiz edilmesini sağlamıştır.
- ***Stop-word Temizliği*** İngilizce dilinde yaygın olarak kullanılan ve anlam taşımayan "the", "is", "at", "which" gibi stop-word'ler çıkarılmıştır. Bu adım, metnin bilgi yoğunluğunu artırmaya yöneliktir.
- ***Lemmatizasyon (Kökleme)*** Kelimelerin temel biçimlerine (örneğin, "running" → "run") indirgenmesi sağlanmıştır. Bu işlem, bağlamdan bağımsız dil modeli oluşturulmasına yardımcı olur.
- ***Tokenizasyon*** E-posta içerikleri, model girdi formatına dönüştürülebilmesi için kelime ve cümle bazında token'lara ayrılmıştır. Token sayısı, modelin giriş uzunluğunu aşmayacak biçimde sınırlanmıştır (maksimum sekans uzunluğu = 128 token).
- **Görsel İçeriklerin Dönüştürülmesi (isteğe bağlı)** Bazı phishing e-postaları görsel öğeler (örneğin sahte CAPTCHA görselleri) içermektedir. Bu içerikler metin açıklamasına dönüştürülerek ("image-based warning", "login button image" vb.) metne entegre edilmiştir.

Bu ön işleme süreci sonucunda, GAN mimarisine uygun, anlamlı ve temsili bir metin kümesi elde edilmiştir. Temizlenmiş ve yapısal olarak dengelenmiş bu veri seti, hem üretici (generator) modelin gerçekçi phishing e-postaları üretmesini hem de ayırt edici (discriminator) modelin gerçek ve sahte veriler arasında ayrım yapmasını kolaylaştırmıştır.

B. GAN Mimarisi

Phishing e-postalarının gerçekçi biçimde üretilebilmesi ve bu e-postaların gerçek ya da sahte olduğunun güvenilir biçimde ayrıştırılabilmesi için bu çalışmada Generative Adversarial Network (GAN) mimarisi uygulanmıştır. GAN'lar, üretici (generator) ve ayırt edici (discriminator) olmak üzere iki bileşenden oluşur ve bu iki bileşen rekabetçi bir biçimde eğitilerek birbirlerini optimize eder. Bu çalışmada metin tabanlı veri üretimi hedeflendiğinden, generator modeli olarak sıralı verilerde dilsel bağımlılıkları öğrenebilen Long Short-Term Memory (LSTM) mimarisi; discriminator modeli olarak ise metin sınıflandırmada yüksek başarı oranlarına sahip olan Bidirectional Encoder Representations from Transformers (BERT) mimarisi tercih edilmiştir. Generator bileşeni, rastgele gürültü vektörlerinden (noise vector) başlayarak insan benzeri phishing e-posta içerikleri üretmeyi hedeflemektedir (Boukhris ve Zaâbi,.2024). Bu üretici modelin temelini, sıralı verilerde uzun vadeli bağıntıları öğrenme kapasitesi ile bilinen LSTM (Long Short-Term Memory) ağı oluşturmaktadır. Model, temel olarak aşağıdaki katmanlardan oluşmaktadır:

- ***Embedding Layer:*** Girdi olarak verilen token ID'lerini kelime vektörlerine dönüştürür. Bu vektörler, kelimeler arasındaki anlamsal ilişkilerin öğrenilmesini sağlar.
- ***LSTM Layer:*** Kelime sıralamaları arasındaki bağlamları öğrenir ve bu bağıntılara göre dil üretimi gerçekleştirir.
- ***Dense Layer (Softmax):*** Çıkış katmanında yer alan bu yapı, her bir adımda en olası kelimeyi seçerek çıktı cümlesini oluşturur.

Modelin çıktısı, kelime dağılımına karşılık gelen bir olasılık dizisidir ve bu diziden örnekleme yapılarak e-posta metni oluşturulur. Eğitim sürecinde, generator'un amacı, discriminator tarafından "gerçek" olarak sınıflandırılacak kadar ikna edici metinler üretmektir. Discriminator bileşeni, verilen bir e-posta metninin gerçek bir kullanıcıdan mı geldiğini yoksa generator tarafından mı üretildiğini sınıflandırmakla görevlidir. Bu amaç doğrultusunda, doğal dil anlayışında yüksek bağlamsal çözümleme yapabilen BERT (Bidirectional Encoder Representations from Transformers) mimarisi kullanılmıştır. Bu model, e-postaları çift yönlü bağlam içerisinde analiz eder ve sınıflandırma katmanına aktarır. Model, bert-base-uncased önceden eğitilmiş ağırlıklar üzerine

inşa edilmiş olup, son katmanda ikili sınıflandırma yapacak şekilde yapılandırılmıştır. Bu yapı sayesinde, discriminator hem phishing hem de meşru e-postalar arasındaki dil farklılıklarını öğrenmekte ve generator tarafından üretilmiş içerikleri tespit etmeye çalışmaktadır. GAN bileşenlerinin eğitimi sırasında çeşitli hiperparametreler optimize edilmiştir. Bu parametreler, modelin öğrenme başarısı üzerinde doğrudan etkili olduğundan dikkatle seçilmiştir. Tablo 1'de kullanılan eğitim parametreleri verilmektedir.

Tablo 1. Eğitim parametreleri.

Parametre	**Değer**	**Açıklama**
Batch Boyutu	16	Her bir eğitim iterasyonunda işlenen örnek sayısı
Epoch sayısı	10	Tam veri seti üzerinden geçilen toplam tur sayısı
Öğrenme Oranı	5e-5	Ağırlık güncelleme hızı
Maksimum Sekans Uzunluğu	128	Token sayısına göre giriş verisinin uzunluk sınırı
Optimizasyon Algoritması	Adam	Momentuma dayalı optimize edici algoritma
Kayıp fonksiyonu- dicriminator	Binary Cross Entropy	Gerçek/sahte ayrımı için kullanılmaktadır.
Kayıp fonksiyonu- Generator	Geri besleme kaybı(feedback loss)	Discriminator yanıtına göre optimize edilir.

Tablo 1'deki parametreler hem LSTM hem de BERT tabanlı modellerin istikrarlı biçimde öğrenmesini sağlamış ve GAN mimarisinin eğitim sürecinde kararlılık elde edilmiştir. Eğitim döngüsü boyunca, önce discriminator eğitilmiş, ardından generator discriminator'u kandırabilecek şekilde güncellenmiştir. Bu süreç her epoch'ta sıralı biçimde tekrarlanmıştır.

C. Eğitim Süreci

Generative Adversarial Network (GAN) eğitimi, birbirine rakip iki yapay sinir ağı modelinin — bir üretici (generator) ve bir ayırt edici (discriminator) — dönüşümlü olarak eğitilmesine dayanır. Bu rekabetçi yapı, her iki modelin de zamanla performansını artırmasını sağlayan bir denge mekanizması oluşturur. Bu çalışmada uygulanan GAN eğitim süreci, belirli adımlar çerçevesinde yapılandırılmıştır.

İlk adımda, discriminator modeli yalnızca gerçek e-posta örnekleri ile eğitilerek, sahte içerikleri tanıma yeteneği kazanmadan önce gerçek verilerin dağılımını öğrenmesi sağlanır. Bu, discriminator'un eğitimin başında sahte veriler tarafından kolay kandırılmamasına yönelik bir ön ısınma (warm-up) evresidir. Daha sonra generator tarafından rastgele gürültü vektörlerinden üretilen sahte e-posta içerikleri oluşturulur ve discriminator'a gerçek verilerle birlikte sunulur. Bu noktadan itibaren discriminator, her bir girdiyi "gerçek" veya "sahte" olarak sınıflandırmayı öğrenmeye başlar. Sonraki aşamada, generator eğitime dahil edilir. Generator'un amacı, discriminator'un karar mekanizmasını yanıltacak kadar gerçekçi içerikler üretmektir. Böylece, discriminator'un sınıflandırma doğruluğunu düşürmeye çalışır. Ancak discriminator da eş zamanlı olarak sahte verileri daha iyi ayırt etmek üzere eğitilmeye devam eder. Bu şekilde iki model arasında bir öğrenme döngüsü oluşur ve modeller her iterasyonda birbirlerine karşı daha güçlü hale gelir.

GAN mimarisinde kullanılan kayıp fonksiyonları, modelin öğrenme başarısının temel belirleyicilerindendir. Bu çalışmada, discriminator ve generator için farklı kayıp fonksiyonları uygulanmıştır. Discriminator için, genellikle kullanılan Binary Cross Entropy (BCE) kayıp fonksiyonu tercih edilmiştir. Bu fonksiyon, modelin tahmin ettiği olasılıkla gerçek etiket (0: sahte, 1: gerçek) arasındaki farkı minimize eder. Discriminator, gerçek e-postaları "1", üretilmiş e-postaları ise "0" olarak etiketlemeye çalışır ve her tahminden sonra ağırlıklarını buna göre günceller. BCE kaybı Eşitlik 1'deki gibi tanımlanmaktadır.

$$\mathcal{L}_D = -\mathbb{E}_{x\sim p_{\text{real}}}[\log D(x)] - \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z)))] \tag{1}$$

Generator için kullanılan kayıp fonksiyonu ise, discriminator'un değerlendirme sonuçlarına dayanan ters BCE formunda geri besleme kaybıdır. Yani, generator'un hedefi, discriminator'un ürettiği sahte örnekleri "gerçek" olarak sınıflandırmasını sağlamaktır. Bu amaçla kullanılan kayıp fonksiyonu Eşitlik 2'de verilmektedir.

$$\mathcal{L}_G = -\mathbb{E}_{z\sim p_z}[\log D(G(z))] \tag{2}$$

Bu yapı sayesinde generator, her iterasyonda discriminator'un karar verme eşiğini aşacak kadar gerçekçi içerikler üretmeye motive edilir. Her iki kayıp değeri de eğitim süresince izlenmiş ve öğrenme eğrileri analiz edilmiştir. GAN eğitimi sırasında belirli sayıda epoch boyunca tekrarlanan bir döngü uygulanır. Her epoch, hem discriminator hem de generator için ayrı ayrı güncelleme adımlarını içerir. Bu çalışmada kullanılan eğitim döngüsü genel olarak şu sırayla gerçekleştirilmiştir:

Discriminator Eğitimi:

Gerçek e-posta örnekleri veri setinden alınarak modele sunulur.
Ardından generator tarafından üretilmiş sahte e-postalar oluşturulur.
Bu iki veri kümesi birleştirilerek etiketlenmiş biçimde discriminator'a giriş olarak verilir.
Binary Cross Entropy kaybı hesaplanarak discriminator'un ağırlıkları güncellenir.

Generator Eğitimi:

Rastgele gürültü vektörleri üretilir.
Generator bu vektörleri anlamlı cümlelere dönüştürür (yani sahte e-posta üretir).
Bu içerikler discriminator'a sunulur.
Discriminator'un geri bildirimi temel alınarak, generator'un ağırlıkları güncellenir. Hedef, discriminator'un yanılmasını sağlamaktır.

Bu eğitim yapısı her epoch boyunca yinelenmiş, 10 epoch boyunca her bir batch için kayıp değerleri kayıt altına alınmıştır. Başlangıçta discriminator kaybı yüksek gözlemlenirken, ilerleyen epoch'larda hızlıca düşerek öğrenme eğrisinin dengelendiği görülmüştür. Öte yandan, generator'un kayıp değeri daha stabil seyretmiş, bu da üretici modelin discriminator'u kandırma becerisinin zamanla geliştiğini göstermiştir. Nihai durumda, discriminator %95'in üzerinde doğruluk oranına ulaşmış ve modelin kararlılığı ispatlanmıştır.

IV. BULGULAR VE TARTIŞMA

GAN mimarisi ile geliştirilen modelin etkinliğini değerlendirmek amacıyla çeşitli doğruluk metrikleri kullanılmıştır. Discriminator'un doğruluğu, precision (kesinlik), recall (duyarlılık) ve F1 skoru gibi klasik sınıflandırma başarı ölçütleri üzerinden analiz edilmiştir. Bu metrikler, modelin hem gerçek hem de sahte e-postaları ne derece doğru şekilde ayırt edebildiğini sayısal olarak ortaya koymaktadır. Tablo 2'de modelin performans metrikleri sunulmuştur.

Tablo 2. Model performans metrikleri.

Metrik	**Değer (%)**
Accuracy	95.2
Prescision	93.8
Recall	96.5
F1 Score	95.1

Tablo 2'deki sonuçlara göre, discriminator modeli oldukça başarılı bir sınıflayıcı olarak çalışmış ve GAN mimarisinin phishing tespiti için güçlü bir temel oluşturduğunu göstermiştir. Özellikle yüksek recall oranı, modelin sahte (phishing) e-postaları kaçırma olasılığının düşük olduğunu ortaya koymaktadır.

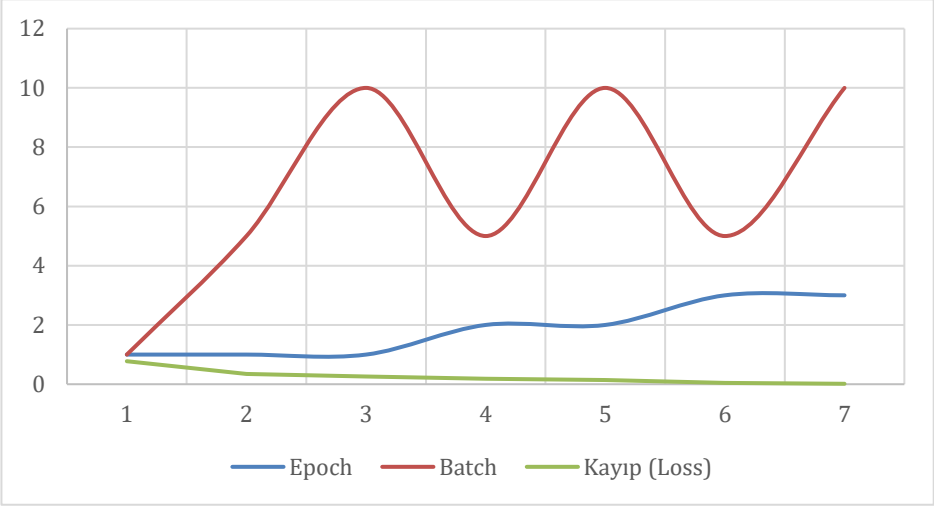
Model tarafından üretilen phishing e-posta örnekleri, insan katılımcılar tarafından da niteliksel olarak değerlendirilmiştir. Katılımcılardan (n=116), sunulan e-postaların gerçek olup olmadığını anlamaya çalışmalarını isteyen bir kör test uygulanmıştır. Katılımcıların yaklaşık %40'ı GAN tarafından üretilmiş e-postaları gerçek e-posta olarak sınıflandırmıştır. Bu sonuç, modelin ürettiği içeriklerin yalnızca otomatik

sistemler değil, aynı zamanda insanlar tarafından da ayırt edilmesinin zor olduğunu göstermektedir. Bu bağlamda GAN mimarisinin sosyal mühendislik saldırılarının simülasyonu açısından etkili ve gerçekçi sonuçlar ürettiği söylenebilir. Tablo 3'te örnek e-postalar verilmektedir.

Tablo 3. Örnek e-postalar.

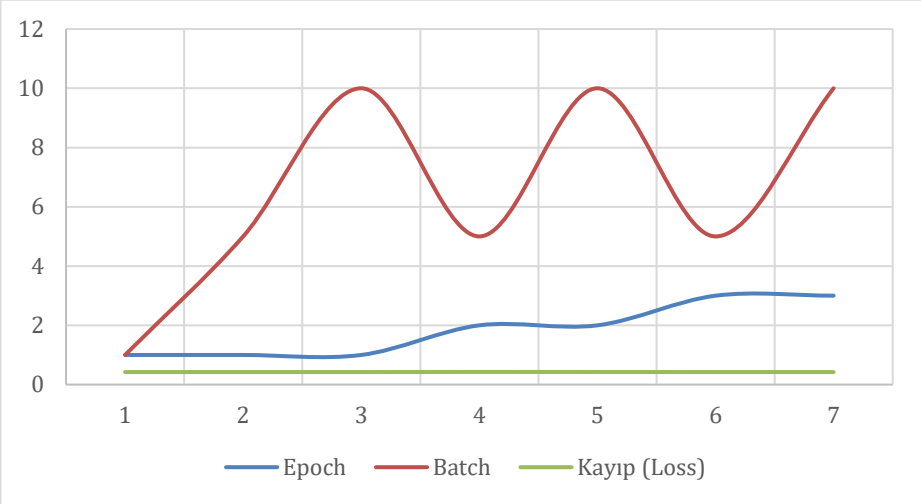
Konu	**İçerik**
Hesap Bildirimi	Değerli kullanıcı, hesabınızda olağan dışı bir etkinlik fark ettik. Lütfen hizmet kesintisi yaşamamak için bilgilerinizi hemen doğrulayın. [Şimdi Doğrulayın]
Acil Güvenlik Uyarısı	Hesabınız geçici olarak kilitlendi. Şimdi oturum açın ve erişimi geri yüklemek için kimliğinizi doğrulayın. [Buraya tıklayın]
Ödeme için Onay Gerekli	Son ödemeniz işlenemedi. Lütfen aşağıdaki bağlantıya tıklayarak fatura bilgilerinizi güncelleyin.

Tablo 3'teki örneklerde kullanılan dil yapısı, tipik phishing e-postalarının ikna edici karakteristiklerini taşımaktadır. Kullanıcıda aciliyet duygusu yaratma, erişim engeli tehdidi, ya da işlem doğrulama gibi psikolojik baskı unsurları, sosyal mühendislik taktiklerinin temelidir. Bu da GAN modelinin bu saldırı tekniklerini başarılı şekilde taklit edebildiğini göstermektedir. Modelin eğitim sürecinde elde edilen log kayıtları, özellikle kayıp fonksiyonlarının (loss) zaman içindeki değişimi açısından incelendiğinde önemli bulgular sunmaktadır. Discriminator başlangıçta yüksek kayıplarla başlamış, ancak her epoch'ta daha düşük değerlere ulaşarak hızlı bir öğrenme eğrisi sergilemiştir. Öte yandan, generator'un kaybı oldukça stabil bir seyir izlemiş ve discriminator'u kandırma başarısında tutarlı bir düzey yakalamıştır. Şekil 1'de, ilk 3 epoch boyunca elde edilen örnek discriminator kayıplarını göstermektedir.



Şekil 1. Dicriminator kayıpları.

Şekil 2'de, ilk 3 epoch boyunca elde edilen örnek generator kayıplarını göstermektedir. Şekil 1 ve 2'ye göre discriminator öğrenmesini hızla tamamlamış, ancak generator sabit kalan kayıp değerleri ile öğrenmenin daha yavaş ve dengeli bir yapıda sürdüğünü göstermiştir. Bu durum GAN mimarisinin doğası gereği beklenen bir eğilimdir. Model, birkaç epoch sonrasında dengeli bir oyun durumuna (Nash Equilibrium) yaklaşmış ve performans stabil hale gelmiştir.



Şekil 2. Generator kayıpları.

Bu çalışmada geliştirilen GAN tabanlı yapı, phishing e-postalarının üretiminde önemli bir başarı göstermiştir. Özellikle LSTM ve BERT mimarilerinin kombinasyonu, modelin hem içerik üretimi hem de içerik sınıflandırması açısından yüksek performans sergilemesini sağlamıştır. LSTM tabanlı generator, dilsel bağlamı dikkate alarak insan benzeri cümleler oluşturabilmiş; bu cümlelerin yapısal tutarlılığı ve semantik gerçekçiliği, hem algoritmik hem de insan temelli değerlendirmelerde tatmin edici bulunmuştur. Öte yandan BERT tabanlı discriminator modeli, çift yönlü bağlam analizi ile çok katmanlı bir ayrıştırma yeteneği sunmuş; bu sayede gerçek ve sahte e-postalar arasında %95'in üzerinde doğrulukla ayrım yapabilmiştir. Bu mimari, sosyal mühendislik saldırılarını anlamada sadece geleneksel sınıflandırıcıların değil, aynı zamanda üreteç-tabanlı yapay zekâ modellerinin de etkili olabileceğini göstermektedir. Ayrıca, GAN'ın ürettiği sahte örnekler, savunma sistemlerinin daha çeşitli senaryolara karşı test edilmesine olanak tanımaktadır. Bu da çalışmayı yalnızca bir sınıflandırma modeli değil, aynı zamanda bir "saldırı simülatörü" olarak değerli kılmaktadır.

Geliştirilen modelin potansiyel uygulama alanları, özellikle siber güvenlik eğitimleri ve tehdit simülasyonu üzerine yoğunlaşmaktadır. Phishing saldırılarının kurum içi farkındalık eğitimlerinde kullanılması oldukça yaygındır; bu bağlamda GAN tarafından üretilen gerçekçi sahte e-postalar, kullanıcıların bu tür saldırılara karşı nasıl tepki vereceklerini değerlendirmek için kullanılabilir. Modelin bu düzeyde gerçekçilik üretme kapasitesi, kurumların güvenlik açıklarını proaktif şekilde değerlendirmesine ve insan hatası kaynaklı zaafların giderilmesine olanak tanır. Ayrıca, anti-phishing yazılımlarının test edilmesi ve performanslarının değerlendirilmesi için yapay olarak üretilmiş çok çeşitli phishing örnekleri kullanılabilir. Böylece, geleneksel güvenlik çözümlerinin yalnızca geçmiş saldırılara değil, gelecekte ortaya çıkabilecek saldırı senaryolarına da karşı ne derece dirençli olduğu ortaya konabilir. Model ayrıca, honeypot sistemlerinde yapay phishing trafiği oluşturmak ve gerçek saldırganların tepkilerini gözlemlemek amacıyla da değerlendirilebilir. Bu yönüyle çalışma, hem ofansif (saldırı) hem de defansif (savunma) siber güvenlik stratejileriyle entegre edilebilir potansiyel taşımaktadır.

Her ne kadar elde edilen sonuçlar umut verici olsa da çalışmanın bazı sınırlılıkları bulunmaktadır. Bunlardan ilki, kullanılan veri setinin boyutu ve çeşitliliğidir. Toplamda 1000 e-posta içeren bir veri kümesi ile çalışılmış olup, bu miktar derin öğrenme mimarileri için göreli olarak küçük bir örneklem olarak değerlendirilebilir. Daha büyük ve farklı kaynaklardan toplanmış çok dilli, sektörel ve görsel içeriğe sahip veri setlerinin kullanılması, modelin genellenebilirliğini önemli ölçüde artıracaktır.

Bir diğer sınırlılık, görsel içeriklerin modelleme sürecine yeterince dahil edilememesidir. Günümüzde birçok phishing e-postası yalnızca metin değil, aynı zamanda logolar, sahte butonlar, HTML temaları veya CAPTCHA görselleri gibi unsurlar da içermektedir. Bu çalışmada görseller metinleştirilmiş betimlemeler (örneğin: "[click here button image]") şeklinde modele entegre edilse de, gerçek görsel veriler ile eğitilmiş multimodal GAN modelleri, saldırı simülasyonlarında çok daha kapsamlı sonuçlar verebilir. Ayrıca, generator'un çıktılarındaki

dilsel tutarlılıklar zaman zaman yapaylık barındırabilir; örneğin bazı cümleler anlamsal olarak tutarsız ya da bağlam dışı olabilir. Bu da GAN mimarisinin daha karmaşık dil modelleriyle (örneğin GPT tabanlı generator'lar) geliştirilmesini gerektirir.

V. SONUÇLAR

Bu çalışma, sosyal mühendislik saldırılarının simülasyonu ve tespiti için GAN (Generative Adversarial Network) tabanlı bir model sunmuştur. LSTM tabanlı bir generator ve BERT tabanlı bir discriminator kullanılarak, gerçekçi phishing e-postaları üretilmiş ve bu e-postaların hem otomatik tespit sistemleri hem de insanlar tarafından ayırt edilme başarısı değerlendirilmiştir. Model, %95'in üzerinde bir doğruluk oranıyla gerçek ve sahte içerikleri sınıflandırabilmiştir. Ayrıca, üretilen e-postaların insan katılımcılar tarafından %40 oranında gerçek olarak algılanması, modelin dilsel ve psikolojik açıdan ikna edici içerikler üretebildiğini göstermiştir.

Çalışmanın en önemli katkılarından biri, siber güvenlik alanında GAN'ların hem saldırı simülasyonu hem de savunma mekanizmalarının test edilmesi için kullanılabileceğini ortaya koymasıdır. Üretilen sahte e-postalar, güvenlik eğitimlerinde kullanılarak kullanıcıların farkındalığını artırabilir veya anti-phishing sistemlerinin performansını değerlendirmek için bir test ortamı sağlayabilir. Ayrıca, açık kaynaklı veri setleri kullanılarak geliştirilen modelin tekrarlanabilir ve geliştirilebilir olması, gelecekteki araştırmalar için bir temel oluşturmaktadır.

Bu çalışmada kullanılan LSTM tabanlı generator, metin üretiminde başarılı sonuçlar vermekle birlikte, daha gelişmiş dil modelleriyle performansın artırılması mümkündür. Özellikle GPT (Generative Pre-trained Transformer) gibi büyük ölçekli dil modelleri, daha tutarlı ve bağlamsal olarak zengin metinler üretebilir. GPT'nin çok katmanlı dikkat mekanizmaları, uzun vadeli bağımlılıkları daha iyi yakalayarak phishing e-postalarının dilbilgisel ve anlamsal tutarlılığını artırabilir.

Diffusion modelleri ise özellikle görsel içeriklerin üretiminde etkili olabilir. Günümüzde phishing e-postaları sıklıkla logo, buton veya sahte CAPTCHA görselleri içermektedir. Diffusion modelleri, bu tür görsel unsurları gerçekçi bir şekilde üreterek multimodal (metin + görsel) phishing simülasyonlarına olanak tanıyabilir. Bu yaklaşım, modelin gerçek dünya saldırılarını daha iyi temsil etmesini sağlayacak ve savunma sistemlerinin görsel tabanlı tehditlere karşı test edilmesine yardımcı olacaktır.

Ayrıca, GPT ve diffusion modellerinin kombinasyonu, hem metin hem de görsel unsurları bir arada üreten hibrit bir sistemin geliştirilmesine öncülük edebilir. Bu tür bir model, phishing saldırılarının çok yönlü doğasını daha iyi yansıtarak siber güvenlik araştırmalarında yeni ufuklar açabilir.

Bu çalışmada geliştirilen modelin bir diğer önemli uygulama alanı, gerçek zamanlı siber savunma sistemlerine entegrasyondur. GAN tabanlı üretilen sahte e-postalar, güvenlik duvarları, e-posta filtreleme sistemleri veya yapay zeka destekli tespit araçları gibi mekanizmaların eğitiminde kullanılabilir. Özellikle dinamik tehdit ortamlarında, saldırganların sürekli evrim geçiren taktiklerine karşı savunma sistemlerinin proaktif olarak güncellenmesi gerekmektedir. Modelin gerçek zamanlı entegrasyonu için şu adımlar önerilebilir:

- ***Sürekli Öğrenme:*** Model, yeni phishing örnekleriyle periyodik olarak eğitilerek güncel kalması sağlanabilir. Bu, saldırganların yeni taktiklerine hızlı adaptasyon için kritik öneme sahiptir.
- ***API Tabanlı Entegrasyon:*** Model, bir API aracılığıyla e-posta filtreleme sistemlerine bağlanarak şüpheli e-postaları analiz edebilir ve gerçek zamanlı uyarılar üretebilir.
- **Honeypot Sistemleriyle İş Birliği:** Üretilen sahte e-postalar, honeypot sistemlerinde yapay phishing trafiği oluşturmak için kullanılabilir. Bu sayede gerçek saldırganların davranışları gözlemlenebilir ve savunma stratejileri buna göre şekillendirilebilir.

Bu bağlamda, bu çalışma GAN tabanlı yaklaşımların siber güvenlik alanındaki potansiyelini ortaya koymuştur. Gelecekte, daha büyük veri setleri, gelişmiş dil modelleri ve multimodal yaklaşımlarla modelin performansının artırılması hedeflenmektedir. Ayrıca, gerçek zamanlı uygulamalara yönelik entegrasyon çalışmaları, modelin pratik değerini daha da artıracaktır. Bu tür çalışmalar, siber güvenlik ekosisteminin saldırganlara karşı daha dirençli hale gelmesine önemli katkılar sağlayacaktır.

BEYANLAR

Teşekkürler: Aileme teşekkür ederim.
Yazarın Katkıları: Kavramsallaştırma, K.A.; metodoloji K.A.; doğrulama, K.A. araştırma, K.A.; kaynaklar, K.A.; veri düzenleme, K.A.; yazma-orijinal taslak hazırlama, K.A; yazma-inceleme ve düzenleme, K.A; denetim, K.A.
Çıkar Çatışması Açıklaması: Yazar herhangi bir çıkar çatışması beyan etmemektedir.

Destekleyen/Destekleyen Kuruluşlar: Bu araştırma herhangi bir dış fon almamıştır.
Etik Onay ve Katılımcı Onayı: Bu çalışmada herhangi bir etik onay gereksinimi söz konusu değildir.
İntihal Beyanı: Bu makale intihal programıyla taranmıştır. İntihal tespit edilmemiştir.
Veri ve Materyallerin Kullanılabilirliği: Veri paylaşımı geçerli değildir.
YZ Araçlarının Kullanımı: Yazar, bu makalenin oluşturulmasında Yapay Zeka (YZ) araçlarını kullanmadığını beyan etmektedir.

KAYNAKLAR

Abdolrazzagh-Nezhad, M., & Langarib, N. (2021). Phishing Detection Techniques: A Review. Data Science: Journal of Computing and Applied Informatics, 9(1). https://doi.org/10.32734/jocai.v9.i1-19904

Aden, I., Child, C. H. T., & Reyes-Aldasoro, C. C. (2024). International Classification of Diseases Prediction from MIMIIC-III Clinical Text Using Pre-Trained ClinicalBERT and NLP Deep Learning Models Achieving State of the Art. Big Data and Cognitive Computing, 8(5), 47. https://doi.org/10.3390/bdcc8050047

Albahadili, A.J.S., Akbas, A. & Rahebi, J. Detection of phishing URLs with deep learning based on GAN-CNN-LSTM network and swarm intelligence algorithms. SIViP 18, 4979–4995 (2024). https://doi.org/10.1007/s11760-024-03204-2

Boukhris, I., Zaâbi, C. A GAN-BERT based decision making approach in peer review. Soc. Netw. Anal. Min. 14, 107 (2024). https://doi.org/10.1007/s13278-024-01269-y

Castaño, F., Fidalgo, E., Alegre, E., Chaves, D., & Sanchez-Paniagua, M. (2021). State of the Art: Content-based and Hybrid Phishing Detection. arXiv preprint arXiv:2101.12723. https://arxiv.org/abs/2101.12723

Elberri, M.A., Tokeşer, Ü., Rahebi, J. et al. A cyber defense system against phishing attacks with deep learning game theory and LSTM-CNN with African vulture optimization algorithm (AVOA). Int. J. Inf. Secur. 23, 2583–2606 (2024). https://doi.org/10.1007/s10207-024-00851-x

Fang, X., Liu, Y., & Zhang, Y. (2019). THEMIS: A Bi-LSTM Based Phishing Email Detection Model. IEEE Access, 7, 122543-122553. https://doi.org/10.1109/ACCESS.2019.2937580

Gan, C.L., Lee, Y.Y. & Liew, T.W. Fishing for phishy messages: predicting phishing susceptibility through the lens of cyber-routine activities theory and heuristic-systematic model. Humanit Soc Sci Commun 11, 1552 (2024). https://doi.org/10.1057/s41599-024-04083-1

Gayathri, M. S., Gokul, S., Mohanshyam, N., & Sudharsan, P. (2023). Detection of Phishing Attack Using GAN with RFC. Rivista Italiana di Filosofia Analitica Junior, 14(2). https://rifanalitica.it/index.php/journal/article/view/252

Kamran, S. A., Sengupta, S., & Tavakkoli, A. (2021). Semi-supervised Conditional GAN for Simultaneous Generation and Detection of Phishing URLs: A Game Theoretic Perspective. arXiv preprint arXiv:2108.01852. https://arxiv.org/abs/2108.01852

Kavya, S., Sumathi, D. Staying ahead of phishers: a review of recent advances and emerging methodologies in phishing detection. Artif Intell Rev 58, 50 (2025). https://doi.org/10.1007/s10462-024-11055-z

Kritika, Er. (2024). A comprehensive literature review on phishing URL detection using deep learning techniques. Journal of Cyber Security Technology, 1–29. https://doi.org/10.1080/23742917.2024.2378552

Nanda, M., Goel, S. URL based phishing attack detection using BiLSTM-gated highway attention block convolutional neural network. Multimed Tools Appl 83, 69345–69375 (2024). https://doi.org/10.1007/s11042-023-17993-0

Pham, T. T. T., Pham, T. D., & Ta, V. C. (2023). Evaluation of GAN-based Models for Phishing URL Classifiers. International Journal of Computer Network and Information Security, 15(2), 1-14. https://doi.org/10.5815/ijcnis.2023.02.01

Thapa, S., Shrestha, A., & Li, J. (2023). An Explainable Transformer-based Model for Phishing Email Detection: A Large Language Model Approach. arXiv preprint arXiv:2402.13871. https://arxiv.org/abs/2402.13871