T.C.
İNÖNÜ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

SİBER TEHDİT İSTİHBARAT VERİLERİNİN MERKEZİ OLARAK İZLENMESİ VE SALDIRILARIN TESPİTİNDE YENİ YAKLAŞIMLARIN GELİŞTİRİLMESİ

DOKTORA TEZİ

Doygun DEMİROL

Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanı: Prof. Dr. Davut HANBAY

MAYIS 2025

T.C
İNÖNÜ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

SİBER TEHDİT İSTİHBARAT VERİLERİNİN MERKEZİ OLARAK İZLENMESİ VE SALDIRILARIN TESPİTİNDE YENİ YAKLAŞIMLARIN GELİŞTİRİLMESİ

DOKTORA TEZİ

Doygun DEMİROL
(23616190004)

Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanı: Prof. Dr. Davut HANBAY
Eş Danışman: Prof. Dr. Resul DAŞ

MAYIS 2025

İNÖNÜ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ MÜDÜRLÜĞÜ

Siber Tehdit İstihbarat Verilerinin Merkezi Olarak İzlenmesi ve Saldırıların Tespitinde Yeni Yaklaşımların Geliştirilmesi

Tezi Hazırlayan: Doygun DEMİROL
Doktora Tezi

TEŞEKKÜR VE ÖNSÖZ

Bu doktora tezinin hazırlanması sürecinde, değerli bilgi birikimi, yol gösterici önerileri ve sabırlı yaklaşımıyla bana rehberlik eden kıymetli danışman hocam Prof. Dr. Davut HANBAY'a en içten teşekkürlerimi sunarım. Ayrıca eş danışmanım Prof. Dr. Resul DAŞ'a, çalışmam boyunca sağladığı bilimsel katkılar, yapıcı eleştiriler ve destekleyici tutumu için gönülden teşekkür ederim. Her iki hocamın da akademik deneyimleri ve anlayışlı yaklaşımları, bu çalışmanın başarıyla tamamlanmasında büyük rol oynamıştır.

Hayatımın her döneminde olduğu gibi, doktora sürecinin zorlu aşamalarında da bana sonsuz sabır, anlayış ve destek gösteren sevgili eşim Nagihan Gülay DEMİROL'a minnettarım. Onun sevgisi, fedakârlığı ve manevi desteği, bu çalışmayı tamamlamamda en büyük güç kaynağım oldu.

Sevgili çocuklarım Kağan ve Ece'ye, varlıklarıyla hayatıma kattıkları mutluluk ve enerji için teşekkür ederim. Sizlerin varlığı, bu uzun ve yorucu süreçte bana ilham veren en değerli motivasyon kaynağı oldu.

Ayrıca, hayatım boyunca bana güvenen, duaları ve sevgileriyle her zaman yanımda olan değerli aileme; yoğun çalışma tempoma rağmen anlayışlarını ve desteklerini esirgemeyen kıymetli çalışma arkadaşlarıma içtenlikle teşekkür ederim. Onların varlığı ve desteği, bu akademik yolculuğu benim için daha anlamlı ve kolay kılmıştır.

Bu çalışmanın ortaya çıkmasında doğrudan ya da dolaylı olarak emeği geçen, ismini burada anamadığım tüm dostlarıma ve meslektaşlarıma da teşekkürlerimi sunarım.

ONUR SÖZÜ

Doktora tezi olarak sunduğum "Siber Tehdit İstihbarat Verilerinin Merkezi Olarak İzlenmesi ve Saldırıların Tespitinde Yeni Yaklaşımların Geliştirilmesi" başlıklı bu çalışmanın bilimsel ahlak ve geleneklere aykırı düşecek bir yardıma başvurmaksızın tarafımdan yazıldığına ve yararlandığım bütün kaynakların hem metin içinde hem de kaynakçada yöntemine uygun biçimde gösterilenlerden oluştuğunu belirtir, bunu onurumla doğrularım.

Doygun DEMİROL

İÇİNDEKİLER

ÇİZELGELER DİZİNİ

ŞEKİLLER DİZİNİ

SEMBOLLER VE KISALTMALAR

BT : Bilişim Teknolojileri

CTI : Cyber Threat Intelligence (Siber Tehdit İstihbaratı)

NLP : Natural Language Processing (Doğal Dil İşleme)

RAG : Retrieval Augmented Generation (Almayla Artırılmış Üretim)

NER : Named Entity Recognition (Adlandırılmış Varlık Tanıma)

APT : Advanced Persistent Threat (Gelişmiş Kalıcı Tehdit)

BERT : Bidirectional Encoder Representations From Transformers (Transformatörlerin Çift Yönlü Kodlayıcı Gösterimleri)

STIX : Structured Threat Information eXpression (Yapılandırılmış Tehdit Bilgisi İfadesi)

OIE : Open Information Extraction (Açık Bilgi Çıkarma)

IOC : Indicator of Compromise (Uzlaşma Göstergesi)

GFS : Google File System (Google Dosya Sistemi)

ETL : Extract, Transform and Load (Çıkartma, Dönüştürme ve Yükleme)

DDoS : Distributed Denial of Service (Dağıtık Hizmet Engelleme)

TAXII : Trusted Automated Exchange of Intelligence Information (Güvenilir Otomatik İstihbarat Bilgi Değişimi)

UCO : Unified Cybersecurity Ontology (Birleşik Siber Güvenlik Ontolojisi)

API : Application Programming Interface (Uygulama Programlama Arayüzü)

IOB : Inside-Outside-Beginning (İç-Dış-Başlangıç)

RDF : Resource Description Framework (Kaynak Tanımlama Çerçevesi)

TF-IDF : Term Frequency-Inverse Document Frequency (Terim Sıklığı-Ters Belge Sıklığı)

MLP : Multi Layer Perceptron (Çok Katmanlı Algılayıcı)

SVM : Support Vector Machine (Destek Vektör Makinesi)

XAI : Explainable AI (Açıklanabilir Yapay Zeka)

AARs : Malware After Action Reports (Kötü Amaçlı Yazılım Eylem Sonrası Raporları)

RE : Relation Extraction (İlişki Çıkarımı)

CVE : Common Vulnerabilities and Exposures (Yaygın Güvenlik Açıkları ve Maruziyetler)

HAG : Hyper Attack Graph (Hiper Saldırı Çizgesi)

GATs : Graph Attention Networks (Çizge Dikkat Ağları)

IDC : International Data Corporation (Uluslararası Veri Şirketi)

IoT : Internet of Things (Nesnelerin İnterneti)

SQL : Structured Query Language (Yapılandırılmış Sorgu Dili)

SOC : Security Operations Center (Güvenlik Operasyon Merkezi)

CPE : Common Platform Enumeration (Yaygın Platform Numaralandırması)

CAPEC : Common Attack Pattern Enumeration and Classification (Yaygın Saldırı Modeli Numaralandırma ve Sınıflandırmaları)

SIEM : Security Information and Event Management (Güvenlik Bilgileri ve Olay Yönetimi)

RNN : Recurrent Neural Networks (Yinelemeli Sinir Ağları)

CRF : Conditional Random Fields (Koşullu Rastgele Alanlar)

GDS : Graph Data Science (Çizge Veri Bilimi)

TTP : Tactics, Techniques and Procedures (Taktikler, Teknikler ve Prosedürler)

GNN : Graph Neural Networks (Çizge Sinir Ağları)

GCN : Graph Convolutional Network (Çizge Evrişimli Ağ)

LSTM : Long Short Term Memory (Uzun Kısa Süreli Bellek)

PAD : Padding (Dolgulama)

MLM : Masked Language Model (Maskeli Dil Modeli)

ÖZET

Doktora Tezi

SİBER TEHDİT İSTİHBARAT VERİLERİNİN MERKEZİ OLARAK İZLENMESİ VE SALDIRILARIN TESPİTİNDE YENİ YAKLAŞIMLARIN GELİŞTİRİLMESİ

DOYGUN DEMİROL

İnönü Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

108 + IX sayfa

2025

Danışman: Prof. Dr. Davut HANBAY

Günümüzün dinamik olarak değişen siber güvenlik ortamı ve artan tehdit ile saldırı karmaşıklığı, ham güvenlik içeriklerinden anlamlı bilgilerin hızlı bir şekilde çıkarılmasını ve eyleme dönüştürülmesini gerekli hale getirmiştir. Siber güvenlik uzmanlarının tehditlere zamanında yanıt verememesi veya tehditleri öngörememesi, kurumların önemli veri kayıpları, itibar zedelenmesi ve operasyonel aksaklıklarla karşı karşıya kalmasını kaçınılmaz hale getirebilmektedir. Bu nedenle, siber tehdit istihbaratının hızlı ve doğru bir şekilde elde edilmesi, savunma stratejilerinin etkinliğini artırmak ve olası zararları en aza indirmek için kuruluşlar için hayati önem taşımaktadır. Güvenlik uzmanlarının bu büyük hacimli içerikleri istihbarat değerlerini kaybetmeden hızlı bir şekilde analiz edebilmesi ve bunlardan içgörü çıkarabilmesi, özellikle artan veri hacmi nedeniyle önemli bir zorluk haline gelmiştir. Siber tehditlerin karmaşık ve dinamik doğası nedeniyle, geleneksel metin analizi yöntemleri siber tehdit verilerinden anlamlı içgörüler çıkarmak için yetersiz kalmaktadır.

Bu çalışmada, ön eğitilmiş bir dil modeli ve siber varlıklar arasındaki ilişkileri tanımlayan bir ontoloji kullanılarak ham tehdit metinlerinden siber varlıkları çıkaran bir dil modeli önerilmiş ve bilgi çizgeleri kullanılarak siber varlıklar ve ilişkileri analiz eden yeni bir yaklaşım sunulmuştur. Bu yaklaşımla, bilgi çizgeleri üzerinde analizler gerçekleştirilerek geleneksel analizlerle öngörülemeyen yeni tehdit ve hedefleri tespit etmek mümkün olmaktadır. Bu tez çalışması, tehdit aktörleri ve kötü amaçlı yazılımlar arasındaki ilişkileri daha anlaşılır hale getirmekte ve farklı tehdit aktörleri ile kötü amaçlı yazılımlar arasında daha önce gizli kalmış ilişkilerin keşfedilmesini kolaylaştırabilmektedir. Sonuçlar, siber güvenlik istihbaratının geliştirilmesinde bilgi çizgelerinin önemini vurgulamakta, siber tehditler hakkında daha fazla içgörü sağlamaktadır. Ayrıca analiz sonuçları etkili savunma stratejilerinin geliştirilmesine de yardımcı olabilmektedir.

Anahtar Kelimeler: Siber tehdit istihbaratı, Büyük dil modeli, Siber güvenlik, Yapay zeka, Büyük veri, Bilgi çizgesi.

ABSTRACT

Phd. Thesis

CENTRALLY MONITORING CYBER THREAT INTELLIGENCE DATA AND DEVELOPMENT OF NEW APPROACHES IN DETECTION OF ATTACKS

Doygun Demirol

İnönü University
Graduate School of Nature and Applied Sciences
Department of Computer Engineering

108 + IX sayfa

2025

Supervisor: Prof. Dr. Davut Hanbay

Today's dynamically changing cyber security environment and increasing threat and attack complexity have made it necessary to quickly extract meaningful information from raw security content and turn it into action. The inability of cyber security experts to respond to or anticipate threats in a timely manner can inevitably expose organisations to significant data loss, reputational damage and operational disruptions. Therefore, obtaining cyber threat intelligence quickly and accurately is vital for organisations to increase the effectiveness of defence strategies and minimise potential damages. The ability of security experts to quickly analyse and extract insights from these large volumes of content without losing their intelligence value has become a significant challenge, especially due to the increasing volume of data. Due to the complex and dynamic nature of cyber threats, traditional text analysis methods are insufficient to extract meaningful insights from cyber threat data.

In this work, we propose a language model that extracts cyber entities from raw threat texts using pre-trained language models and an ontology describing the relationships between these entities, and we present a new approach to analyse cyber entities and relationships using knowledge graphs. With this new approach, it is possible to detect new threats and targets that cannot be predicted by traditional analyses by performing analyses on knowledge graphs. By analysing cyber threat data with knowledge graphs, this thesis provides a better understanding of the relationships between threat actors and malware, and can also help discover previously hidden relationships between different threat actors and malware. The results highlights the importance of knowledge graphs in the development of cyber security intelligence, providing greater insight into cyber threats. In addition, the results of the analysis can also help in the development of effective defence strategies.

Keywords: Cyber threat intelligence, Large language model, Cyber security, Artificial intelligence, Big data, Knowledge graph.

1. GİRİŞ

Bilgi Teknolojileri (BT) sistemlerinin son yıllarda hızlı bir şekilde yaygınlaşmasıyla birlikte, pek çok sektör geleneksel yöntemlerden güncel BT altyapılarına geçiş yapmıştır. Bu dönüşüm, BT sistemlerini saldırganlar için daha çekici ve hedef alınabilir hale getirmiştir. Siber saldırılar da bu gelişmelere paralel olarak hem sayıca artmış hem de yöntem açısından daha karmaşık ve etkili hale gelmiştir. Bu durum, BT altyapıları ve genel toplum açısından kritik bir tehdit olarak görülmeye başlanmıştır. Bu tehditlerle başa çıkabilmek amacıyla siber güvenlik uzmanları, yapay zeka tabanlı çözümlere yönelmiştir. Yapay zeka destekli yöntemler sayesinde, siber saldırılara karşı daha güçlü ve etkili savunma stratejileri geliştirilebilmektedir. Ancak yapay zeka tabanlı güvenlik çözümlerinin başarısı, kullanılan eğitim verilerinin kalitesine ve güncelliğine bağlıdır. Eğitim verilerinin kalitesini artırmak ve bilgi tabanını güncel tutmak amacıyla güncel güvenlik dokümanlarının ve siber varlıkların etkin biçimde analiz edilmesi önem arz etmektedir. Bu dokümanların analiziyle elde edilen sonuçların güvenlik topluluklarıyla eş zamanlı olarak paylaşılması, siber tehdit istihbaratının verimliliğini artırmaktadır. Bu paylaşım sayesinde siber güvenlik sistemlerinin sürekli olarak güncellenen ve doğru bilgilerle eğitilmesi mümkün olmakta, böylece sistemlerin tehditlere karşı dayanıklılığı ve savunma yeteneği önemli ölçüde geliştirilmektedir.

Dijital çağda siber güvenlik, sadece teknik bir zorluk olmanın ötesinde, küresel güvenlik ve mahremiyetin korunması açısından kritik bir eşik olarak karşımıza çıkmaktadır. Dijital verilerin katlanarak büyümesi ve karmaşıklıklarının artması, çok sayıda güvenlik zaafiyetini ortaya çıkarmış ve siber tehditlere karşı korunmayı daha da zorlaşmıştır. Bu durum, tehditlerin tanımlanmasını ve analiz edilmesini daha karmaşık ve zorlu hale getirmektedir. Geleneksel savunma sistemleri çoğunlukla önceden tanımlanmış tehdit imzalarına dayanmakta ve bu durum onları yeni nesil saldırı yöntemlerinde kullanılan gelişmiş kalıcı tehditlere (Advanced Persistent Threat-APT'ler) karşı savunmasız bırakmaktadır. Siber savunmadaki bu boşluk, yalnızca bilinen tehditleri tespit etmekle kalmayıp aynı zamanda ortaya çıkan yeni tehditleri ve potansiyel hedefleri de öngörebilen yenilikçi yaklaşımlara duyulan ihtiyacı vurgulamaktadır.

Siber tehdit istihbaratı (Cyber Threat Intelligence-CTI), kuruluşların altyapılarını siber tehditlerden korumalarına yardımcı olabilecek kritik unsurlardır. Siber güvenlik uzmanlarının yeni saldırılara karşı hazırlıklı olabilmeleri için, güncel CTI raporlarını analiz ederek bilgi tabanlarını yeni kötü amaçlı yazılımlar (Malware) ve saldırı senaryoları konusunda güncel tutmaları hayati önem taşımaktadır. Symantec, McAfee, Trend Micro, FireEye vb. güvenlik kuruluşları tarafından paylaşılan çok sayıda yapılandırılmış CTI verisi bulunmakla birlikte, siber güvenlik bilgi paylaşım platformları, güvenlik raporları gibi halka açık kaynaklarda büyük miktarda, yapılandırılmamış biçimde CTI verileri bulunmaktadır. Yapısal olmayan bu büyük miktardaki siber tehdit istihbarat verilerinin toplanması ve analiz edilmesi oldukça zor ve zaman almaktadır. Bu durum, siber güvenlik tehditlerine karşı savunmayı verimsiz hale getirmektedir. Bununla birlikte, CTI bilgilerini içeren açık kaynak belgelerinin hızla artması, süreçlerin izlemesini ve verilerin işlenmesini zorlaştırmaktadır. Bu durum, bu konuda çalışanlar için önemli sınırlılıklar ve zorluklar içermektedir.

Bu tez çalışmasında, bahsedilen zorlukların üstesinden gelmek için verileri otomatik olarak toplayan, yapılandırılmamış güvenlik metinlerinden siber güvenlik varlıklarını çıkaran ve bu varlıklar arasındaki ilişkileri analiz eden, çizge tabanlı bir yaklaşım önerilmektedir. Önerilen hibrit yaklaşımda, doğal dil işleme, yapay zeka yöntemleri, kural tabanlı örüntü tanıma ve bilgi çizgeleri kullanılmıştır. Bu çok yönlü hibrit yaklaşımda, doğal dil işleme teknikleri ve yapay zeka algoritmaları aracılığıyla siber varlıkları tespit etmeden önce, toplanan verilerin temizlenmesi ve standartlaştırılması amacıyla bir dizi ön işleme adımının uygulanması gerekmektedir. Bu ön işleme adımları kapsamında, veri kaynağının türüne bağlı olarak toplanan metin verilerinin temiz, anlamlı cümle yapılarına ve ortak bir biçime dönüştürülmesi gerekmektedir. Bu süreçte, her bir veri kaynağına özgü yöntemler uygulanarak metinler gürültüden arındırılmış, analiz edilebilir nitelikte temiz cümle yapıları elde edilmiştir. Ancak metnin yapısal uygunluğunu sağlamanın yanı sıra, verinin değerini ve güncelliğini korumak da ayrı bir zorluk alanı olarak öne çıkmaktadır. Büyük veri kavramı içinde sıkça vurgulanan "değer" boyutu, zamana duyarlı bir özelliktir ve zaman ilerledikçe verinin işlevselliği azalmaktadır. Bu nedenle siber istihbarat verilerinin sürekli güncel tutulması gerekmektedir. Bu nedenle, önerilen sistemde siber istihbarat (CTI) verilerinin güncel kalması amacıyla belirli aralıklarla yeni raporlar toplanarak veri havuzu sürekli güncellenmiştir. Ayrıca bu çalışmanın karşılaştığı bir diğer zorluk ise, siber güvenlik alanına özgü etiketlenmiş bir CTI verisetinin bulunmamasıdır. Bu nedenle, adlandırılmış varlık tanıma (Named Entity Recognition–NER) sürecinde kullanılmak üzere özel olarak elle

etiketlenmiş bir veri kümesi oluşturulmuştur. Bu verisetini oluşturmak amacıyla, PDF formatındaki siber güvenlik raporları, güvenlik blog sayfalarından elde edilen içerikler ve MITRE ATT&CK [1] verileri kullanılmıştır. Elde edilen bu veriler, ön işleme adımlarından geçirilerek temizlenmiş ve ardından veri etiketleme sürecinde kullanıma hazır hale getirilmiştir.

1.1 Tezin Amacı

Bu tez çalışmasının temel amaçları aşağıda listelenmiştir.

- Farklı kaynaklardan elde edilen ve biçimsel olarak çeşitlilik gösteren metin tabanlı CTI (Cyber Threat Intelligence) verilerinin, merkezi bir veri tabanında toplanması ve işlenebilmesi süreçlerini otomatikleştirmek.
- Otomatik olarak toplanan metin tabanlı CTI verilerinin, bilgi çıkarımı süreçlerinde daha etkili kullanılabilmesi için dilsel ve yapısal açıdan ön işleme adımlarıyla dönüştürülerek anlamlı hale getirmek.
- Yapay zeka tabanlı yöntemler kullanarak, siber tehdit istihbaratına yönelik metinlerden varlık ve ilişki çıkarımı yapmak ve tehdit unsurları arasındaki etkileşimlerin analiz edilebilmesini sağlayacak semantik bir yapı inşa etmek.
- Çıkarılan siber varlıklar ve bunlar arasındaki ilişkiler üzerinden çizge tabanlı analizler gerçekleştirerek, tehdit aktörlerinin davranışsal örüntülerine ve siber saldırıların yapısal özelliklerine dair anlamlı içgörüler elde etmek.
- Farklı platformlarda yer alan güvenlik bültenleri, blog yazıları ve haber içerikleri gibi metin tabanlı siber tehdit dokümanlarında bulunan ancak çoğu zaman atıl durumda kalan güvenlik verilerini analiz ederek anlamlı ve kullanılabilir bilgiye dönüştürmek.

Bu doğrultuda, önerilen yaklaşımlarla siber tehdit istihbarat verilerinin anlamlandırılması, görselleştirilmesi ve kullanılabilir bilgiye dönüştürülmesi hedeflenmekte; böylece çalışmanın siber güvenlik alanında literatüre özgün ve uygulanabilir katkılar sağlaması öngörülmektedir.

1.2 Tezin Motivasyonu

Her gün tehdit raporları, blog yazıları, günlükler ve yapılandırılmış siber tehdit verileri gibi birçok kaynaktan büyük miktarda güvenlik verisi üretilmektedir. Bu veriler;

saldırı teknikleri, istismar edilen zafiyetler, hedef alınan sistemler ve tehdit aktörlerinin davranışlarına ilişkin önemli bilgiler içermektedir. Ancak bu verilerin yüksek hacmi, biçimsel çeşitliliği ve düzensizliği, analiz süreçlerinde ciddi zorluklar yaratmaktadır. Geleneksel metin işleme yaklaşımları, bu ölçekteki verilerde gizli kalan kalıpları ve karmaşık ilişkileri ortaya çıkarmada yetersiz kalmakta ve dolayısıyla potansiyel tehdit bilgileri çoğu zaman işlenmeden atıl durumda kalmaktadır. İşte tüm bu zorluklar, mevcut analiz yaklaşımlarının sınırları ve kullanılmadan kalan veri potansiyeli, bu çalışmanın temel motivasyonunu oluşturmaktadır. Siber güvenlik uzmanlarının tehdit ortamını daha derinlemesine anlayabilmelerine ve veri değeri zamanında ortaya çıkarabilmelerine katkı sağlamak amacıyla, bu tezde doğal dil işleme teknikleri ile bilgi çizgelerinin olanaklarını bir araya getiren bütüncül bir yaklaşım önerilmektedir. Bilgi çizgeleri, siber varlıklar ile bunlar arasındaki ilişkileri çizge tabanlı yapılarla temsil ederek; gizli örüntülerin keşfi, saldırı davranışlarının modellenmesi ve çok boyutlu ilişkilerin analizini kolaylaştırmaktadır.

1.3 Tez Organizasyonu ve Katkılar

Bu tez çalışması, beş bölümden oluşmaktadır:

1. Bölümde, tez konusu hakkında bilgiler, tezin amacı, motivasyonu ve katkılar ile tezin organizasyonu özetlenmiştir.

2. bölümde, literatür taraması ve değerlendirilmesi detaylı bir şekilde sunulmuştur.

3. bölümde, çalışmanın dayandığı temel kavramlar olan büyük veri, siber tehdit istihbaratı, doğal dil işleme ve adlandırılmış varlık tanıma, çizge veritabanları ve çizge tabanlı veri analizi hakkında temel bilgiler verilmiştir.

4. bölümde, farklı kaynaklardan elde edilen CTI verilerinin otomatik olarak toplanması, ön işleme adımlarıyla anlamlı hale getirilmesi, yapay zekâ tabanlı varlık ve ilişki çıkarımı yapılması ve çizge tabanlı analizlerle tehdit örüntülerinin ortaya konulması ele alınmıştır. Bu bağlamda, tez kapsamında geliştirilen ve önerilen yaklaşımın uygulaması tüm detaylarıyla sunulmuştur.

5. bölümde ise, gerçekleştirilen uygulamaya ait bulgular sunulmuş ve elde edilen sonuçlar kapsamlı şekilde tartışılmıştır.

Tez çalışmasından üretilen akdemik çıktılar aşağıda listelenmiştir:

- Demirol, D., Das, R., & Hanbay, D. (2019, Eylül). Büyük veri üzerine perspektif bir bakış. International Conference on Artificial Intelligence and Data Processing (IDAP) (pp. 1-9). IEEE. https://doi.org/10.1109/IDAP.2019.8875902 (Uluslararası konferans bildirisi)
- Demirol, D., Das, R., & Hanbay, D. (2023). A key review on security and privacy of big data: issues, challenges, and future research directions. Signal, Image and Video Processing, 17 (4), 1335-1343. https://doi.org/10.1007/s11760-022-02341-w (SCI uluslararası dergi makalesi)
- Demirol, D., Das, R., & Hanbay, D. (2025). A Novel Approach for Cyber Threat Analysis Systems Using BERT Model from Cyber Threat Intelligence Data. Symmetry, 17(4), 587. https://doi.org/10.3390/sym17040587 (SCI uluslararası dergi makalesi)

2. LİTERATÜR TARAMASI VE DEĞERLENDİRİLMESİ

Bu bölümde, siber tehdit istihbaratının önemi, mevcut tehdit istihbaratı kaynakları ve bu kaynaklardan tehdit bilgilerinin çıkarılması ile ilgili literatür çalışmaları incelenmiştir.

2.1 Mevcut Literatürün Değerlendirilmesi

Siber tehdit istihbaratı, kuruluşların ve güvenlik topluluklarının kendi sistemlerini, hızla gelişen siber tehditlere karşı etkin bir şekilde korumaları için hayati önem taşımaktadır. Özellikle APT saldırılarına yönelik siber tehdit istihbaratı, saldırganlar, hedefler ve saldırı teknikleri-taktikleri hakkında ayrıntılı teknik bilgiler içermektedir. Ayrıca, geleneksel yöntemler kullanılarak tehditlerle ilgili bilgilerin çıkarılması ve analiz edilmesi genellikle çok zaman alıcı olabilir ve güvenlik analistleri için önemli ölçüde elle çalışma gerektirebilmektedir. Bu nedenle, yapısal olmayan verilerden, siber tehdit istihbaratının çıkarılması ve kullanılması önemli ölçüde zorluk içeren karmaşık bir süreçtir. Bu zorluklar nedeniyle, çoğu güvenlik uzmanı, kamuya açık veri kaynaklarından tehdit istihbaratı elde etme süreçlerini otomatikleştirmeye yönelik çözümlere odaklanmıştır. ThreatExchange [2], Symantec [3], Kaspersky, bilgisayar korsan forumları ve sosyal medya gibi platformlar, tehdit istihbaratı elde edilmesi için yararlı kaynaklar sunmaktadır. Bununla birlikte, belirtilen tehdit kaynakları çoğunlukla yapılandırılmamış metin biçimindedir. Yapılandırılmamış siber tehdit metinleri, tehdit istihbarat bilgilerinin paylaşılmasını kolaylaştırmak ve saldırıların hızlı ve etkili bir şekilde önlenmesini ve tespit edilmesini sağlamak için STIX [4] ve MAEC [5] gibi formatlarda makine tarafından okunabilir şekilde standartlaştırılmıştır [6].

Bu bölümde, siber güvenlik alanında açık veri kaynaklarından CTI verilerinin çıkarılmasına yönelik mevcut çalışmalar detaylı olarak incelenmektedir. Yinhai ve diğerleri [7] yapılandırılmamış APT metinlerinden CTI verilerini çıkarmak ve bunları otomatik olarak analiz etmek amacıyla CTI View adlı bir analiz çerçevesi önermiştir. Araştırmacılar, tehdit varlıklarını çıkarmak için transformatörlerden çift yönlü kodlayıcı temsillerine (Bidirectional Encoder Representations from Transformers - BERT) dayanan hibrit bir model kullanmışlardır. BERT-GRU-BiLSTM-CRF mimarisiyle eğitilen bu model, %72'nin üzerinde bir doğruluk oranıyla varlık çıkarımı yapmıştır. Jones ve arkadaşları [8], ağ güvenliği ile ilgili siber varlıkları ve bunlar arasındaki ilişkileri çıkarmak için bir önyükleme (bootstrapping) tabanlı bir algoritma, yarı denetimli makine öğrenimi teknikleri ve aktif öğrenme yaklaşımlarının bir kombinasyonunu kullanmıştır. Sekiz farklı ilişki türü içeren bir

veri kümesi üzerinde ortalama %82'lik bir F1 puanı elde etmişlerdir. Husari ve diğerleri [6], yapılandırılmamış CTI metinlerini analiz eden ontoloji tabanlı TTPDrill adlı bir araç geliştirmiştir. Sistem, destek vektör makinesi tabanlı bir belge sınıflandırıcı kullanarak ilgili içerikleri filtrelemekte ve güvenlik terimlerinin yoğunluğu gibi özelliklerden yararlanmaktadır. Araç, raporlardaki özne-fiil-nesne yapılarını tanıyarak bunları ontolojideki tehdit eylemleriyle eşleştirmektedir. CTI metinleri üzerinde yapılan testler ile TTPDrill'in tehdit eylemlerini yüksek doğrulukla tespit edebildiğini göstermiştir. Alves ve arkadaşları [9] ise, hedeflerle ilgili tehditlerin bir özetini düzenli olarak güncelleyen bir Twitter akış tehdit sistemi sunmaktadır. Araştırmacılar, siber güvenlik olaylarıyla ilgili tweetleri topladıktan sonra bu tweetlerden ve terim sıklığı-ters belge sıklığı (Term Frequency-Inverse Document Frequency - TF-IDF) yöntemlerini kullanarak özellikler çıkarmışlardır. Ardından toplanan tweetler için sınıflandırıcı olarak hem çok katmanlı bir algılayıcı (Multi-Layer Perceptron - MLP) hem de destek vektör makinesi (Support Vector Machine - SVM) kullanmışlardır. Bir başka çalışmada, Kim ve arkadaşları [10], veri havuzlarından yapılandırılmış CTI verilerini kullanan CyTIME adlı bir çerçeve önermektedir. Bu çerçeve, istihbarat verilerini sürekli olarak toplayıp, insan müdahalesi olmadan otomatik olarak kurallar oluşturmakta ve ardından bunları gerçek zamanlı ağ siber tehditlerini azaltmak için yapılandırılmış tehdit bilgi ifadesi (Structured Threat Information eXpression - STIX) olarak bilinen bir JSON biçimine dönüştürmektedir. Zhang ve arkadaşları [11], doğal dil işleme yöntemlerini kullanarak CTI raporlarından siber tehdit eylemlerini çıkarmak için bir çerçeve önermiştir. Önerdikleri sistem siber eylemlerin çıkarılmasına ek olarak, varlıklar arasındaki ilişkileri de otomatik olarak belirlemektedir. Piplai ve arkadaşları [12] ise, olay sonrası değerlendirme raporlarından siber bilgileri çıkarmak, siber güvenlik analistlerine kapsamlı ve anlayışlı analizler sunmak ve bunları bir bilgi çizgesinde temsil etmek için bir çerçeve geliştirmiştir. Sistem, siber güvenlik ile ilgili varlıkları tanımlamak için adlandırılmış varlık tanıma ve düzenli ifadeleri kullanmaktadır. Sarhan ve arkadaşları [13], yapılandırılmamış metinlerden siber istihbarat elde etmek için sinir ağı tabanlı açık bilgi çıkarma (Open Information Extraction - OIE) yöntemine dayalı bir sistem sunmaktadır. Önerilen yaklaşım, tehdit raporlarından bilgi çizgesi temsili oluşturmakta ve OIE kullanarak adlandırılmış varlık tanıma işlemini gerçekleştirmektedir. Alam ve arkadaşları [14] ise, siber varlıkları çıkarmak amacıyla adlandırılmış varlık tanıma sistemini gerçekleştirmek için dönüştürücü (transformer) tabanlı bir çerçeve tasarlamıştır. Bu çerçeve, tehdit raporları üzerinde ön eğitimli XLM RoBERTa olarak bilinen bir sinirsel dil modeli kullanmaktadır. Zhu ve arkadaşları [15], siber güvenlik makalelerinden saldırı

göstergelerini (Indicator of Compromise – IOC) çıkarmak ve makaleleri kampanya aşamalarına göre sınıflandırmak için doğal dil işleme yöntemlerini kullanan bir sistem önermiştir. Önerilen sistemde, saldırı göstergelerini çıkarma aşamasını geliştirmek için çeşitli kural tabanlı yöntemler kullanılmıştır. Ayrıca, bu yöntemlerin etkinliğini değerlendirmek amacıyla kapsamlı deneysel analizler gerçekleştirilmiştir. Bu çalışmalar siber tehdit istihbaratının çıkarılmasına yönelik çeşitli yaklaşımlar ortaya koymaktadır. Bununla birlikte, son gelişmeler, transformer tabanlı modelleri özel bilgi temsilleriyle birleştiren daha sofistike yöntemlere doğru açık bir eğilim göstermektedir. Mevcut araştırmalar ağırlıklı olarak, dil modellerini çizge tabanlı yaklaşımlarla birleştiren ve geleneksel yöntemlere kıyasla daha yüksek performans ölçütleri elde eden hibrit mimarilere odaklanmaktadır. Bu hibrit yaklaşımlar, tehdit istihbaratı çıkarma süreçlerinde otomasyonu artırmakta ve güvenlik analistlerinin elle yapılan iş yükünü azaltmaktadır. Ayrıca bu yöntemlerin, tehditlerin erken aşamada tespit edilmesi ve önlenmesi konusunda önemli avantajlar sağladığı görülmektedir. Yukarıda incelenen çalışmaların yanı sıra, literatürde siber tehdit istihbaratının çıkarılması ve analiz edilmesi üzerine çok sayıda farklı yaklaşım bulunmaktadır. Bu yaklaşımlar, kullanılan yöntemler, veri kaynakları, uygulama alanları ve elde edilen sonuçlar açısından farklılık göstermektedir. Bu bağlamda, mevcut literatürdeki önemli çalışmaların kapsamlı bir değerlendirmesi Çizelge 2.1'de sunulmuştur.

Çizelge 2.1 : Siber tehdit istihbaratı ile ilgili literatür çalışmalarının incelenmesi.

Kaynak	Veri	Metot	Çıkarım	Sonuç
[16]	Açık kaynaklardan toplanan CTI raporları	Doğal Dil İşleme (NLP), Gizli Anlamsal İndeksleme (Latent Semantic Analysis - LSA), Tekil Değer Ayrıştırması (Singular Value Decomposition - SVD), Naive Bayes, K-En Yakın Komşu (KNN), Karar Ağaçları, Rastgele Orman (Random Forest), Derin Sinir Ağları (Deep Neural Network - DNN)	Metinlerden tehdit aktörleri ve üst düzey bağlamsal tehdit göstergeleri (high-level IOCs) çıkarılarak, sınıflandırma ve analiz doğruluğu artırılmıştır.	Derin sinir ağı (DNN), %94 doğruluk oranı ile en başarılı model olarak öne çıkmış; diğer makine öğrenimi algoritmalarına kıyasla daha yüksek tutarlılık ve performans göstermiştir.
[17]	Freebuf web sitesi ve WooYun güvenlik açığı veri tabanından elde edilen metin tabanlı güvenlik açığı bildirileri.	CRF, LSTM tabanlı modeller (LSTM-CRF, BiLSTM-CRF), konvolüsyonel sinir ağlarıyla desteklenmiş birleşik yapılar (CNN-BiLSTM-CRF, FT-LSTM-CRF, FT-BiLSTM-CRF, FT-CNN-BiLSTM-CRF)	Yerel (bağlama özgü) ve küresel (genel) bağlam özelliklerinin bir araya getirilmesiyle, adlandırılmış güvenlik varlıklarının daha doğru ve tutarlı biçimde tespit edildiği gözlemlenmiştir.	Önerilen FT-CNN-BiLSTM-CRF modeli, %93.31 doğruluk oranı ve %86 F1 skoru ile en yüksek performansı göstermiş; diğer modellere kıyasla üstün başarı sergilemiştir.
[18]	Twitter (X) platformundan elde edilen siber güvenlik odaklı sosyal medya paylaşımları.	Terim Frekansı–Ters Doküman Frekansı (Term Frequency–Inverse Document Frequency - TF-IDF), Yoğunluk Tabanlı Kümeleme Algoritması (Density-Based Spatial Clustering of Applications with Noise - DBSCAN), Anahtar Kelime Sıralama Algoritması (TextRank), Doğal Dil Anlamlandırma Servisi (TextRazor)	Önerilen yöntem, sosyal medya akışlarından yeni ve gelişmekte olan siber tehdit olaylarının tespit edilmesini sağlamaktadır.	Geliştirilen sistem, %93.75 doğruluk oranı ile yüksek başarı göstermiş; bu da X verilerinin tehdit algılama süreçlerinde etkin bir biçimde kullanılabileceğini ortaya koymuştur.
[19]	CVE (Common Vulnerabilities and Exposures) listesi, Adobe ve Microsoft güvenlik bültenleri ile çeşitli güvenlik bloglarından elde edilen metinler.	Stanford Adlandırılmış Varlık Tanıma Aracı (Stanford Named Entity Recognizer - NER), Uzun Kısa Süreli Bellek Ağı + Yoğun Katman (Long Short-Term Memory + Dense), LSTM + Koşullu Rastgele Alanlar (Conditional Random Fields - CRF), Genişletilmiş İki Yönlü LSTM + CRF (Extended Bidirectional LSTM - CRF)	Önerilen XBiLSTM-CRF modeli, farklı varlık kategorileri arasındaki ayrımları daha iyi öğrenerek adlandırılmış varlık tanıma (NER) görevinde üstün başarı sağlamıştır.	Model, %90.54 doğruluk oranı ve %89.38 F1 skoru ile diğer yöntemlere kıyasla en yüksek performansı göstermiştir.
[20]	APT (Advanced Persistent Threat) güvenlik raporları ve kötü amaçlı yazılım (malware) depolarından elde edilen analiz verileri.	Düzenli İfade (Regular Expression - Regex) tabanlı Göstergelerin (Indicator of Compromise - IoC) ayrıştırılması, kötü amaçlı yazılım analiz hizmetlerinden elde edilen çıktıların bütünleştirilmesi .	CTIMiner sistemi, açık kaynaklı tehdit raporları ve zararlı yazılım analiz verilerini bir araya getirerek yüksek kaliteli, yapılandırılmış bir CTI veri seti üretmiştir.	Sistem, IoC'lerin otomatik çıkarımı ve malware analizlerinin entegrasyonu yoluyla CTI verilerinin hızla elde edilmesini sağlamış; tehdit analizi süreçlerinde zaman ve kaynak tasarrufu sağlamıştır.

Çizelge 2.1 (devam) : Siber tehdit istihbaratı ile ilgili literatür çalışmalarının incelenmesi.

Kaynak	Veri	Metot	Çıkarım	Sonuç
[21]	CVE (Common Vulnerabilities and Exposures) kayıtları ve çeşitli güvenlik bloglarından elde edilen metinler.	İki Yönlü Uzun Kısa Süreli Bellek Ağı + Dikkat Mekanizması + Koşullu Rastgele Alanlar (Bidirectional Long Short-Term Memory + Attention + Conditional Random Fields - BiLSTM-Attention-CRF), Kalabalık Kaynaklı CRF modeli (CRF-crowd), Çoklu Anotasyon CRF modeli (CRF-MA), Dawid ve Skene temelli LSTM modeli, BiLSTM-Attention-CRF-VT.	Önerilen yöntemler, düşük kaliteli ve gürültülü metinlerden doğru bilgi çıkarımını ve güvenlik varlıklarının güvenilir şekilde sınıflandırılmasını mümkün kılmıştır.	BiLSTM-Attention-CRF-VT modeli, %89.1 doğruluk ve %89 F1 skoru ile en yüksek performansı göstermiş; diğer yaklaşımlara kıyasla daha başarılı sonuçlar üretmiştir.
[22]	CVE (Common Vulnerabilities and Exposures) kayıtlarından elde edilen güvenlik açığı metinleri.	Adlandırılmış Varlık Tanıma (Named Entity Recognition - NER), Ontoloji Tabanlı Veri Modellemesi (Ontology-Based Data Modeling)	IoT (Nesnelerin İnterneti) güvenlik durumlarının belirlenmesi ve değerlendirilmesi amacıyla NER tabanlı bir analiz sistemi geliştirilmiştir.	Önerilen model, IoT ağlarındaki güvenlik açıklarını tespit etmede başarılı olmuş; semantik analiz ve ontolojik yapıların katkısıyla daha anlamlı ve bütüncül sonuçlar üretmiştir.
[23]	Açık kaynak tehdit istihbarat platformlarından elde edilen makale ve içerikler.	Çok Katmanlı Algılayıcı (Multi-Layer Perceptron - MLP), Terim Frekansı–Ters Doküman Frekansı (Term Frequency–Inverse Document Frequency - TF-IDF) tabanlı özellik çıkarımı.	Önerilen yöntem, tehdit istihbaratına yönelik metinlerden doğru ve anlamlı bilgi çıkarmada etkili bir yapı sunmuştur.	Geliştirilen sistem, %94.85 doğruluk oranı ile makaleleri başarılı şekilde sınıflandırmış; çok kaynaklı veri ile zenginleştirilmiş bilgi çizgeleri üretmiştir.
[24]	Siber tehdit raporları, Çin kaynaklı CTI (Cyber Threat Intelligence) raporları ve zafiyet bilgileri.	Koşullu Rastgele Alanlar (Conditional Random Fields - CRF), İki Yönlü Uzun Kısa Süreli Bellek Ağı (Bidirectional Long Short-Term Memory - BiLSTM), BiLSTM-CRF, BiLSTM-CRF + Ontoloji Tabanlı Düzeltme Modülü	BiLSTM-CRF modelinin ontoloji tabanlı düzeltme modülüyle entegrasyonu, bağlamsal doğruluğu artırmış ve özellikle karmaşık güvenlik varlıklarının doğru şekilde çıkarılmasında etkili olmuştur.	BiLSTM-CRF + Ontoloji tabanlı düzeltme modeli, tüm varlık türlerinde en yüksek F1 skorlarına ulaşmış; siber tehdit istihbaratı için güvenilir ve bütüncül bilgi çıkarımı sağlamıştır.
[25]	Siber tehdit raporları ve güvenlik açıklarına ilişkin metin tabanlı bilgiler.	Koşullu Rastgele Alanlar (Conditional Random Fields - CRF), Uzun Kısa Süreli Bellek Ağı + CRF (LSTM-CRF), Konvolüsyonel Sinir Ağı + CRF (CNN-CRF), Tekrarlayan Sinir Ağı + CRF (RNN-CRF), Kapı Kontrollü Tekrarlayıcı Ünite + CRF (Gated Recurrent Unit - GRU-CRF), Çift Yönlü GRU + CRF (BiGRU-CRF), BiGRU + CNN + CRF birleşik modeli.	Derin öğrenme tabanlı BiGRU + CNN + CRF modeli, bağlamsal bilgi modelleme ve varlıklar arası ilişkilerin anlamlandırılmasında üstün başarı sergilemiştir.	BiGRU + CNN + CRF modeli, %93.4 F1 skoru ile en yüksek performansı göstermiş; varlık tanımlamada diğer yaklaşımlara kıyasla daha doğru ve tutarlı sonuçlar üretmiştir.

Çizelge 2.1 (devam) : Siber tehdit istihbaratı ile ilgili literatür çalışmalarının incelenmesi.

Kaynak	Veri	Metot	Çıkarım	Sonuç
[26]	Siber tehdit raporları ve CVE (Common Vulnerabilities and Exposures) kayıtlarından elde edilen metinler.	Karakter Düzeyinde RNN + Çift Yönlü LSTM (Char-RNN-BiLSTM), Char-RNN-BiLSTM + Koşullu Rastgele Alanlar (CRF), Karakter Düzeyinde CNN + BiLSTM (Char-CNN-BiLSTM), Char-CNN-BiLSTM + CRF, Çanta-İçerik Temsili + BiLSTM (Bag of Characters - BOC-BiLSTM), BOC-BiLSTM + CRF.	BOC-BiLSTM-CRF modeli, karakter düzeyindeki temsilleri bağlamsal bilgilerle birleştirerek doğruluk ve verimlilik açısından en yüksek başarıyı göstermiştir. CRF katmanı genel olarak doğruluğu artırmış; özellikle Char-CNN yapısı yerel özellikleri etkili biçimde modellemiştir.	%75.05 F1 skoru ile önerilen BOC-BiLSTM-CRF modeli en iyi performansı sergilemiş; CRF katmanının tüm modellerde önemli bir katkı sağladığı gözlemlenmiştir.
[27]	Siber tehdit raporlarından elde edilen metin tabanlı veriler.	CTIBERT modeli (Bidirectional Encoder Representations from Transformers - BERT, Çift Yönlü Gated Recurrent Unit - BiGRU, Koşullu Rastgele Alanlar - CRF, Çok Başlı Dikkat Mekanizması - Multi-head Attention)	Hyper Attack Graph (HAG) çerçevesi, siber tehdit istihbaratını hiper-çizge (hypergraph) yapısı üzerinden modelleyen ilk yaklaşımlardan biridir. CTIBERT modeli ise karmaşık bağlamsal bilgileri etkili biçimde çıkarmada yüksek başarı göstermiştir.	HAG modeli, taktiksel ve teknik tehdit bilgilerini sezgisel biçimde yapılandırmış; CTIBERT modeli ise çok başlı dikkat mekanizması sayesinde doğruluk oranını artırarak bağlamsal analizlerde üstün performans sergilemiştir.
[28]	Otomatik olarak etiketlenmiş siber güvenlik metinlerinden oluşan veri kümesi.	GloVe gömme vektörleri + Çift Yönlü Uzun Kısa Süreli Bellek Ağı + Koşullu Rastgele Alanlar (GloVe + BiLSTM + CRF), FastText + BiLSTM + CRF, BERT-base-cased + BiLSTM + CRF, BERT-large-cased + BiLSTM + CRF, Whole Word Masking ile eğitilmiş BERT-large-cased + BiLSTM + CRF, BERT-base-cased + Tam Bağlantılı Katman (Feedforward Network - FFN), BERT-large-cased + FFN, BERT-large-cased-wwm + FFN	BERT-large-cased-wwm + FFN modeli, %97.4 F1 skoru ile en yüksek performansı göstermiş; bağlamsal bilgi çıkarımında diğer tüm modelleri geride bırakmıştır.	BERT tabanlı yaklaşımlar, özellikle Whole Word Masking (WWM) ile doğruluk oranını artırmış; CRF katmanı ise etiket bağımlılıklarını modellemede önemli katkı sağlamıştır.
[29]	Microsoft, Cisco, McAfee, Kaspersky, Fortinet, CrowdStrike ve diğer kaynaklardan elde edilen siber tehdit istihbaratı (CTI) raporları .	Konvolüsyonel Sinir Ağı + Çift Yönlü Uzun Kısa Süreli Bellek Ağı + Koşullu Rastgele Alanlar (CNN-BiLSTM-CRF), CNN-BiGRU-CRF, BERT tabanlı BiLSTM-CRF ve BiGRU-CRF, RoBERTa tabanlı BiLSTM-CRF ve BiGRU-CRF	RoBERTa-BiGRU-CRF modeli, bağlamsal bilgi çıkarımında %83.2 F1 skoru ile en yüksek başarıyı sağlamış; CRF katmanı ise etiketler arası ilişkisel bağımlılıkları etkili biçimde modelleyerek doğruluk oranını artırmıştır.	RoBERTa tabanlı modeller, örtüşmeli veri testlerinde karmaşık ilişkileri daha iyi çözümlemiş; tam veri kümesi üzerinde yapılan testlerde ise diğer modellere kıyasla daha yüksek doğruluk oranlarına ulaşarak öne çıkmıştır.

Çizelge 2.1 (devam) : Siber tehdit istihbaratı ile ilgili literatür çalışmalarının incelenmesi.

Kaynak	Veri	Metot	Çıkarım	Sonuç
[30]	Microsoft, Cisco, McAfee, Kaspersky ve diğer kaynaklardan toplanan siber tehdit istihbaratı (CTI) raporları	Çift Yönlü Uzun Kısa Süreli Bellek Ağı (Bidirectional Long Short-Term Memory - BiLSTM), Bidirectional Encoder Representations from Transformers (BERT)	Geliştirilen yöntem, CTI raporlarından saldırı davranışlarını otomatik olarak çıkarmakta etkili olmuştur.	Yöntem, tehdit tespitinde yüksek doğruluk oranları elde etmiş; özellikle karmaşık yapıya sahip CTI raporlarından anlamlı ve kullanılabilir bilgiler çıkarılmasını sağlamıştır.
[13]	Microsoft, Cisco, McAfee, Kaspersky gibi siber güvenlik firmalarına ait raporlar ile açık kaynaklı veri setlerinden elde edilen metinler.	Çift Yönlü Kapı Kontrollü Tekrarlayıcı Ünite (Bidirectional Gated Recurrent Unit - BiGRU), BiGRU + Dikkat Mekanizması (Attention), BiLSTM + Koşullu Rastgele Alanlar (BiLSTM + CRF), BiGRU + CRF	Geliştirilen yöntem, yapılandırılmamış metinlerden bağlamsal olarak doğru bilgi çıkarımı yaparak siber tehdit analizi için yenilikçi bir yaklaşım sunmuştur. Ayrıca, canonicalization süreci bilgi çizgesi oluşturmayı standartlaştırarak bilgi birliğini artırmıştır.	Open-CyKG-BiGRU+CRF modeli, ilişkilerin doğru çıkarımında %98.9 F1 skoru ve %80.8 duyarlılık (recall) oranı ile üstün başarı göstermiştir. Canonicalization süreçleri ise %82.6 F1 skoru ile ilişki eşleştirmede yüksek doğruluk sağlamıştır.
[31]	CVE açıklamaları, APT (Advanced Persistent Threat) raporları, güvenlik bültenleri, MITRE ATT&CK çerçevesi, MISP, Unit 42 ve WatcherLab kaynaklarından elde edilen tehdit verileri.	Bidirectional Encoder Representations from Transformers (BERT), Çift Yönlü Gated Recurrent Unit (BiGRU), Dikkat Mekanizması (Attention Mechanism)	Varlıklar ve bunlar arasındaki ilişkilerin eşzamanlı çıkarımı amacıyla geliştirilen yöntem, tehdit istihbaratının doğruluğunu ve bağlamsal bütünlüğünü artırmıştır. Ontoloji tabanlı eşleştirme işlemi, semantik uyumu güçlendirmiştir.	Oluşturulan bilgi çizgesi, %81.37 F1 skoru ile güvenilir bilgi ve ilişkiler sunmuş; BERT tabanlı model bağlamsal analiz görevlerinde üstün performans göstermiştir.
[32]	CTI (Cyber Threat Intelligence) raporları, MITRE ATT&CK verileri, NVD (National Vulnerability Database) ve açık kaynaklı veri kaynakları.	BERT + Çift Yönlü Uzun Kısa Süreli Bellek Ağı + Koşullu Rastgele Alanlar (BERT + BiLSTM + CRF), BERT tabanlı ilişki çıkarımı modeli.	Önerilen sistem, yapılandırılmamış metinlerden tehdit varlıkları ve bunlar arasındaki ilişkileri daha yüksek doğrulukla çıkarmış; bağlamsal bilgi eşleştirme ile tehdit analiz süreçlerini optimize etmiştir.	Varlık çıkarımı görevinde %97.2 ve ilişki çıkarımı görevinde %98.5 F1 skoru elde edilerek, mevcut yöntemlerden üstün performans sergilemiştir.

Çizelge 2.1 (devam) : Siber tehdit istihbaratı ile ilgili literatür çalışmalarının incelenmesi.

Kaynak	Veri	Metot	Çıkarım	Sonuç
[33]	OpenCVE, MITRE ATT&CK, CAPEC ve çeşitli web sitelerinde yer alan yapılandırılmış siber güvenlik verileri.	Bidirectional Encoder Representations from Transformers (BERT), Grafik Dikkat Ağları (Graph Attention Networks - GAT)	Tehdit istihbaratı çıkarımı kapsamında geliştirilen model, varlık ve ilişki tespiti görevlerinde yüksek doğruluk sağlamıştır.	Varlık çıkarımı görevinde %90.16 ve ilişki çıkarımı görevinde %81.83 F1 skorları elde edilmiş; GAT modeli, bilgi çizgesi oluşturma sürecinde bağlamsal ilişkilerin doğru modellenmesini desteklemiş ve tehdit analizlerini optimize etmiştir.
[148]	Blog yazıları, APT (Advanced Persistent Threat) raporları, güvenlik bültenleri ve MITRE ATT&CK veri seti.	Bidirectional Encoder Representations from Transformers (BERT), çizge algoritmaları, çizge veri tabanı (Neo4j), tehdit analizi yaklaşımları.	BERT tabanlı model ile adlandırılmış varlıklar yüksek doğrulukla çıkarılmış; bu varlıklar arasındaki ilişkiler belirlenerek bilgi çizgeleri oluşturulmuştur.	Varlık çıkarımında %98.97 doğruluk ve %93.57 F1 skoru elde edilmiş; Neo4j tabanlı bilgi çizgeleri aracılığıyla tehdit aktörlerinin davranış örüntüleri analiz edilerek gelecekteki olası tehditler tahmin edilmiştir.

2.2 Çalışmanın Literatüre Katkısı

Bu çalışma, literatürde siber tehdit istihbaratı, doğal dil işleme, adlandırılmış varlık tanıma, derin öğrenme modelleri ve çizge tabanlı analiz alanlarına katkı sağlamayı hedeflemektedir. Literatürde mevcut çalışmalar incelendiğinde, siber tehdit istihbaratı süreçlerinin çoğunlukla geleneksel yöntemlerle ele alındığı ve çizge tabanlı analiz yaklaşımlarının bu alanda sınırlı bir şekilde uygulandığı görülmektedir. Ayrıca, BERT gibi bağlamsal dil modellerinin siber tehdit düzeyinde özelleştirilmesi ve ontoloji tabanlı yaklaşımın entegre edilmesi üzerine yapılan çalışmalar sınırlıdır. Bu bağlamda, çalışmamız hem akademik hem de pratik alanda çeşitli boşlukları doldurmaktadır.

İlk olarak, çalışmada, siber tehdit istihbaratı verilerinin analizinde derin öğrenme tabanlı BERT modelinin etkili bir şekilde kullanımı ele alınmıştır. Bu çalışma, BERT modeline siber tehdit terminolojisiyle ilgili ince ayar yaparak, bu alandaki bağlamsal varlık tanıma süreçlerini geliştirmiştir. Bu bağlamda, literatürdeki doğal dil işleme ve adlandırılmış varlık tanıma alanına önemli bir katkı sağlanmıştır.

İkinci olarak, çalışma, siber tehdit istihbaratına çizge tabanlı bir yaklaşım getirmektedir. Literatürde, tehdit aktörleri ve saldırı kampanyaları arasındaki ilişkilerin analizinde genellikle statik yöntemler kullanılmaktadır. Çalışmada, Neo4j tabanlı çizge veri modelleri kullanılarak siber tehdit varlıkları arasındaki ilişkiler modellenmiştir. Çizge tabanlı analiz yöntemleriyle, tehdit verilerinin düğüm ve kenarlar aracılığıyla görselleştirilmesi ve analiz edilmesi sağlanmıştır. Bu yöntem, tehdit aktörleri arasındaki potansiyel bağlantıların belirlenmesi, tehdit unsurlarının haritalanması ve tehdit zincirlerinin ortaya çıkarılması açısından yenilikçi bir yaklaşım sunmaktadır.

Son olarak, bu çalışma ontoloji tabanlı bir model önererek, varlıklar arasındaki bağlantıların daha sistematik bir şekilde yapılandırılmasına katkı sağlamaktadır. Bu yaklaşım, varlıklar ve ilişkiler arasındaki bağların daha anlamlı bir şekilde ortaya konulmasını sağlamıştır. Ayrıca, önerilen ontoloji tabanlı model sayesinde, tehditlerin ve ilişkili varlıkların görselleştirilmesi ve yorumlanması kolaylaşmakta, böylece güvenlik analistlerinin karar alma süreçleri desteklenmektedir. Bu çalışma, hem doğal dil işleme hem de çizge tabanlı analiz yaklaşımlarını bir araya getirerek siber tehdit istihbaratı alanında geniş ve etkili bir çözüm sunmaktadır. Bu birleşik yaklaşım, mevcut yöntemlerin sınırlılıklarını aşarak daha kapsamlı ve bütünleşik bir tehdit analizi sağlamaktadır.

3. ÇALIŞMANIN DAYANDIĞI TEMEL KAVRAMLAR

Bu bölümde, tez çalışmasının temelini oluşturan kavramlar sistematik bir şekilde ele alınmıştır. Öncelikle büyük veri kavramı, büyük verinin gelişimi ve temel özellikleri ele alınmıştır. Ardından siber güvenlik alanının önemli bir bileşeni olan siber tehdit istihbaratı kavramı incelenmiştir. Sonrasında, doğal dil işleme ve adlandırılmış varlık tanıma konularına değinilmiş ve son olarak, çizge veritabanları ve çizge tabanlı veri analizi yaklaşımları hakkında temel bilgilere yer verilmiştir.

3.1 Büyük Veri

Büyük veri, genellikle farklı kaynaklardan gelen, heterojen veri formatlarını da içeren, geniş ve sürekli büyüyen yapılandırılmış, yapılandırılmamış ve yarı yapılandırılmış büyük veri kümeleri olarak tanımlanmaktadır. Bu heterojen verilerin analizi, işlenmesi ve depolanması büyük veri alanının yetenekleriyle mümkün olmaktadır. Geleneksel veri analizi, işleme ve depolama teknolojileri ile yeterli performansı sağlayamadığında, büyük veri çözümleri ve uygulamaları devreye girmektedir. Özellikle büyük veri birden fazla ilgisiz veri kümesinin birleştirilmesi, büyük miktarlarda yapılandırılmamış verilerin işlenmesi ve gizli bilgilerin zamana duyarlı bir şekilde ortaya çıkarımı gibi farklı gereksinimleri ele almaktadır. Bu bağlamda büyük veri çözümleri, geleneksel yöntemlerin sınırlılıklarını aşarak daha hızlı ve etkili sonuçlar sunmaktadır.

3.1.1 Büyük verinin gelişimi

Verilerin depolanması ve analizi için özel bir çözüm olarak “veri tabanı” kavramı 1970'lerin sonlarında geliştirilmiştir. Artan veri miktarı karşısında geleneksel bilgisayar sistemlerinin kapasite sınırlarına ulaşılması, yeni çözümlerin arayışlarını başlatmıştır. Bu problemin üstesinden gelmek amacıyla araştırmacılar, 1980'lerde "Shared-Nothing" adı verilen paralel bir veri tabanı mimarisi geliştirmiştir [34]. Bu alanda ilk ticari başarıya ulaşan Teradata şirketi, 1986'da büyük bir firmanın veri ambarını genişletme projesine yardımcı olarak ilk paralel veri tabanı sistemini başarıyla uygulamıştır [35]. Paralel veri tabanı sistemlerinin sağladığı avantajlar 1990'ların sonunda geniş kabul görmüş ve bu teknoloji veri tabanı alanında yaygın olarak kullanılmaya başlanmıştır.

Bilimin önceki dönemlerinde bilimsel çalışmalar deneysel verilere, kuramsal bilgilere ve hesaplamalı verilere dayanırken, günümüzde bunlar birlikte kullanılarak “veri

keşfi" adı altında gerçekleştirilmektedir [36]. E-bilim oalrak adlandırılan bu yöntem ile toplanan bilimsel veriler analiz yazılımları ile analiz edilmekte ve depolanmaktadır. İnternet kullanımının yaygınlaşmasıyla birlikte depolanan veri miktarı da önemli ölçüde artmıştır. Bu gelişme, özellikle arama motoru şirketlerini yeni çözümler üretmeye yöneltmiştir. Örneğin Google, artan veri haciminin yönetim zorluklarının üstesinden gelmek için GFS (Google File System) [37] ve MapReduce [38] gibi teknolojiler geliştirmiştir. Büyük veri kavramının gelişiminde önemli bir dönüm noktası, Haziran 2011'de IDC (International Data Corporation) tarafından yayımlanan "Kaostan Değer Çıkarma" (Extracting Value from Chaos) başlıklı araştırma raporudur. Bu rapor, büyük veri kavramını ve potansiyelini ilk kez kapsamlı bir şekilde ele almıştır. Günümüzde birçok kuruluş, büyük veri ile ilgili projelere odaklanmış ve büyük veri kavramı birçok disiplinde yerini almıştır.

3.1.2 Büyük verinin boyutları

Büyük veri, geleneksel veri işleme araçlarının yetersiz kaldığı karmaşık ve büyük ölçekli veri kümelerini yönetmek için geliştirilen yeni bir teknolojik yaklaşımdır. Bu yaklaşımın en önemli özelliği, yapılandırılmamış verileri gerçek zamanlı olarak işleyebilme yeteneğidir. Bir veri kümesinin "büyük veri" olarak nitelendirilebilmesi için, analitik sistemlerin tasarım ve mimarisinde özel düzenlemeler gerektiren belirli özelliklere sahip olması gerekmektedir [39]. Analist Doug Laney'in geliştirdiği ve sektörde geniş kabul gören 3V modeli, büyük veriyi üç temel boyutta tanımlamaktadır. Bu boyutlardan hacim (Volume) veri kümesinin büyüklüğünü, hız (Velocity) verilerin oluşturulma ve işlenme hızını ve çeşitlilik (Variety) ise verilerin farklı türlerde ve farklı kaynaklardan geldiğini ifade etmektedir [40]. IBM ve Microsoft gibi teknoloji şirketleri uzun süre bu modeli benimsemiş olsalar da, bazı araştırmacılar modelin yetersiz kaldığını düşünerek dördüncü bir boyut olarak büyük veriye Değer (Value) tanımını da eklemişlerdir. Bu şekilde büyük veri için 4V modeli ortaya çıkmıştır. Bu yeni boyut, verilerden anlamlı içgörüler elde edilmesi gerekliliğine vurgu yapmaktadır. Bu genişletilmiş model, büyük veri analiz süreçlerinin daha kapsamlı ve etkili bir şekilde gerçekleştirilmesine olanak sağlamaktadır. Büyük verilerin dört özelliğine ek olarak bazı araştırmacılar gerçeklik (veracity), özelliğinden de bahsetmektedirler. Bu özellik büyük verilerin kalitesini ifade etmektedir [39]. Büyük verinin bahsedilen bu boyutlarını aşağıdaki gibi tanımlayabiliriz:

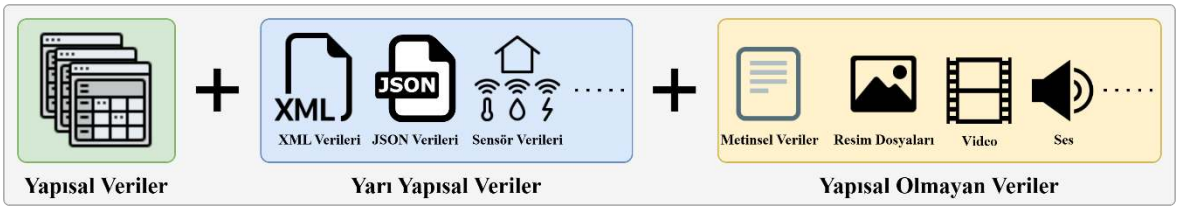
Hacim (Volume): Hacim, şu anda mevcut olan ve sürekli artan büyük veri miktarını ifade etmektedir. Teknolojik gelişmelerle beraber otomatikleşen cihazlar etrafımızdaki herşeyi kayıt altına almaktadır. Bu veriler örnek olarak aşağıdaki veri kaynaklarından elde edilebilir;

- Satış noktası ve bankacılık gibi çevrimiçi işlemlerden,
- Bilimsel araştırma deneylerinden,
- GPS sensörleri, RFID'ler, akıllı metreler, akıllı şehirler ve telematik gibi sensörler
- Facebook, Twitter, Instagram, LinkedIn ve YouTube gibi sosyal medya platformlarından,
- Çevresel veriler, iş verileri, medikal veriler, izleme verileri ve eğitim verileri gibi çeşitli kaynaklardan,
- Banka ATM'leri tarafından gerçekleştirilen tüm finansal işlemler, havaalanı sistemlerindeki yolcu hareketleri, otomatik kapılardaki giriş-çıkış kayıtları ve karayollarında kullanılan kameralar tarafından tespit edilen aşırı hız ihlalleri, trafik kazaları ve sürücü hataları gibi olaylardan elde edilen verilerden,
- Mobil cihazlar ve uygulamalardan elde edilen konum, kullanım ve etkileşim verilerinden.

Hız (Velocity): Hız kavramı, büyük veri bağlamında iki temel boyutta ele alınmaktadır. Bunlardan ilki, verilerin üretilme hızıdır [41]. Günümüzde her dakika devasa boyutlarda veri üretilmekte ve bu durum veri hacmini sürekli arttırmaktadır. İkincisi ise, bu verilerin işlenme ve analiz edilme hızıdır. Verinin ticari değerinin en üst düzeyde elde edilebilmesi için veri toplama, analiz ve bilgi çıkarım süreçlerinin mümkün olan en kısa sürede gerçekleştirilmesi gerekmektedir. Bu özellik, büyük veri yaklaşımını geleneksel veri madenciliği yöntemlerinden ayıran önemli faktörlerden biridir. Özellikle gerçek zamanlı analiz gerektiren uygulamalarda, verilerin hızlı bir şekilde işlenmesi kritik önem taşımaktadır.

Çeşitlilik (Variety): Çeşitlilik, büyük veri ekosisteminde bulunan farklı veri türlerini ifade etmektedir. Büyük veri, geleneksel yapılandırılmış verilerin ötesinde, ses kayıtları, video dosyaları, web sayfaları ve metin belgeleri gibi yarı yapılandırılmış ve yapılandırılmamış verileri de içermektedir. Yapılandırılmış veriler finansal işlem kayıtları gibi düzenli formattaki verileri kapsarken, e-postalar yarı yapılandırılmış verilere, görüntüler ise yapılandırılmamış verilere örnek olarak gösterilebilir. Ayrıca sosyal medya paylaşımları, sensör verileri, log kayıtları ve mobil uygulama verileri gibi farklı kaynaklardan elde edilen

çeşitli veri türleri de büyük veri kapsamına girmektedir. Bu farklı veri türlerinin varlığı, veri analizi süreçlerinde çeşitli teknolojilerin ve yaklaşımların bir arada kullanılmasını zorunlu kılmaktadır. Şekil 3.1, veri çeşitliliğini finansal işlemlerden imgelere, e-postalardan yapısal olmayan verilere kadar görsel olarak temsil etmektedir.

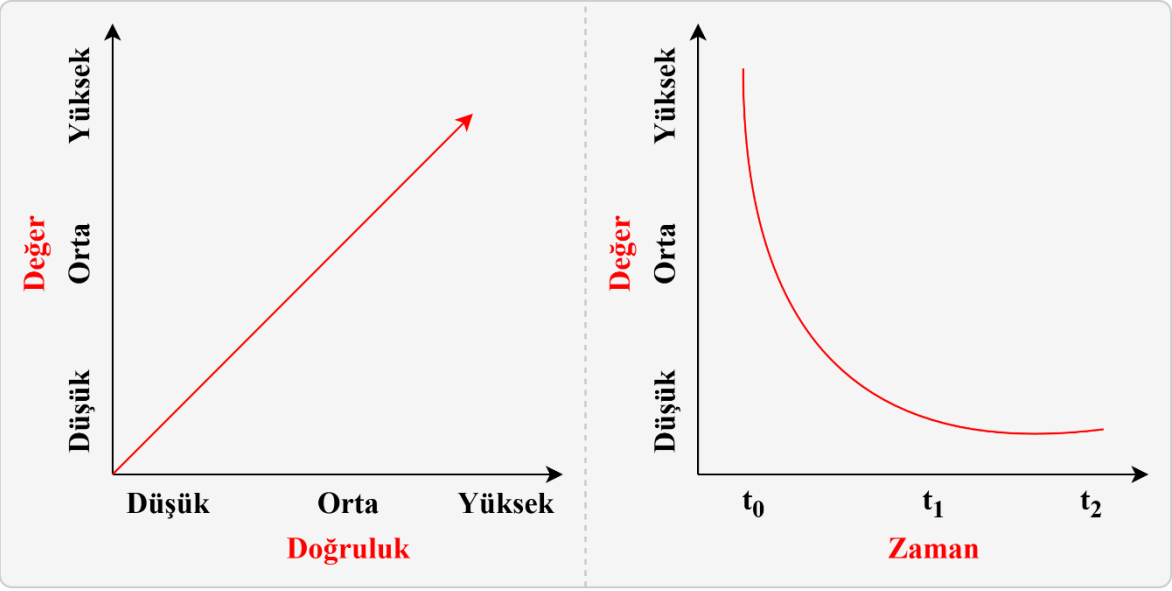


Şekil 3.1 : Yapılandırılmışlık türüne göre büyük veri türleri.

Doğruluk (Veracity): Doğruluk veya gerçeklik kavramı, verilerin kalitesi ve doğruluğu ile ilgilidir. Büyük veri ortamlarında verilerin kalite açısından değerlendirilmesi büyük önem taşımaktadır. Bu değerlendirme sürecinde, geçersiz verilerin ayıklanması ve veri gürültüsünün giderilmesi için titiz bir veri işleme süreci uygulanmaktadır. Veriler, bir veri kümesinde sinyal veya gürültü olarak sınıflandırılabilir. Gürültü, bilgiye dönüştürülemeyen ve dolayısıyla değer taşımayan verileri temsil ederken, sinyaller anlamlı bilgi içeren ve değer taşıyan verilerdir. Son yıllarda, yapay zeka, makine öğrenimi ve derin öğrenme tabanlı yaklaşımlar, veri kalitesini artırmak ve gürültüyü azaltmak amacıyla yaygın olarak kullanılmaktadır. Yüksek sinyal ve gürültü oranına sahip veri kümeleri, düşük oranlı veri kümelerine göre daha yüksek doğruluk seviyesine sahiptir. Örneğin, müşterilerden çevrimiçi kayıt yoluyla toplanan veriler, blog gönderileri gibi kontrol edilmeyen kaynaklardan elde edilen verilere kıyasla daha az gürültü içermektedir. Bu durum, verilerin analiz hızını ve doğruluğunu doğrudan etkilemektedir. Gürültü oranı yüksek veri kümelerinde kapsamlı temizleme işlemleri gerektiğinden, veri analizi ve bilgi çıkarımı süreci uzamaktadır. Bu gecikme, büyük verinin temel özelliklerinden biri olan hız karakteristiğiyle çelişmektedir. Bu nedenle, günümüzde veri kalitesini artırmak için otomatik veri temizleme ve doğrulama süreçleri büyük veri sistemlerinin ayrılmaz bir parçası haline gelmiştir.

Değer (Value): Değer kavramı, yapılandırılmış ve yapılandırılmamış verilerin kayıpsız bir şekilde anlamlı bilgilere dönüştürülmesi sürecini ifade etmektedir. Bu kavram aynı zamanda, bir işletmenin veri analizi sonucunda elde ettiği fayda olarak da tanımlanmaktadır. Büyük verinin değer boyutu, doğruluk boyutu ile yakından ilişkilidir, çünkü verinin doğruluk seviyesi arttıkça, elde edilen bilginin değeri de artmaktadır. Son yıllarda, büyük veri analitiği ve yapay zekâ yöntemlerinin gelişmesiyle birlikte, verilerden değer elde etme süreçleri

önemli ölçüde hızlanmış ve iyileşmiştir. Ayrıca, değer boyutu veri analiz işlemlerinin ne kadar sürdüğüne de bağlıdır çünkü analiz sonuçlarının belirli bir kullanım süresi vardır. Örneğin, hisse senedi işlemlerinde 20 dakikalık gecikmeyle verilen bir emir, 20 milisaniyelik gecikmeyle verilen bir emre kıyasla çok daha düşük bir değere sahiptir ve çoğu durumda hiçbir değer taşımamaktadır [39, 42]. Örnekte de görüldüğü gibi veriler için değer ve zaman ters orantılıdır. Verilerin anlamlı bilgilere dönüştürülmesi ne kadar uzun sürerse, verilerden çıkarılan değer de o kadar az olur. Bu nedenle, günümüzde gerçek zamanlı veri analizi ve hızlı karar alma süreçleri, büyük veri uygulamalarında kritik önem kazanmıştır. Şekil 3.2'de, verilerin doğruluğu ve çıkarılan bilgilerin güncelliği ile değerin nasıl etkilendiğine ilişkin iki örnek grafik gösterilmektedir.



Şekil 3.2 : Verinin değerinin, doğruluk ve zaman ile ilişkisi.

3.1.3 Büyük veri ile ilgili zorluklar

Büyük veri alanındaki hızlı veri artışı, veri yönetimi süreçlerinde önemli zorluklar yaratmaktadır. Geleneksel veri yönetimi sistemleri, temelde ilişkisel veri tabanı yönetim sistemlerine dayanmaktadır. Ancak bu sistemler sadece yapılandırılmış veriler için uygun olup, yarı yapılandırılmış veya yapılandırılmamış verileri işlemekte yetersiz kalmaktadır. Ayrıca, bu sistemler giderek daha maliyetli donanım gereksinimleri ortaya çıkarmakta ve büyük verilerin hacmi ve çeşitliliği karşısında performans sorunları yaşamaktadır. Bunun yanı sıra, geleneksel sistemlerin ölçeklenebilirlik ve gerçek zamanlı veri işleme yetenekleri sınırlı olduğundan, büyük veri uygulamalarının ihtiyaçlarını karşılamakta güçlük çekmektedirler. Bu sorunlara çözüm olarak araştırmacılar alternatif yaklaşımlar

geliştirmişlerdir. Örneğin bulut bilişim, büyük veri altyapı gereksinimlerini maliyet verimliliği, esneklik ve kolay güncellenebilirlik açısından karşılamaktadır. Büyük ölçekli düzensiz veri setlerinin depolanması ve yönetimi için bulut sistemlerde dağıtılmış dosya sistemleri ve NoSQL veritabanları kullanılmaktadır. Son yıllarda, Apache Hadoop ve Apache Spark gibi açık kaynaklı büyük veri platformları da bu zorlukların üstesinden gelmek için yaygın olarak tercih edilmektedir. Bu çalışmada, büyük veri alanında karşılaşılan zorluklar veri, veri işleme ve yönetimsel zorluklar olarak üç ana kategoride altında incelenmiştir.

3.1.3.1 Verilerle ilgili zorluklar

Veri ile ilgili zorluklar, verilerin karakteristiklerine ve özelliklerine ilişkin zorlukları ifade etmektedir. Bazı araştırmacılar veriyi 3V olarak ifade ederken, bazıları 4V, bazıları 6V ve bazı araştırmacılar ise 7V kavramlarını öne sürmüşlerdir. Bu çalışmada literatürde en sık ifade edilen zorluklar ele alınmıştır.

Hacim: Dünya genelinde depolanan veri miktarı her dakika katlanarak artmaktadır. Meta, X, Instagram ve YouTube gibi sosyal medya platformları günlük olarak terabaytlar seviyesinde veri üretmektedir. Günümüzde çevresel verilerden iş verilerine, medikal verilerinden izleme verilerine kadar her türlü veri, kullanılmak üzere kayıt altına alınmaktadır. Teknolojinin gelişmesiyle birlikte otomatikleşen cihazlar çevremizdeki tüm aktiviteleri kaydetmektedir. Banka ATM'lerindeki finansal işlemler, havaalanlarındaki yolcu işlemleri, otomatik kapılardaki giriş-çıkış kayıtları ve karayollarındaki kameraların kaydettiği aşırı hız ihlalleri, kazalar ve sürücü hataları bu duruma örnek olarak gösterilebilir. Bu veri artışı iki temel problemi beraberinde getirmektedir. Bunlardan birincisi, veri yığınları artarken işlenebilen veri yüzdesinin azalması; ikincisi ise, büyük ham veri yığınlarının düşük maliyetle depolanması gerekliliğidir. Geleneksel veritabanları, bu büyüklükteki verileri saklama ve yönetme konusunda yetersiz kalmaktadır. Bu nedenle, Hadoop, Apache Spark ve bulut tabanlı depolama çözümleri gibi ölçeklenebilir ve dağıtık sistemler yaygın olarak tercih edilmektedir.

Çeşitlilik: Büyük verinin temel zorluklarından biri, verilerin çok çeşitli formatlarda bulunmasıdır. Bu veriler yapılandırılmış ve yapılandırılmamış metin, görüntü, çoklu ortam içeriği, ses dosyaları ve sensör verileri gibi farklı biçimlerde olabilmektedir. Literatür çalışmaları, bu veri türlerinin tutarlı bir format yapısına sahip olmadığını göstermektedir. Bu durum, veri bilimcilerin farklı kaynaklardan gelen verileri birleştirme sürecinde önemli

zorluklarla karşılaşmasına neden olmaktadır. Bu veri çeşitliliğine örnek olarak metin mesajları, e-postalar, tweetler, blog yazıları gibi mesajlaşma verileri; kullanıcı içerikleri; web günlükleri ve ticari işlem kayıtları gibi işlem verileri; veri ağırlıklı deneylerden elde edilen bilimsel veriler ve sağlık verileri; ayrıca sosyal medyada paylaşılan resim, video ve sensör verileri gibi web verileri gösterilebilir. Son yıllarda IoT (Nesnelerin İnterneti) cihazlarından elde edilen veriler ve artırılmış gerçeklik uygulamalarından gelen karmaşık veri türleri de bu çeşitliliğe eklenmiştir.

Hız: Büyük verinin bir diğer önemli zorluğu, homojen olmayan yapıdaki yüksek hızlı veri akışının yönetilmesidir. Bu zorluk özellikle, dijital cihazların yaygınlaşmasıyla elde edilen ve gerçek zamanlı analiz gerektiren büyük, karmaşık ağlar üzerinden üretilen veri kümeleri için geçerlidir. Örneğin Twitter'ın saatte bir milyondan fazla işlem gerçekleştirmesi bu duruma örnek gösterilebilir. Mobil cihazlar, uygulamalar ve mağaza kartlarından elde edilen veriler, müşteriler için gerçek zamanlı ve kişiselleştirilmiş teklifler oluşturulmasına olanak sağlayan bilgi akışları üretmektedir. Bu veriler aynı zamanda müşterilerin coğrafi konumları, satın alma davranışları ve örüntüleri gibi önemli bilgileri içermekte ve müşteri profillerinin gerçek zamanlı olarak oluşturulmasında kullanılmaktadır. Bu nedenle, Apache Kafka, Apache Flink ve gerçek zamanlı veri akışı platformları gibi teknolojiler, yüksek hızlı veri akışlarını yönetmek için yaygın olarak kullanılmaktadır.

Doğruluk: Verilerin doğruluğu sadece veri kalitesiyle sınırlı olmayan, daha geniş kapsamlı bir zorluktur. Bu özellik, verilerin kesinliğini ve analiz için kullanılabilirlik düzeyini ölçmektedir. Toplanan verilerde doğal farklılıklar bulunabildiğinden, verilerin doğru anlaşılması kritik önem taşımaktadır. Sistemlerde depolanan veri kümelerinin doğruluğu, analiz çalışmalarından elde edilecek sonuçları ve değerleri doğrudan etkilediğinden, doğruluk özelliği büyük veri alanındaki en kritik zorluklardan biri olarak kabul edilmektedir. Son dönemde, veri doğruluğunu artırmak için makine öğrenimi tabanlı otomatik veri temizleme ve doğrulama yöntemleri yaygınlaşmıştır.

Oynaklık: Oynaklık, verilerin geçerlilik süresini ve veritabanlarında tutulması gereken süreyi ifade etmektedir. Verilerin ne zaman geçerliliğini yitirdiğini belirlemek, büyük veri yönetiminin önemli zorluklarından biridir. Bu nedenle, verilerin analizi için artık geçerli olmadığı zamanı doğru tespit etmek büyük önem taşımaktadır. Özellikle KVKK gibi veri koruma düzenlemeleri nedeniyle, verilerin saklama sürelerinin belirlenmesi ve yönetilmesi kritik hale gelmiştir.

Görselleştirme: Veri görselleştirmesi, anahtar bilgilerin çeşitli görsel biçimler kullanılarak grafiksel düzende daha sezgisel ve etkili bir şekilde sunulmasıdır. Chen ve Zhang'in araştırmalarına göre, veri görselleştirmesi için kullanılan birçok uygulama, işlevsellik, ölçeklenebilirlik ve tepki süresi açısından yetersiz performans göstermektedir. Bu performans sorunlarının temel nedeni olarak büyük veri hacimleri ve boyutları gösterilmektedir [43]. Son yıllarda, Tableau, Neo4j, Power BI ve QlikView gibi gelişmiş görselleştirme araçları, büyük veri kümelerinin görselleştirilmesi ve analiz edilmesi için yaygın olarak tercih edilmektedir. Ayrıca, etkileşimli ve gerçek zamanlı görselleştirme yöntemleri, büyük veri analiz süreçlerinde giderek daha fazla önem kazanmaktadır.

3.1.3.2 Veri işleme ile ilgili zorluklar

Bu zorluk, verilerin nasıl ele alınacağı, entegre edileceği, dönüştürüleceği ve doğru analiz tekniğinin nasıl seçileceği gibi konuları kapsamaktadır. Bu kapsamda karşılaşılan zorluklar, verilerin toplanmasından işlenip analiz edilmesine, sonuçların çıkarılmasından görselleştirilip sunulmasına kadar olan tüm süreçleri içermektedir. Büyük veri kümelerinin genellikle ilişkisel olmayan veya yapılandırılmamış verilerden oluşması, bu verilerin işlenmesini özellikle zorlaştırmaktadır. Bu nedenle, geleneksel veri işleme yöntemleri yerine, dağıtık ve paralel işlem yeteneklerine sahip yeni nesil büyük veri platformları tercih edilmektedir.

Veri Toplama ve Depolama: Çeşitli kaynaklardan veri elde edilmesi ve bu verilerin depolanması süreçleriyle ilgilidir. Büyük verilerin karmaşık yapısı ve sürekli artan hacmi, veri toplama ve depolama aşamalarında önemli problemler oluşturmaktadır [44]. Değerli bilgilerin elde edilebilmesi için yararlı verilerin ayıklanması ve gereksiz veya tutarsız verilerin elenmesi amacıyla güçlü ve akıllı filtreleme sistemlerinin kullanılması gerekmektedir. Bu filtreleme ihtiyacı, veri toplama ve depolama sürecinin zorluklarından biri olarak değerlendirilmektedir. Son yıllarda, makine öğrenimi ve yapay zekâ tabanlı otomatik filtreleme ve veri temizleme yöntemleri, bu süreçlerde etkin çözümler sunmaktadır. Bunun yanında, veri kaynaklarının doğru anlaşılması, geniş akış verilerinin işlenmesi ve depolama öncesinde veri boyutunun azaltılması için verimli analitik algoritmaların geliştirilmesi de büyük önem taşımaktadır [45]. Bu bağlamda Apache Kafka, Apache NiFi gibi gerçek zamanlı veri akışı platformları ve Apache Hadoop, Apache Spark gibi dağıtık veri işleme sistemleri, veri toplama ve depolama süreçlerinde yaygın olarak

kullanılmaktadır. Ayrıca, bulut tabanlı depolama çözümleri, ölçeklenebilirlik ve maliyet etkinliği açısından önemli avantajlar sağlamaktadır.

Veri Madenciliği ve Veri Temizleme: Bu süreç, büyük miktardaki yapılandırılmamış verilerin düzenlenmesi, arındırılması ve analiz için hazır hale gelmesini kapsamaktadır. Büyük veri uzmanları, verilerin titizlikle analiz edilmesi ve temizlenmesinin, elde edilen bilginin kalitesini ve değerini önemli ölçüde artırdığını vurgulamaktadır [46]. Büyük veri kaynaklarında sıklıkla gürültülü, hatalı veya eksik verilerle karşılaşılır. Temel zorluk, bu geniş veri kümelerinin nasıl temizleneceğini belirlemek ve verilerin güvenilirliğini değerlendirmektir. Son yıllarda, veri temizleme süreçlerinde makine öğrenimi ve derin öğrenme tabanlı otomatik yöntemler yaygın olarak kullanılmaktadır. Bu yöntemler, büyük veri kümelerindeki hatalı ve eksik verilerin daha hızlı ve doğru bir şekilde tespit edilmesini sağlamaktadır. Hatalı verilerin tespiti kritik öneme sahiptir, çünkü bu veriler analiz edildiğinde yanıltıcı sonuçlar ortaya çıkarabilir ve hatalı kararlara yol açabilir. Örneğin, akıllı bir trafik kavşağındaki arızalı sensörler, araç sayısının yanlış ölçülmesine neden olabilir. Bu durum, trafik ışıklarının zamanlamasının hatalı hesaplanmasına ve trafik sıkışıklığına yol açabilir. Benzer şekilde, sağlık sektöründe eksik veya hatalı hasta verileri, yanlış teşhis ve tedavi kararlarına neden olarak ciddi sonuçlar doğurabilir. Bu nedenle, veri temizleme ve doğrulama süreçleri, büyük veri analitiğinin temel bileşenlerinden biri olarak kabul edilmektedir.

Veri Birleştirme ve Veri Entegrasyonu: Bu süreç, farklı kaynaklardan elde edilen temizlenmiş ve analiz için hazır hale getirilmiş yapılandırılmamış verilerin bir araya getirilmesi ve bütünleştirilmesini kapsamaktadır. Büyük veri genellikle sosyal medya etkileşimleri (X, Meta, Instagram ve YouTube gönderileri/beğenileri gibi) gibi farklı anlamlar taşıyan çevrimiçi aktiviteleri içermektedir [47]. Bu farklı karakteristikteki verilerin doğal yapısı gereği ortak bir bağlantı noktası veya standardı bulunmamaktadır. Diğer bir deyişle bu verileri birleştirmek için belli bir standart yoktur.

Büyük veri alanındaki temel zorluklardan biri, çeşitli kaynaklardan gelen yapılandırılmamış verilerin toplanması ve mevcut verilerle entegrasyonudur. Sistemler arası uyumluluğun sağlanması için standartların belirlenmesi ve bilginin etkili bir şekilde bütünleştirilmesi hayati önem taşır. Bu noktada en büyük engel, farklı veri havuzlarında kullanılan meta verilerdeki tutarsızlıklardır. Meta veriler için ortak standartların olmaması, büyük veri uygulamalarında üretilen verilerin entegrasyonunu önemli ölçüde zorlaştırmaktadır. Son yıllarda, bu entegrasyon sorunlarını aşmak için Apache NiFi, Apache

Airflow gibi veri akışı yönetim araçları ve bulut tabanlı veri entegrasyon platformları yaygın olarak kullanılmaktadır. Ayrıca, veri gölleri (data lakes) ve veri ambarları (data warehouses) gibi merkezi veri depolama çözümleri, farklı kaynaklardan gelen verilerin entegrasyonunu kolaylaştırmakta ve standartlaştırılmış veri yönetimini desteklemektedir. Bununla birlikte, veri entegrasyonu süreçlerinde yapay zekâ ve makine öğrenimi tabanlı otomatik eşleştirme ve entegrasyon yöntemleri de giderek daha fazla tercih edilmektedir.

Veri Analizi: Büyük veri, heterojen yapısı ve büyük hacmi ile ifade edilmektedir. Ayrıca büyük veriler dinamik bir yapıya da sahip olabilmektedir. Büyük verinin bu özellikleri, özellikle gerçek zamanlı analizlerde, verilerin modellenmesi ve analizi sürecini zorlaştırmaktadır. Farklı kaynaklardan gelen heterojen ve değişken yapıdaki verilerin modellenmesi, ancak verilerin alt kategorilere ayrılması ve her kategori için özel modeller geliştirilmesi ile mümkün olmaktadır. Alternatif olarak, makine öğrenmesinin yüksek boyutlu veri işleme kapasitesinden yararlanılarak, derin öğrenme algoritmaları ile büyük veri modellenebilmektedir [48]. Son yıllarda, büyük veri analizi süreçlerinde Apache Spark, TensorFlow ve PyTorch gibi ölçeklenebilir ve dağıtık makine öğrenimi platformları yaygın olarak kullanılmaktadır. Ayrıca, otomatik makine öğrenimi yöntemleri, büyük veri kümelerinin analizinde model geliştirme süreçlerini hızlandırmakta ve kolaylaştırmaktadır. Bununla birlikte, büyük veri analizinde karşılaşılan temel zorluklardan biri, analiz sonuçlarının yorumlanabilirliği ve açıklanabilirliğidir. Bu nedenle, açıklanabilir yapay zekâ (Explainable AI - XAI) yöntemleri, büyük veri analiz süreçlerinde giderek daha fazla önem kazanmaktadır.

3.1.3.3 Yönetimsel zorluklar

Birçok veri ambarında kişisel veriler, finansal bilgiler, hasta bilgileri gibi hassas veriler bulunmaktadır. Bu verilere erişimle ilgili yasal ve etik kurallar, verilerin korunmasını ve yalnızca yetkili kişiler tarafından kullanılmasını zorunlu kılmaktadır. Bu nedenle, verilerin güvence altına alınması, yetkisiz erişimlerin engellenmesi ve erişimlerin sürekli kontrol edilmesi büyük önem taşımaktadır. Son yıllarda kişisel verilerin korunması kanunu gibi veri koruma düzenlemeleri, veri yönetimi süreçlerinde uyulması gereken standartları ve yükümlülükleri artırmıştır.

Verilerin yönetiminde en sık karşılaşılan problemler; veri gizliliğinin sağlanması, veri güvenliğinin korunması ve etik kurallara uyumdur. Bu problemlerin çözümü için verilerin doğru kullanıldığından emin olmak, amaçlanan kullanımların dışına çıkılmamasını

sağlamak ve ilgili yasal düzenlemelere tam uyum göstermek gereklidir. Ayrıca, verilerin nasıl kullanıldığını, dönüştürüldüğünü ve türetildiğini izlemek, verinin yaşam döngüsünü etkin bir şekilde yönetmek, güvenilir ve sürdürülebilir bir veri yönetimi için kritik bir adımdır. Bu bağlamda, veri yönetişimi politikalarının oluşturulması ve uygulanması, veri kalitesi yönetimi, veri erişim kontrolü ve denetim mekanizmalarının geliştirilmesi büyük önem kazanmıştır. Ayrıca, blok zinciri teknolojisi gibi yenilikçi yaklaşımlar, veri yönetiminde şeffaflık, güvenlik ve izlenebilirlik sağlamak amacıyla giderek daha fazla tercih edilmektedir. Sonuç olarak, büyük veri yönetiminde karşılaşılan zorlukların üstesinden gelmek için, teknolojik çözümlerin yanı sıra organizasyonel politikaların ve süreçlerin de sürekli olarak güncellenmesi ve geliştirilmesi gerekmektedir.

3.1.4 Büyük veri tipleri

Büyük veri platformları tarafından işlenen veriler, insan ve makineler tarafından üretilmektedir. İnsanlar tarafından üretilen veriler, sosyal medya platformları, çevrimiçi hizmetler, e-ticaret siteleri ve mobil uygulamalar gibi sistemlerle insan etkileşimlerinin sonucu oluşan verilerdir. Makine tarafından üretilen veriler, gerçek dünya olaylarına yanıt olarak yazılım programları ve donanım cihazları tarafından üretilir. Bu tür verilere sensör verileri, IoT cihazlarından gelen veriler, log kayıtları ve otomatik sistemlerin ürettiği işlem verileri örnek olarak gösterilebilir. İnsanlar ve makineler tarafından üretilen veriler çeşitli kaynaklardan elde edilebilir ve çeşitli formlarda veya türlerde temsil edilebilir. Bu farklı veri türlerinin yönetimi ve analizi için farklı teknikler ve araçlar kullanılmaktadır. Bu çalışmada veri türleri üç grupta ele alınmıştır. Bu veri türlerinin karşılaştırmalı özellikleri Çizelge 3.1'de sunulmuştur.

Çizelge 3.1 : Yapısal, yarı yapısal ve yapısal olmayan verilerin karşılaştırılması.

Özellikler	Yapısal Veriler	Yarı Yapısal Veriler	Yapısal Olmayan Veriler
Veri Modeli	Belirli bir şemaya uygun	Kısmen şemaya uygun	Şemaya uygun değil
Depolama Biçimi	İlişkisel veritabanları (tablolar)	XML, JSON gibi metin tabanlı dosyalar	Dosya sistemleri, bulut depolama
Sorgulama Yöntemi	SQL ile sorgulanabilir	Özel araçlar ve sorgu dilleri (XPath, JSONPath)	Özel analiz araçları (NLP, görüntü işleme)

3.1.4.1 Yapısal veriler

Yapısal veriler, belirli bir veri modeline veya şemaya uygun olarak düzenlenmiş, genellikle tablo biçiminde saklanan verilerdir. Bu veriler, ilişkisel veritabanlarında depolanır ve SQL kullanılarak sorgulanabilir. Yapısal veriler, önceden tanımlanmış uzunluk, tip ve formata sahiptir. Yapısal verilere ait örnekler Çizelge 3.2'de sunulmuştur.

Çizelge 3.2 : Yapısal veriler.

Üretim Şekli	Veri Örnekleri
Makine Tarafından Üretilen	GPS verileri, RFID etiket verileri, tıbbi cihaz verileri, akıllı sayaç verileri, web sunucu logları, finansal işlem verileri
İnsan Tarafından Üretilen	Kullanıcı giriş bilgileri (isim, adres, yaş vb.), tıklama akışı verileri (müşteri davranışları), oyun içi hareket verileri

3.1.4.2 Yapısal olmayan veriler

Bir veri modeline veya veri şemasına uymayan veriler, yapılandırılmamış veriler olarak bilinir. Yapılandırılmamış veriler yapılandırılmış verilere göre daha hızlı bir büyüme oranına sahiptir [49]. Genellikle değişken uzunluk, tip veya formata sahip verilerdir. Dünyadaki verilerin çoğu yapılandırılmamış verilerdir. Yapısal olmayan verilere ait örnekler Çizelge 3.3'te sunulmuştur.

Çizelge 3.3 : Yapısal olmayan veriler.

Üretim Şekli	Veri Örnekleri
Makine Tarafından Üretilen	Uydu görüntüleri (hava durumu, coğrafi görüntüler), bilimsel veriler (sismik, atmosferik veriler), güvenlik ve trafik kameralarından elde edilen fotoğraf ve videolar, radar ve sonar verileri
İnsan Tarafından Üretilen	Belgeler (günlükler, anket sonuçları, e-postalar), sosyal medya içerikleri (YouTube, Facebook, X vb.), mobil cihazlardan gelen metin mesajları ve konum bilgileri, web sitelerindeki yapılandırılmamış içerikler

3.1.4.3 Yarı yapısal veriler

Yarı yapısal veriler, yapılandırılmış formattaki bazı özelliklerin yapılandırılmamış şekilde temsil edildiği yapısal ve yapısal olmayan verilerin birleşiminden oluşmaktadır. Yarı yapısal verilerde bilgi kolayca ayrıştırılabilen alanlar halinde düzenlenmiştir. Ancak veriler analizlerde kullanılmadan önce veri alanlarının ön işlemeye ihtiyacı olabilir [50]. Bu tür veriler genellikle metin dosyalarında saklanır. XML ve JSON dosyaları yarı yapısal veriler

için örnek olarak gösterilebilir. Bu tür verilerin metinsel niteliği ve bazı yapılara uygunluğu nedeniyle, yapısal olmayan verilere göre daha kolay işlenir. Yarı yapısal verilere ait örnekler Çizelge 3.4'te sunulmuştur.

Çizelge 3.4 : Yarı yapısal veriler.

Üretim Şekli	Veri Örnekleri
Makine Tarafından Üretilen	Log dosyaları (kısmen yapılandırılmış), sensör verileri (XML, JSON formatında)
İnsan Tarafından Üretilen	XML dosyaları, JSON dosyaları, HTML içerikleri, e-posta içerikleri (başlık ve gövde yapısı olan)

3.1.5 Büyük veri üretimi

Günümüzde küresel veri hacmi sürekli artış göstermektedir. Çevremizdeki sensörler, yazılım kayıtları, kameralar, mikrofonlar ve kablosuz sensör ağları gibi çeşitli cihazlar sürekli olarak yeni veriler üretmekte ve toplamaktadır. Küresel ölçekte günlük 2,5 kentilyon bayt veri üretilmekte ve bu miktar yıllar geçtikçe katlanarak artmaya devam etmektedir [51].

Veri üretimi, büyük veri yaşam döngüsünün başlangıç aşamasını oluşturmaktadır. Veri kaynakları çeşitlilik göstermekte olup sensörler, video kayıtları, kullanıcı tıklama verileri ve diğer uygun veri kaynaklarını kapsamaktadır. Büyük veri kaynakları arasında işletmelerin ticari verileri, bilgi teknolojilerindeki lojistik veriler, internet üzerindeki insan etkileşimleri, sosyal medya paylaşımları, konum bilgileri ve bilimsel araştırmalardan elde edilen veriler bulunmaktadır. Bu verilerin hacmi, mevcut bilişim teknolojilerinin depolama kapasitesini aşmaktadır. Ayrıca, büyük verilerin gerçek zamanlı işlenmesi ihtiyacı, mevcut depolama ve işlem kapasitesi üzerinde önemli bir baskı oluşturmaktadır. Bu verilerin hacmi, mevcut bilişim teknolojilerinin depolama kapasitelerini aşmakta ve sürekli olarak yeni depolama çözümlerine ihtiyaç duyulmaktadır. Ayrıca, büyük verilerin gerçek zamanlı olarak işlenmesi gerekliliği, mevcut depolama ve işlem kapasiteleri üzerinde önemli bir baskı oluşturmakta ve daha güçlü, ölçeklenebilir ve hızlı veri işleme sistemlerinin geliştirilmesini zorunlu kılmaktadır. Bu nedenle, bulut bilişim, dağıtık veri işleme platformları ve yüksek performanslı hesaplama sistemleri gibi yenilikçi teknolojiler büyük veri üretimi ve yönetiminde giderek daha fazla önem kazanmaktadır.

3.1.6 Büyük verilerin elde edilmesi

Veri edinimi, büyük veri yaşam döngüsünün ikinci aşamasını oluşturur ve verilerin depolama sistemlerine aktarılmadan önce toplanması, filtrelenmesi ve temizlenmesi

süreçlerini kapsamaktadır [52]. Toplanan veri kümeleri genellikle gereksiz bilgiler içerebilmekte, bu durum hem depolama alanının verimsiz kullanımına hem de sonraki analiz süreçlerinin etkinliğinin azalmasına neden olabilmektedir. Bu nedenle verilerin aktarılmadan önce filtreleme ve temizleme işlemlerinden geçirilerek depolanması önem arz etmektedir.

Büyük verilerin elde edilmesi ve depolama altyapısına aktarımı genellikle iki aşamada gerçekleştirilir [53]. Bunlar; IP omurga iletimi ve veri merkezi aktarımıdır. IP omurgası, belirli bir bölgedeki büyük verilerin üretildiği kaynaktan veri merkezine aktarımı için yüksek kapasiteli iletim hatları sağlar. Bu iletim hatlarının hızı ve kapasitesi, kullanılan fiziksel ortam (örneğin fiber optik kablolar) ve bağlantı yöntemlerine bağlı olarak değişkenlik gösterir. Veri merkezi aktarımı ise, verilerin ön işleme ve analiz gibi süreçler için veri merkezlerine transfer edilmesini içerir [54].

Etkili karar verme süreçleri için kaliteli ve güvenilir veriler gerekmektedir. Bu nedenle, büyük verilerin analiz öncesinde dikkatli bir şekilde aktarılması ve kalite doğrulamasından geçirilmesi kritik önem taşımaktadır. Bu kapsamda, verilerin analiz öncesinde ön işleme sürecinden geçirilmesi gereklidir. Ön işleme sürecinde özellikle dikkat edilmesi gereken hususlar; bozuk kayıtlar, hatalı içerikler, eksik değerler ve format tutarsızlıklarıdır. Örneğin, bir veri setinde müşteri yaşı tam sayı olarak kaydedilirken (25), başka bir veri setinde ondalıklı sayı (25.0) veya metin formatında ("yirmi beş") kaydedilmiş olabilir. Benzer şekilde, ülke bilgisi bir veri setinde "Türkiye" olarak kaydedilirken, diğerinde "TR" veya "TUR" gibi farklı kod sistemleriyle kayıt altına alınmış olabilir. Bu tür tutarsızlıklar, analiz süreçlerinde hatalara ve yanlış yorumlamalara neden olabilmektedir. Ön işleme aşamasındaki temel zorluk, farklı kaynaklardan gelen verilerin standart formatlara dönüştürülerek analiz süreçlerine hazır hale getirilmesidir. Bu aşamada yapılacak hatalar, sonraki analiz süreçlerini olumsuz etkileyerek hatalı kararlara yol açabilmektedir. Bu nedenle, ön işleme aşaması büyük veri analiz süreçlerinin başarısı açısından kritik öneme sahiptir.

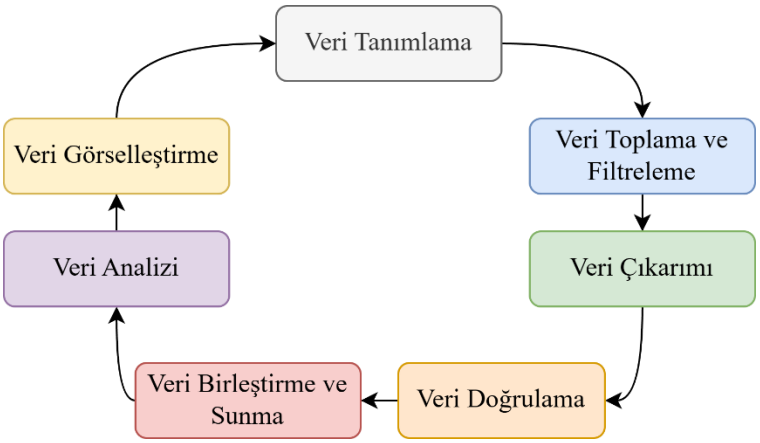
3.1.7 Büyük veri analizi

Geleneksel veri analizinde veriler genellikle ilişkisel veritabanlarında saklanmaktadır. İlişkisel veritabanlarının birçoğu, çıkarma-dönüştürme-yükleme (Extract-Transform-Load - ETL) işlemlerinden ve veri tabanı tablolarına veri girerken kullanılan kısıtlamalardan dolayı verileri düzgün bir formatta, temizlenmiş ve meta verilerine uygun

bir şekilde saklamaktadır. Bu durum, geleneksel veri analiz süreçlerinde verilerin daha kolay ve tutarlı bir şekilde analiz edilmesini sağlamaktadır. Büyük veri analizlerinde ise durum farklıdır. Çeşitli kaynaklardan gelen farklı formatlardaki büyük veriler genellikle yapılandırılmamış veya yarı yapılandırılmış olabilmekte ve hata içerebilmektedir. Bu durum büyük verilerin analiz edilmesini ve depolanmasını zorlaştırmakla birlikte, aynı zamanda bu verilerden daha fazla bilgi çıkarımı için yeni fırsatlar da sunmaktadır [55]. Geleneksel veri analizi, ilişkisel veri modelleri üzerine kurulmuştur. Bu sistemlerde ilgili alanlar arasında ilişkiler kurulur ve analizler bu ilişkilere bağlı olarak gerçekleştirilir. Ancak büyük veri analizinde elde edilen veriler çoğunlukla yapılandırılmamış veya yarı yapılandırılmış olduğundan, geleneksel yöntemler büyük veri analizi için yetersiz kalmaktadır. Ayrıca, geleneksel veri analizi yöntemleri genellikle yığın (batch) odaklıdır ve ETL işlemlerinden dolayı analiz süreçlerinde gecikmeler yaşanabilmektedir. Büyük veri analizinde ise verilerin değişimi oldukça dinamik olduğundan, verilerin gerçek zamanlı veya gerçek zamana yakın (near real-time) olarak analiz edilmesi gerekmektedir. Bu nedenle, büyük veri analiz süreçlerinde Apache Spark, Apache Flink gibi gerçek zamanlı veri işleme platformları ve NoSQL veritabanları gibi yeni nesil teknolojiler tercih edilmektedir.

3.1.8 Büyük veri yaşam döngüsü

Büyük veri analizi, geleneksel veri analizinden temel olarak üç karakteristik özelliği ile ayrılır: hacim, hız ve çeşitlilik. Bu özellikler, büyük veri projelerinin yönetimini ve analiz süreçlerini daha karmaşık hale getirmekte ve geleneksel yöntemlerin yetersiz kalmasına neden olmaktadır. Bu nedenle, büyük veri projelerinde verilerin tanımlanmasından başlayarak analiz edilip sonuçların görselleştirilmesine kadar geçen süreçleri kapsayan sistematik bir yaşam döngüsüne ihtiyaç duyulmaktadır. Bu süreç, büyük veri projelerinin başarılı bir şekilde yürütülmesi ve yönetilmesi için kritik öneme sahiptir. Literatürde farklı kaynaklarda çeşitli şekillerde ifade edilse de, genel olarak büyük veri yaşam döngüsü Şekil 3.3'teki aşamalardan oluşmaktadır.



Şekil 3.3 : Büyük veri yaşam döngüsü.

Büyük veri yaşam döngüsüne ait yaşam döngüsü takip eden alt başlıklarda detaylıca açıklanmıştır. Bu süreç akışı, büyük veri projelerinin yönetiminde ve analiz süreçlerinin planlanmasında yol gösterici bir çerçeve sunmaktadır.

3.1.8.1 Veri tanımlama

Bu aşama, büyük veri analizi için gerekli veri setlerinin belirlenmesi ve tanımlanması sürecini kapsamaktadır. Veri tanımlama süreci, analiz edilecek verilerin kapsamının, türünün ve kaynaklarının net bir şekilde ortaya konulmasını içermektedir. Ayrıca bu süreç, aynı varlığa ait farklı kaynaklarda bulunan bilgi parçalarının birbiriyle ilişkilendirilmesini içermektedir. Yapılandırılmış veriler söz konusu olduğunda tanımlama süreci genellikle basit ve doğrudan gerçekleştirilebilir. Ancak büyük veri kümelerinin çoğunlukla yapılandırılmamış veya yarı yapılandırılmış verilerden oluşması nedeniyle, veri tanımlama işlemi daha karmaşık bir hal almaktadır. Bu karmaşıklık, verilerin farklı formatlarda, farklı kaynaklardan ve farklı kalitede gelmesinden kaynaklanmaktadır.

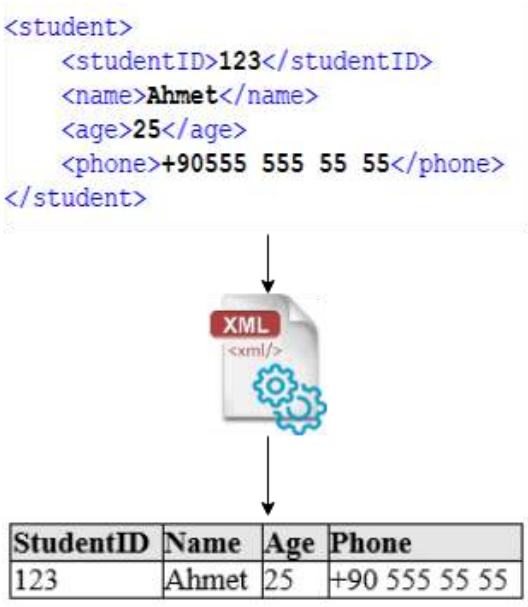
3.1.8.2 Veri toplama ve filtreleme

Bu aşama, önceden tanımlanan kaynaklardan verilerin toplanmasını ve ardından filtreleme işleminin gerçekleştirilmesini kapsamaktadır. Veri toplama sürecinde, belirlenen kaynaklardan elde edilen ham veriler, analiz süreçlerine aktarılmak üzere bir araya getirilmektedir. Toplanan veriler, analiz açısından değer taşımadığı düşünülen bozuk verilerden arındırılmak üzere filtreleme sürecinden geçirilir. Özellikle yapılandırılmamış veriler söz konusu olduğunda, elde edilen verilerin önemli bir kısmı ilgisiz, hatalı veya gürültülü olabilmektedir. Filtreleme süreci sonunda, bu tür veriler veri kümesinden çıkarılır. Ancak burada dikkat edilmesi gereken önemli bir nokta, bir analiz için değersiz kabul

edilerek filtrelenen verilerin, başka bir analiz için değerli olabileceğidir. Bu nedenle, veri yönetimi açısından iki önemli öneri bulunmaktadır: Filtreleme işlemine başlamadan önce orijinal veri kümesinin yedeklenmesi ve depolama alanından tasarruf sağlamak amacıyla verilerin sıkıştırılarak saklanmasıdır [39]. Gerçek zamanlı veri analizi söz konusu olduğunda ise süreç farklılık göstermektedir. Gerçek zamanlı analizlerde, veriler öncelikle analiz edilir ve ardından sıkıştırma ve yedekleme işlemleri gerçekleştirilir. Bu yaklaşım, analiz sonuçlarının mümkün olan en kısa sürede elde edilmesini sağlamakta ve gerçek zamanlı karar verme süreçlerini desteklemektedir.

3.1.8.3 Veri çıkarımı

Büyük veri analizi için toplanan verilerin bir bölümü, analiz için uygun olmayan formatlarda bulunabilir. Bu nedenle, ham veri kümelerinden analiz süreçlerinde kullanılabilecek anlamlı bilgilerin ayıklanması ve uygun formatlara dönüştürülmesi gerekmektedir. Veri çıkarımı ve dönüşüm süreçlerinin kapsamı, büyük veri analizinin hedeflerine göre şekillenmektedir. Bazı durumlarda, örneğin web günlüklerinde olduğu gibi belirli formatlara göre düzenlenmiş metinsel veriler, büyük veri analizinde doğrudan işlenebilir nitelikteyse, veri çıkarımı işlemine gerek duyulmayabilir [39]. Ancak, Şekil 3.4'te ifade edildiği gibi JSON, XML gibi yarı yapılandırılmış dokümanlarda, analiz için gerekli olmayan komut ifadeleri, etiketler ve sözdizimsel öğelerin ayıklanarak yalnızca anlamlı verilerin çıkarılması zorunludur. Bu süreç, verilerin analiz edilebilir hale getirilmesi açısından kritik öneme sahiptir.



Şekil 3.4 : Veri çıkarımı.

3.1.8.4 Veri doğrulama

Veri doğrulama ve temizleme aşaması, büyük veri yaşam döngüsünün kritik bir bileşenidir; çünkü geçersiz veya hatalı veriler, analiz sonuçlarının da geçersiz olmasına yol açabilmektedir. Bu aşamanın temel amacı, yalnızca uygun ve güvenilir verilerin analiz sürecine dahil edilmesini sağlamaktır. Veri kalitesi açısından problemli veriler genel olarak iki kategoriye ayrılmaktadır: beklenen formata uygun olmasına rağmen yanlış veya hatalı kabul edilen geçersiz veriler ve beklenen formata hiç uymayan bozuk veya tutarsız veriler. Bu tür problemli verileri analiz için uygun hale getirmek amacıyla ayrıştırma, standardizasyon, kısaltma, genişletme ve eksik bilgileri güncelleştirme gibi çeşitli yöntemler kullanılmaktadır. Veri doğrulama ve temizleme sürecinde, verilerin doğruluğunu ve tutarlılığını sağlamak için önceden belirlenmiş koşullar ve yöntemler sistematik olarak uygulanmaktadır. Ayrıştırma, verilerdeki kalıpları tanımlama ve analiz etme sürecidir. Bu yöntemde veriler, geçerli veya geçersiz verileri ayırt etmek amacıyla düzenli ifadeler veya önceden tanımlanmış desenlerle karşılaştırılmaktadır [56]. Standardizasyon yöntemi, ayrıştırma yöntemini bir adım ileriye götürerek verileri standart bir formata dönüştürmektedir. Bu yöntem ile verilerde bulunan kelimeler, kısaltmalara veya doğru yazım şekillerine dönüştürülebilmektedir. Örneğin, "Ankara" kelimesi veri kümesi içerisinde farklı biçimlerde temsil ediliyorsa, tüm bu farklı temsiller (örneğin "Ank.", "başkent" vb.) standart bir forma dönüştürülerek veri tutarlılığı sağlanabilir [56]. Kısaltma yöntemi ise, bilinen bir verinin daha kısa ve standart bir temsilini oluşturma işlemidir. Özellikle adres bilgilerinde sıkça kullanılan bu yönteme örnek olarak, "sokak" kelimesinin "sk.", "mahalle" kelimesinin "mah." şeklinde kısaltılması gösterilebilir. Bu yöntem, veri doğrulama ve temizleme aşamasında verilerin tutarlı ve standart hale getirilmesine yardımcı olmaktadır.

3.1.8.5 Veri birleştirme ve sunma

Veri analizi sürecinde, farklı kaynaklardan toplanan verilerin işlenebilmesi için ortak bir formata dönüştürülmesi ve birleştirilmesi gerekmektedir. Bu süreç, çeşitli veri kümelerinin tek bir bütünleşik yapıda toplanmasını sağlamaktadır. Veri birleştirme işlemi, farklı veri setlerine tekrar tekrar erişim ihtiyacını ortadan kaldırarak analiz sürecini önemli ölçüde hızlandırmakta ve verimliliği artırmaktadır.

3.1.8.6 Veri analizi

Günümüz dijital dünyasında verilerin sadece toplanılması değil aynı zamanda anlamlı bilgilere dönüştürülmesi de kritik önem taşımaktadır. Verilerin derinlemesine anlaşılması, stratejik karar alma süreçlerinde belirleyici rol oynamaktadır. Büyük veri analizi, karmaşık algoritmaların kullanılmasıyla veriler içindeki gizli örüntülerin, ilişkilerin ve değerli bilgilerin ortaya çıkarılması sürecini ifade etmektedir [57]. Büyük veri kümelerinden anlamlı sonuçlar elde edebilmek için gelişmiş analitik hizmetler, programlama araçları ve özel algoritmaların kullanılması gerekmektedir [58]. Bu analiz süreci, veri kümeleri arasındaki bilinmeyen bağlantıları keşfetmeye, önemli ilişkileri belirlemeye ve gizli örüntüleri ortaya çıkarmaya olanak tanımaktadır. Büyük veri kümelerinin ön işleme adımlarından geçirilmeden doğrudan analiz edilmesi, araştırmacılar için verimsiz ve zaman alıcı bir süreç oluşturur. Bu nedenle, verilerin hem hızlı hem de doğru yöntemlerle zamanında analiz edilmesi, etkili sonuçlar elde etmek için temel gereksinimdir.

3.1.8.7 Veri görselleştirme

Veri görselleştirme, karmaşık ve büyük hacimli bilgileri çeşitli grafik türleri ve görsel araçlar kullanarak daha anlaşılır ve etkili biçimde sunmayı amaçlamaktadır [59]. Bu yöntem, karar vericilerin verileri daha hızlı yorumlamasına ve görsel desteklerle daha isabetli kararlar almasına olanak sağlamaktadır. Ancak büyük verinin sahip olduğu yüksek hacim, çeşitlilik ve boyutsal karmaşıklık, görselleştirme sürecini zorlaştırmakta ve özel araçların kullanımını gerekli kılmaktadır. Büyük veri görselleştirme alanında önemli bir uygulama örneği olarak eBay gösterilebilir. Her ay milyonlarca aktif kullanıcı ve milyarlarca ürün satışıyla büyük miktarda veri üreten eBay, bu verileri anlamlandırmak ve analiz etmek için Tableau görselleştirme aracını kullanmaktadır. Tableau, karmaşık ve büyük ölçekli veri kümelerini sezgisel ve anlaşılır görsel unsurlara dönüştürme yeteneğine sahiptir. eBay çalışanları, Tableau sayesinde müşteri geri bildirimlerini takip edebilmekte, kullanıcıların arama davranışlarını analiz edebilmekte ve böylece müşteri deneyimini iyileştirmeye yönelik stratejik kararlar alabilmektedir [43].

3.1.9 Büyük veri analizinde kullanılan metotlar

Farklı kaynaklardan toplanan yapısal, yarı yapısal ve yapısal olmayan karmaşık veri kümeleri; hacim, hız ve çeşitlilik gibi özellikler açısından geleneksel veri kümelerinden farklılık göstermektedir. Geleneksel veri madenciliği yöntemleri genellikle sabit olarak depolanmış ve yapılandırılmış veriler üzerine odaklanırken, büyük veri analizinde

depolanmış verilerin yanı sıra sürekli olarak akan (streaming) ve dinamik yapıya sahip veriler de söz konusu olmaktadır. Bu tür büyük ölçekli ve dinamik veri kümelerini kısa süre içerisinde etkin ve verimli bir şekilde işleyip analiz edebilmek için gelişmiş ve güçlü veri analizi tekniklerine ihtiyaç duyulmaktadır. Bu nedenle büyük veri analiz süreçlerinde, geleneksel veri madenciliği yöntemlerinin yanı sıra makine öğrenmesi, derin öğrenme, doğal dil işleme ve istatistiksel analiz gibi farklı disiplinlerden ve alanlardan da yararlanılmaktadır [60].

3.1.9.1 Veri madenciliği metotları

Veri madenciliği, büyük veri kümeleri içerisindeki gizli örüntüleri ve anlamlı bilgileri ortaya çıkarmak amacıyla kullanılan; ilişkilendirme kuralları, kümeleme, sınıflandırma ve regresyon analizi gibi çeşitli yöntemleri içeren bir süreçtir [43]. Bu süreçte, veri kümelerinden yararlı bilgilerin çıkarılması için makine öğrenmesi algoritmaları ve istatistiksel yöntemlerden yararlanılmaktadır. Son yıllarda veri hacminin hızla artması ve verilerin karmaşıklığının yükselmesi nedeniyle, geleneksel veri madenciliği yöntemleri yetersiz kalmakta ve yeni nesil büyük veri madenciliği tekniklerine ihtiyaç duyulmaktadır. Bu kapsamda, özellikle derin öğrenme tabanlı yöntemler büyük veri analizinde yaygın olarak kullanılmakta ve başarılı sonuçlar vermektedir.

3.1.9.2 Görselleştirme metotları

Görselleştirme, büyük veri kümelerinin genel bir görünümünün elde edilmesini ve verilerdeki anlamlı bilgilerin keşfedilmesini sağlayan önemli bir araçtır. Geleneksel veri görselleştirme yöntemlerinin büyük veri uygulamalarında etkin bir şekilde kullanılabilmesi için geliştirilmesi ve uyarlanması gerekmektedir. Ancak büyük hacimli verilerin görselleştirilmesi, verilerin yüksek boyutluluğu ve karmaşıklığı nedeniyle geleneksel veri kümelerinin görselleştirilmesine göre daha zordur. Bu nedenle, büyük veri görselleştirme süreçlerinde, yüksek boyutlu verileri etkin biçimde temsil edebilen gelişmiş görselleştirme teknikleri ve araçları kullanılmaktadır.

3.1.9.3 Makine öğrenmesi metotları

Makine öğrenmesi, yapay zeka alanının bir alt dalı olup, bilgisayarların verilerden öğrendikleri davranışları geliştirerek farklı problemlere uygulamalarına dayanmaktadır [43]. Belirli bir veri kümesindeki anlamlı desenlerin otomatik olarak tespit edilmesi yeteneği, makine öğrenmesini büyük veri analizi açısından kritik öneme sahip hale getirmektedir.

Ancak mevcut makine öğrenmesi yöntemlerinin büyük veri kümelerinde etkin bir şekilde kullanılabilmesi için, diğer yöntemlerde olduğu gibi bu yöntemlerin de geliştirilmesi ve uyarlanması gerekmektedir. Büyük veri kümeleriyle çalışmak üzere makine öğrenmesi algoritmaları modifiye edilerek çeşitli yöntemler geliştirilmiştir [61]. Bu yöntemlere örnek olarak, büyük ölçekli verilerin paralel ve dağıtık olarak işlenmesini sağlayan MapReduce [62] gibi veri işleme teknolojileri ve Hadoop [62] gibi dağıtık işlem mimarileri bunlara örnek olarak verilebilir.

Yapay sinir ağları, uygulama alanı oldukça geniş olan bir makine öğrenmesi tekniğidir. Örüntü tanıma, görüntü analizi, adaptif kontrol ve daha birçok alanda başarılı bir şekilde uygulanmaktadır. Özellikle derin öğrenme yöntemleriyle birlikte kullanıldığında, konuşma tanıma, görsel nesne sınıflandırma, bilgi çıkarımı ve doğal dil işleme gibi alanlarda büyük hacimli ve yapılandırılmamış verilerin analizinde oldukça etkili sonuçlar vermektedir.

3.1.9.4 Sosyal ağ analizi metotları

Sosyal ağ analizi, sosyal sistem tasarımı, insan davranış modellemesi [63], sosyal ağ görselleştirme [64], çizge sorgulama ve çizge madenciliği gibi çeşitli yöntemleri içermektedir [65]. Son yıllarda sosyal medya platformlarının yaygınlaşmasıyla birlikte, sosyal medya analizi de önemli bir araştırma alanı haline gelmiştir. Sosyal medya platformları aracılığıyla büyük hacimli ve yüksek hızda veri üretilmekte, bu durum ise verilerin analiz edilmesini zorlaştırmaktadır. Ayrıca, milyarlarca nesnenin birbirine bağlı olduğu sosyal ağların analiz edilmesi, yüksek hesaplama maliyetleri ve karmaşık analiz süreçleri gerektirmektedir. Ancak bu maliyetlere rağmen, sosyal ağların analiz edilmesi pazarlamacılar ve kurumlar için önemli faydalar sağlamaktadır. Sosyal ağ analizi sayesinde, bir topluluk veya kurum içerisindeki bireyler arasındaki bağlantılar ortaya çıkarılmakta, bilginin ağ içerisinde nasıl yayıldığı ve hangi kullanıcı grupları üzerinde daha fazla etkiye sahip olduğu belirlenebilmektedir. Bu bilgiler, kurumların pazarlama stratejilerini geliştirmelerine ve hedef kitlelerine daha etkin bir şekilde ulaşmalarına yardımcı olmaktadır.

3.2 Siber Tehdit İstihbaratı

Siber tehdit istihbaratı, kurumsal siber güvenlik ekosisteminin kritik bir bileşenidir. CTI, kuruluşların karşı karşıya kalabileceği potansiyel siber tehditlerin sistematik bir şekilde tanımlanması, analiz edilmesi ve değerlendirilmesi süreçlerini kapsamaktadır [66]. Bu süreç, tehdit aktörlerinin davranış kalıplarının, taktiklerinin ve prosedürlerinin derinlemesine

incelenmesini içermektedir. CTI'nin temel amacı, kuruluşların siber tehditlere karşı önleyici bir savunma stratejisi geliştirmelerine yardımcı olmaktır. Siber tehdit istihbaratı süreci, analistler, karar vericiler ve tehdit ortamı arasındaki dinamik bir etkileşimi ifade etmektedir. Analistler, tehdit ortamını detaylı bir şekilde inceleyerek elde ettikleri bilgileri yorumlamakta ve karar vericilere sunmaktadır. Karar vericiler ise bu bilgiler doğrultusunda siber tehditleri kontrol altına almak veya riskleri azaltmak için stratejik kararlar almaktadırlar. Aynı zamanda karar vericiler, analistlerin hangi konulara odaklanması gerektiğini belirleyerek bu süreci şekillendirmektedirler. Bu şekilde, siber tehdit bilgileri döngüsel olarak sürekli analiz edilerek ve yönetilerek daha güvenli bir ortam oluşturulmaktadır.

Siber tehdit istihbaratı, yapılandırılmış ve yapılandırılmamış veri kaynaklarından elde edilen bilgilerin sistemli bir şekilde toplanması, analizi ve sentezlenmesi süreçlerini kapsamaktadır. Bu yaklaşımla, tehdit aktörlerinin davranışlarının, yeteneklerinin ve hedeflerinin hem teknik hem de operasyonel olarak kapsamlı bir şekilde anlaşılması sağlanmaktadır. Siber tehdit istihbaratı, modern siber güvenliğin temel yapı taşlarından biridir ve birbiriyle bütünleşik üç ana süreçten oluşmaktadır. Birinci süreçte, güvenlik olaylarına ilişkin sistemli olarak veri toplama ve analiz faaliyetleri gerçekleştirilmektedir. Bu süreçte, çeşitli kaynaklardan elde edilen ham veriler, gelişmiş analitik yöntemler kullanılarak işlenir ve anlamlı örüntüler tespit edilmeye çalışılır. İkinci aşamada, toplanan verilerin birleştirilip birbirine entegre edilmesi ve potansiyel tehditlere ilişkin anlamlı istihbarat bilgisine dönüştürülmesi gerçekleştirilir. Bu dönüşüm süreci, tehdit aktörlerinin davranış örüntülerini, saldırı vektörlerini ve olası hedeflerini belirlemeye yönelik derinlemesine bir analizi kapsamaktadır. Elde edilen anlamlı bilgiler, kuruluşların kendi risklerini hesaplamada kritik bir rol oynamaktadır. Üçüncü aşamada ise, üretilen istihbarat bilgisinin kuruluş içindeki ilgili birimlere etkili bir şekilde aktarılması ve mevcut savunma mekanizmalarının bu doğrultuda inşa edilmesinin sağlanmasıdır. Bu savunma mekanizmaları, ilgili kuruluşun siber tehditlere karşı direncini artırmakta ve güvenlik açıklarının minimize edilmesine katkıda bulunmaktadır. Siber tehdit istihbaratının temel amacı, kuruluşların siber güvenliğe yaklaşımlarında değişiklik sağlamaktır. Bu değişim, saldırı gerçekleştikten sonra tepki vermeye dayalı olan geleneksel güvenlik anlayışından, tehditleri öngörüp önceden önlem almaya dayalı modern güvenlik yaklaşımına geçişi amaçlamaktadır. Belirtilen bu dönüşüm, üç temel bileşenden oluşmaktadır. İlk olarak, potansiyel siber tehditler için erken siber tespit mekanizmalarının geliştirilmesidir. İkinci

olarak, tespit edilen siber tehditlere yönelik risk azaltma yaklaşımlarının uygun zamanda devreye alınmasının sağlanması ve son olarak da mevcut güvenlik açıklarının sistemli ve etkin bir şekilde yönetilmesi süreçlerini içermektedir. Belirtilen bu dönüşümün nihai amacı, kuruluşların, organizasyonların vb. siber tehditlere karşı dirençlerini artırmaktır. Bu sayede kuruşlar, dijital çağın karmaşık tehdit ortamına daha hazırlıklı ve dayanıklı bir yapıya kavuşmaktadır. Bu yapı, kuruluşların siber güvenlik farkındalık seviyelerinin artmasına ve kritik altyapıların daha etkin bir şekilde korunmasına olanak sağlamaktadır.

3.2.1 Siber tehdit istihbaratının türleri

Siber tehdit istihbaratı, kuruluşların farklı seviyelerindeki ihtiyaçlarına cevap verebilmek için dört temel kategoride sınıflandırılmaktadır. Her kategori, belirli bir kullanıcı grubunun ihtiyaçlarını karşılamak üzere tasarlanmıştır. Siber tehdit istihbaratı dört farklı grupta incelenmektedir [67].

3.2.1.1 Stratejik tehdit istihbaratı

Kuruluşların üst düzey karar alma mekanizmaları için destekleyici görev gören ve en kapsamlı istihbarat türüdür. Bu istihbarat türü, kuruluşun gelecekteki güvenlik kararlarını şekillendirmede önemli bir rol oynamaktadır [68]. Örneğin, bir bankadaki üst düzey yöneticiler, finans sektörüne yönelik gerçekleştirilen siber saldırıların detaylı analizlerini içeren raporları inceledikten sonra saldırganların kullandığı yöntemleri, hedef aldıkları sistemleri ve saldırıların banka altyapıları ile ilgili detaylı bilgiler alabilir. Banka yönetimi bu bilgilerle beraber, mobil bankacılık uygulamalarının veya müşteri verilerinin korunması için yeni güvenlik önlemleri alarak istihbarata dayalı olarak savunma stratejilerini bu şekilde belirleyebilmektedirler. Stratejik tehdit istihbaratı aynı zamanda kuruluşun karşı karşıya olduğu risklerin belirlenmesinde ve uzun vadeli güvenlik planlarının yapılmasında da temel bir kaynak görevi görmektedir. Bu sayede, gelecekteki olası tehditlere karşı daha hazırlıklı olmakta ve kaynaklarını bu yönde daha verimli kullanabilmektedirler. Örneğin, yeni sistemler geliştirilirken hangi güvenlik önlemlerinin alınması gerektiği konusunda önceki istihbaratlara dayanarak daha stratejik kararlar alınabilmektedir.

3.2.1.2 Operasyonel tehdit istihbaratı

Günlük güvenlik olaylarının yönetiminde kullanılan pratik bilgileri kapsayan istihbarat türüdür. Güvenlik Operasyon Merkezleri (Security Operations Center - SOC) için hayati önem taşıyan bu istihbarat türü devam eden veya yaklaşan siber tehditlerin zamanı,

hedefi ve kullandıkları yöntem gibi kritik bilgileri içermektedir [69]. Örneğin, devam eden bir dağıtık hizmet engelleme (Distributed Denial of Service - DDoS) saldırısı hakkında anlık bildirimler ve saldırı detayları bu kategori altında değerlendirilmtekdir.

3.2.1.3 Teknik tehdit istihbaratı

Siber güvenlik savunmasının en ayrıntılı seviyesini oluşturmaktadır. Bu istihbarat türünde, tehdit aktörlerinin kullandığı teknik altyapılara ait somut bilgiler yer almaktadır. IP adresleri, dosya hash değerleri, kötü amaçlı yazılım örnekleri ve zararlı alan adları gibi teknik detaylar bu kategoride incelenmtekdir [69]. Örneğin, bir fidye yazılımı saldırısında kullanılan zararlı dosyaların hash değerleri, güvenlik sistemlerinin bu tehditleri otomatik olarak tespit etmesini ve engellemesini sağlamaktadır.

3.2.1.4 Taktiksel tehdit istihbaratı

Tehdit aktörlerinin kullandığı yöntemlerin detaylı analizine odaklanmaktadır. Bu istihbarat türü, özellikle güvenlik ekiplerinin savunma stratejilerini geliştirmelerine yardımcı olmaktadır. Tehdit aktörlerinin kullandığı taktikler, teknikler ve prosedürler hakkında derinlemesine bilgi sağlamaktadır [69]. Örneğin, yeni ortaya çıkan bir zararlı yazılımın kullandığı saldırı vektörlerinin analizi, güvenlik ekiplerinin etkili savunma mekanizmaları geliştirmesine olanak tanımaktadır.

Bahsedilen bu dört istihbarat türü, birbirleriyle iletişim halinde çalışmaktadırlar. Her bir istihbarat türü, kendi içinde belirli bir görevi ihya ederken, bir bütün olarak bu dört istihbarat türü kuruluşların siber tehditlere karşı kapsamlı bir şekilde savunma stratejisi geliştirmelerine yardımcı olmaktadır.

3.2.2 Siber tehdit istihbaratı standartları ve paylaşım protokolleri

Siber tehditlere karşı etkin ve hızlı bir şekilde karşılık verebilmek adına süreçlerin otomatikleştirilmesi ve anlaşılır ortak bir çerçevenin oluşturulması kritik bir rol oynamaktadır. Bu kapsamda, küresel olarak siber güvenlik birimleri arasındaki iş birliğinin sağlanması için standartlaştırma girişimleri hayata geçirilmiştir. Bu standartlardan bazıları olay verileriyle ilgili genel bilgiler sunarken, bazıları ise tehdit istihbaratlarını makine tarafından okunabilir bir biçimde sunarak farklı sistemler arasında kolayca paylaşılmasını mümkün kılmaktadır. Bu standartlar, verilerin nasıl yapılandırılacağını belirlerken, aynı zamanda da bu bilgilerin nasıl iletileceğini belirlemek için protokol geliştirilmesini de gerekli kılmaktadır. Yapılandırılmış tehdit bilgi ifadesi [4] (STIX) standardı ve güvenilir

otomatik istihbarat paylaşımı [70] (TAXII) protokolü, açık tehdit verilerinin yapılandırılmasını ve paylaşılmasını kolaylaştırmak amacıyla 2012 yılında geliştirilmiştir. Bu çalışmaları yöneten MITRE, tehditlerle ilgili veri paylaşımını standartlaştırmak amacıyla çeşitli standartlar ve yöntemler geliştirmiştir. Bu standartlardan bazıları, ortak platform numaralandırması [71] (Common Platform Enumeration - CPE), ortak saldırı deseni sınıflandırması [72] (Common Attack Pattern Enumeration and Classification - CAPEC) olarak sıralanabilir. Tüm bu girişimlerin temel hedefi, siber güvenliğin daha ölçülebilir ve yönetilebilir hale getirilmek istenmesidir. Ayrıca bu çalışmalar, tehditlerin daha verimli bir şekilde analiz edilmesini ve paylaşılmasını sağlamak için önemli bir altyapı oluşturmaktadır. Günümüzde bu standartlar ve protokoller, bazı araç ve eklentiler ile desteklenmektedir. Bu tür ek araç ve eklentiler bahsedilen bu standart ve protokollerin kullanımını kolaylaştırmaktadır.

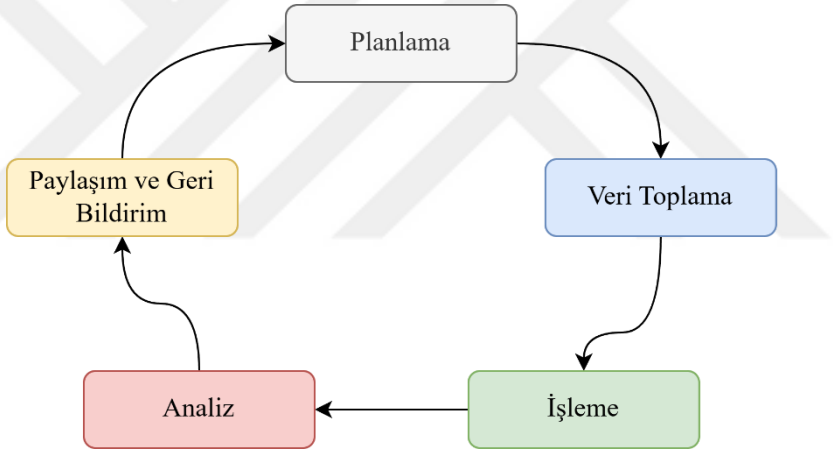
3.2.3 Siber tehdit istihbaratı platformları ve servisleri

Siber tehdit istihbaratı standartları ve protokolleri, yalnızca araç ve eklentilerle değil, aynı zamanda özel olarak tasarlanmış platformlar ve servislerle de desteklenmektedir. Bu platformlar ve servisler, tehdit istihbaratını toplama, analiz etme ve paylaşma süreçlerini kolaylaştırarak güvenlik operasyonlarının etkinliğini artırmaktadır. Örneğin, ThreatConnect [73], tehdit istihbaratı verilerini toplama ve analiz etme için kapsamılı bir altyapı sunmaktadır. Platform, tehdit aktörleri, siber saldırı teknikleri ve zafiyetlerle ilgili bilgileri derinlemesine analiz ederek güvenlik ekiplerine kritik destek sağlamaktadır. Aynı zamanda, manuel analiz ihtiyacını azaltan otomasyon yetenekleriyle dikkat çekmektedir. Benzer şekilde, Recorded Future [74], tehdit istihbaratı verilerini gerçek zamanlı olarak sağlamakta ve tehdit aktörlerinin davranışlarını analiz ederek güvenlik ekiplerine karar alma süreçlerinde önemli bir avantaj sunmaktadır. Bu platform, özellikle tehditlerin önceden tespit edilmesi ve öngörülmesi için kullanılmaktadır. Servisler tarafında, IBM X-Force Exchange [75], tehdit verilerini analiz etmek ve bu bilgileri diğer paydaşlarla paylaşmak için kullanılan bulut tabanlı bir platformdur. Bu servis, topluluk tabanlı bilgi paylaşımına olanak tanırken, zenginleştirilmiş veri setleriyle tehditlere karşı etkili çözümler sunmaktadır. Bir diğer önemli servis olan Cisco Talos [76], küresel tehdit ortamını sürekli izleyerek kullanıcılara erken uyarılar sağlamaktadır. Bu servis, potansiyel tehditleri öngörerek güvenlik ekiplerinin daha önceden hazırlıklı olmasını sağlamaktadır. Kaspersky Threat Intelligence Services [77] ise, tehdit istihbaratını derinlemesine analiz etmekte ve özelleştirilmiş raporlarla kurumsal kullanıcıları desteklemektedir. Bu servis, özellikle

hedefli saldırılara karşı önlem alınmasında kritik bir rol oynamaktadır. Son olarak, CrowdStrike Falcon X [78], tehdit istihbaratını otomatikleştirilmiş bir şekilde analiz etmekte ve saldırganların taktik, teknik ve prosedürlerini detaylı bir şekilde ortaya çıkarmaktadır. Bu, olay müdahale ekipleri için çok değerli bir bilgi kaynağı sunmaktadır. Bu platform ve servisler, tehdit istihbaratını operasyonel süreçlere entegre ederek hem standartların hem de protokollerin etkinliğini artırmaktadır. Ayrıca, otomasyon ve yapay zeka gibi teknolojilerle desteklenen bu sistemler, tehditlerin erken tespit edilmesi ve hızlı müdahale edilmesi için bir altyapı sunmaktadır.

3.2.4 Siber tehdit istihbaratı yaşam döngüsü

Siber tehdit istihbaratı, genellikle planlama, veri toplama, işleme, analiz, paylaşım ve geri bildirim süreçlerini içeren bir yaşam döngüsü ile uygulanmaktadır. Şekil 3.5'te gösterilen bu döngü aşağıdaki aşamalardan oluşmaktadır.



Şekil 3.5 : Siber tehdit istihbaratı yaşam döngüsü.

Planlama: İhtiyaç duyulan tehdit istihbaratının türü ve kapsamının belirlendiği aşamadır. Hedefler, ilgili paydaşların gereksinimlerine göre oluşturulmaktadır.

Veri toplama: Bu aşamada tehditlere yönelik veriler, açık kaynak istihbarat kaynaklarından, tehdit istihbaratı platformlarından, siber güvenlik araçlarından ve özel kaynaklardan toplanmaktadır.

İşleme: Siber tehdit istihbarat kaynaklarından toplanan veriler, analiz edilebilir bir formata dönüştürüldüğü aşama işleme aşamasıdır. Bu aşamada gereksiz ve ilgisiz veriler elenmekte ve yapılandırılmış bir forma getirilmektedir.

Analiz: Tehdit verilerinin anlamlı bilgilere dönüştürülerek tehditlerin tespiti, değerlendirilmesi ve sınıflandırılması gibi işlemlerin yapıldığı aşamadır.

Paylaşım ve geri bildirim: Analiz sonucunda elde edilen tehdit istihbaratı bilgilerinin, ilgili paydaşlara zamanında ve doğru bir formatta iletildiği aşamadır. Tehdit istihbaratı genellikle, TAXII ve STIX gibi standartlar kullanılarak paylaşılmaktadır. Güvenlik sistemleri bu istihbaratlar doğrultusunda güncellenerek güvenlik riskleri azaltılmaktadır.

3.2.5 Tehdit istihbaratının uygulama alanları

Siber tehdit istihbaratı, kuruluşların güvenlik açıklarını önleyici bir şekilde tespit edip ele almasına yardımcı olan çeşitli kullanım alanlarına sahiptir. Bu kullanım alanları, farklı güvenlik operasyonlarından stratejik risk analizlerine kadar geniş bir yelpazeye yayılmakta ve güvenlik ekiplerine kritik bilgiler sunarak tehditlere karşı daha etkili bir savunma sağlamaktadır.

3.2.5.1 Güvenlik operasyonları için tehdit istihbaratı

Güvenlik operasyonları, tehdit istihbaratı ile güçlendirilerek olayların daha hızlı tespit edilmesini ve yanıtlanmasını sağlamaktadır. İstihbarat verileri, güvenlik bilgisi ve olay yönetimi (Security Information and Event Management - SIEM) sistemlerine entegre edilerek anomali algılama ve tehdit önleme süreçlerini daha etkin hale getirmektedir [79].

3.2.5.2 Olay müdahalesi için tehdit istihbaratı

Olay müdahale ekipleri, tehdit istihbaratı sayesinde olayların kaynağını ve etkisini daha hızlı analiz edebilmektedir. Bu bilgiler, olayın etkisini sınırlamak ve saldırganların gelecekteki girişimlerini önlemek için kritik kararların alınmasına olanak tanımaktadır [79].

3.2.5.3 Zafiyet yönetimi için tehdit istihbaratı

Tehdit istihbaratı, zafiyet yönetimi süreçlerini destekleyerek kuruluşların güvenlik açıklarını önceliklendirmesine yardımcı olmaktadır. Bu sayede, kuruluşlar sınırlı kaynaklarını en kritik açıkların kapatılmasına harcayarak risklerini azaltabilmektedir. Dinamik ve statik risk analizleri, zafiyetlerin gerçek zamanlı olarak izlenmesini sağlamakta ve bu da güvenlik stratejilerinin daha etkili bir şekilde uygulanmasına katkı sunmaktadır [80].

3.2.5.4 Güvenlik liderleri için tehdit istihbaratı

Siber güvenlik liderleri, tehdit istihbaratını stratejik kararlar almak için kullanır. Bu bilgiler, güvenlik yatırımlarının hangi alanlara yönlendirilmesi gerektiği ve güvenlik stratejilerinin nasıl şekillendirileceği konusunda rehberlik sağlamaktadır [79].

3.2.5.5 Risk analizi için tehdit istihbaratı

Tehdit istihbaratı, risk analizi süreçlerinin daha doğru ve etkin bir şekilde gerçekleştirilmesine olanak tanımaktadır. Tehdit istihbaratı, kuruluşlara gerçek zamanlı veriler sağlayarak potansiyel tehditlerin tespit edilmesine ve bu tehditlerin iş süreçleri üzerindeki etkisini değerlendirmesine olanak sağlamaktadır. Bu yaklaşım, risklerin önceliklendirilmesi ve kritik varlıkların korunmasına yönelik stratejik kararların alınmasında önemli bir rol oynamaktadır [81].

3.2.5.6 Dolandırıcılığı önleme için tehdit istihbaratı

Tehdit istihbaratı, dolandırıcılık girişimlerinin önceden tespit edilmesini sağlayacak bilgiler sağlayarak, finans kuruluşlarının bu tehditlere karşı önlemler almasını sağlamaktadır. Gerçek zamanlı veriler ve geçmiş bilgiler, dolandırıcılık tespit algoritmalarının güncellenmesini destekleyerek kaynakların daha verimli kullanılmasını ve risklerin azaltılmasını sağlamaktadır. Bu yaklaşım, dolandırıcılık önleme stratejilerinin etkinliğini artırmaktadır [82].

3.2.6 Siber tehdit istihbaratının çalışmadaki rolü

Siber tehdit istihbaratı, bu tez çalışmasının temel bileşenlerinden birini oluşturmaktadır. Çalışma kapsamında, siber tehdit istihbaratı hem verilerin toplanmasında hem de bu verilerin analiz edilip anlamlandırılmasında yönlendirici bir çerçeve sağlamaktadır. Bu bağlamda, siber tehdit istihbaratı verileri ile ilgili süreçler, siber tehdit istihbaratı ham verilerinin eyleme dönüştürülebilir bilgilere dönüştürülmesini mümkün kılarak çalışmanın genel metodolojisine önemli bir katkı sunmuştur. Çalışmada kullanılan siber tehdit istihbarat verileri, tehdit aktörleri, zafiyetler, zararlı yazılımlar, saldırı teknikleri vb. geniş bir yelpazeyi kapsamaktadır. Toplanan ham veriler, siber tehdit istihbaratı yaşam döngüsü doğrultusunda işlenmiş ve analiz edilebilir bir forma dönüştürülmüştür. Bu süreç, verilerin sistematik bir şekilde düzenlenmesi ve kullanılabilir hale getirilmesini sağlamıştır. Verilerin işlenmesinin ardından, tehdit raporları ve ilgili belgeler üzerinde adlandırılmış varlık tanıma süreçleri uygulanmıştır. Bu süreç, siber tehdit varlıklarının metinlerden

otomatik olarak çıkarılmasını sağlamıştır. Adlandırılmış varlık tanıma, tehdit unsurlarının doğru bir şekilde belirlenmesini sağlayarak, sonraki analiz süreçleri için sağlam bir temel oluşturmuştur. Bu süreçte elde edilen sonuçlar, siber tehditlerin daha sistematik bir şekilde değerlendirilmesine olanak tanımıştır.

Siber tehdit istihbaratı, çalışmada çizge tabanlı analizlere temel veri sağlamada da önemli bir rol oynamıştır. Toplanan ve işlenen veriler, çizge tabanlı veri modellerinde düğümler ve kenarlar şeklinde temsil edilmiştir. Bu çizge modeller, tehdit aktörleri, zararlı yazılımlar ve hedeflenen zafiyetler arasındaki ilişkilerin detaylı bir şekilde incelenmesini mümkün kılmıştır. Çizge analizleri sayesinde, tehdit aktörleri arasındaki bağlantılar görselleştirilmiş ve bu ilişkiler derinlemesine analiz edilmiştir. Siber varlıklar arasındaki ontoloji tabanlı ilişkiler ise, siber tehdit istihbaratını çalışmanın daha anlamlı bir bağlamda ele almasına katkı sağlamıştır. Çalışmada, siber tehdit verileri ontoloji tabanlı bir yapıya dönüştürülerek varlıklar arasındaki ilişkiler tanımlanmıştır. Bu ilişkiler, çizge veri modellerine entegre edilerek tehdit aktörleri ve tehdit unsurları arasındaki etkileşimler daha net bir şekilde ortaya konmuştur. Bu bağlamda, siber tehdit istihbaratı bu çalışmada bir temel sunarak tüm süreçleri yapılandırmış ve yönlendirmiştir. Verilerin toplanması, işlenmesi ve çizge analizlerinde modellenmesi aşamalarında, siber tehdit istihbaratı ilkeleri temel alınmıştır. Bu yaklaşım, çalışmanın sistematik bir şekilde ilerlemesini ve elde edilen sonuçların güvenilirliğini sağlamıştır.

3.3 Doğal Dil İşleme ve Adlandırılmış Varlık Tanıma

Doğal Dil İşleme, insan dilini bilgisayarlar aracılığıyla işlemek, anlamak ve analiz etmek için kullanılan yapay zekanın bir alt alanıdır. Bu alan, dilin yapısal karmaşıklığını çözümlemek için dilbilim, istatistik ve derin öğrenme tekniklerinden yararlanmaktadır. Doğal dil işleme, metin sınıflandırma, duygu analizi, bilgi çıkarımı, otomatik çeviri ve konuşma tanıma gibi görevlerde geniş bir kullanım alanına sahiptir [83]. Özellikle siber güvenlikte doğal dil işleme, tehdit istihbaratı oluşturmak ve anormal davranışları tespit etmek gibi görevlerde kritik bir rol oynamaktadır. Örneğin, kötü amaçlı yazılım tespitinde [84] veya sosyal mühendislik saldırılarının tespitinde [85], doğal dil işleme algoritmaları tehdit aktörlerinin dil kalıplarını analiz edebilmektedir.

Doğal dil işlemenin bir alt dalı olan adlandırılmış varlık tanıma, doğal dil işlemenin birçok görevine temel oluşturmaktadır. Adlandırılmış varlık tanıma, bir metindeki kişi, yer, organizasyon gibi özel isimleri tespit ederek bu varlıkları belirli kategorilere sınıflandırma

görevidir. Örneğin, bir müşteri şikayetini analiz eden bir sistem, adlandırılmış varlık tanıma kullanarak şikayetin hangi ürün ya da hizmetle ilgili olduğunu otomatik olarak belirleyebilmektedir. Doğal dil işlemenin genel yapısında, adlandırılmış varlık tanıma genellikle bilgi çıkarımı süreçlerinin bir parçası olarak işlev görmekte ve metinlerin anlamlandırılmasına katkıda bulunmaktadır. Bu nedenle, adlandırılmış varlık tanıma, doğal dil işlemenin birçok uygulaması için kritik bir adımdır ve özellikle siber güvenlik alanında tehdit istihbaratı elde etmek için kullanılmaktadır [86].

3.3.1 Doğal dil işleme ve temel kavramlar

Doğal dil işleme, insan dilinin karmaşık yapısını anlama ve işleme yeteneği kazandırmayı amaçlayan disiplinlerarası bir araştırma alanıdır. İnsan dilinin doğal yapısındaki belirsizlikler ve bağlamsal değişkenlik, bu alanı yapay zeka araştırmalarının en zorlu konularından biri haline getirmektedir. Dilbilim ve bilgisayar bilimi yaklaşımlarının kesişiminde yer alan doğal dil işleme, son yıllarda derin öğrenme tekniklerinin gelişimiyle beraber önemli ilerlemeler kaydetmiştir.

3.3.2 Adlandırılmış varlık tanıma

Adlandırılmış varlık tanıma, doğal dil işleme alanında metin içerisindeki özel isimleri, tarihleri, yer isimlerini, organizasyonları ve benzeri varlıkları otomatik olarak tanımlama ve sınıflandırma işlemidir. NER, metin madenciliği, bilgi çıkarımı, soru-cevap sistemleri ve bilgi erişimi gibi birçok uygulama alanında kritik bir rol oynamaktadır [87]. NER sistemleri, genellikle önceden belirlenmiş kategorilere göre varlıkları sınıflandırmaktadır. Bu kategoriler arasında kişi isimleri (PER), yer isimleri (LOC), organizasyon isimleri (ORG), tarihler (DATE), zaman ifadeleri (TIME), parasal değerler (MONEY) ve yüzdelik ifadeler (PERCENT) gibi türler bulunmaktadır [88]. Siber güvenlik alanında ise IP adresleri, alan adları, zararlı yazılım isimleri, saldırı türleri ve tehdit aktörleri gibi özel varlık türleri de sıklıkla kullanılmaktadır [89].

NER yöntemleri genel olarak üç temel yaklaşıma ayrılmaktadır: kural tabanlı yöntemler, makine öğrenmesi tabanlı yöntemler ve derin öğrenme tabanlı yöntemlerdir [90]. Kural tabanlı yöntemler, dilbilimsel kurallar ve sözlükler kullanılarak varlıkları tanımlarken, makine öğrenmesi yöntemleri, etiketlenmiş veri kümelerinden öğrenilen modeller aracılığıyla varlıkları tanımlamaktadır. Son yıllarda derin öğrenme yöntemleri, özellikle tekrarlayan sinir ağları (Recurrent Neural Networks - RNN) ve Transformer tabanlı modeller (örneğin BERT, RoBERTa) ile yüksek başarımlar elde etmektedir [91]. Makine öğrenmesi

tabanlı NER yöntemleri arasında en yaygın kullanılan algoritmalardan biri koşullu rastgele alanlar (Conditional Random Fields - CRF) algoritmasıdır. CRF, metin içerisindeki kelimelerin bağlamını dikkate alarak varlıkları sınıflandırmakta ve özellikle sıralı etiketleme problemlerinde başarılı sonuçlar vermektedir [92]. Ancak CRF gibi geleneksel yöntemler, büyük veri kümelerinde ve karmaşık dil yapılarında sınırlı kalabilmektedir.

Son yıllarda derin öğrenme yöntemlerinin NER alanında kullanımı yaygınlaşmıştır. Özellikle transformer tabanlı modeller, bağlam bilgisini daha iyi yakalayarak varlık tanıma performansını önemli ölçüde artırmıştır. Devlin ve arkadaşları [91] tarafından geliştirilen BERT modeli, ön eğitimli dil modelleri arasında en popüler olanlardan biridir ve birçok NER görevinde en iyi sonuçları vermektedir. BERT, büyük ölçekli metin verileri üzerinde ön eğitim yaparak, bağlam bilgisini etkin bir şekilde kullanabilmekte ve sınırlı etiketli veri ile yüksek doğruluk oranları elde edebilmektedir.

Siber güvenlik alanında NER uygulamaları, tehdit istihbaratı raporları, güvenlik günlükleri (loglar), sosyal medya paylaşımları ve forumlar gibi çeşitli kaynaklardan kritik varlıkların otomatik olarak çıkarılmasını sağlamaktadır. Örneğin, Bridges ve arkadaşları [89], siber güvenlik metinlerinden zararlı yazılım isimleri, saldırı türleri ve tehdit aktörleri gibi varlıkları otomatik olarak tanımlamak için makine öğrenmesi tabanlı NER yöntemlerini kullanmışlardır. Bu çalışma, siber tehdit istihbaratı alanında NER yöntemlerinin etkinliğini ve önemini ortaya koymaktadır.

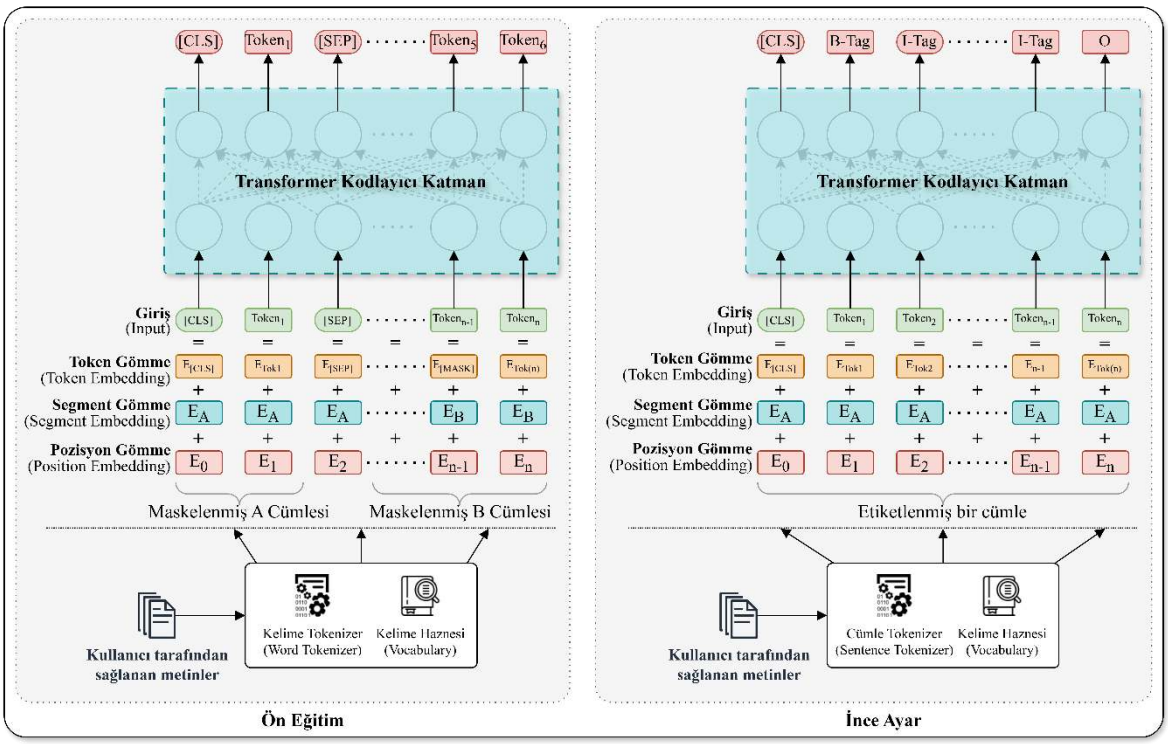
NER yöntemlerinin başarımı, kullanılan veri kümesinin büyüklüğüne, kalitesine ve dil özelliklerine bağlı olarak değişiklik göstermektedir. Özellikle siber güvenlik alanında kullanılan metinlerin teknik terimler içermesi, standart dışı dil kullanımı ve kısaltmaların yoğunluğu gibi faktörler, varlık tanıma sürecini zorlaştırmaktadır [93]. Bu nedenle, siber güvenlik alanına özgü veri kümelerinin oluşturulması ve bu veri kümeleri üzerinde eğitilmiş modellerin kullanılması, NER yöntemlerinin başarımını artırmak için kritik öneme sahiptir.

Sonuç olarak, adlandırılmış varlık tanıma yöntemleri, metinlerden anlamlı bilgilerin otomatik olarak çıkarılması için temel bir araçtır. Özellikle siber güvenlik alanında, tehdit istihbaratı ve güvenlik analizleri için kritik öneme sahip varlıkların otomatik olarak tanımlanması ve sınıflandırılması, güvenlik uzmanlarının iş yükünü azaltmakta ve tehditlere karşı daha hızlı ve etkin müdahale edilmesini sağlamaktadır.

3.3.3 BERT dil modeli ve çalışma prensibi

Son yıllarda doğal dil işleme alanında yaşanan en önemli gelişmelerden biri, derin öğrenme tabanlı büyük dil modellerinin ortaya çıkmasıdır. Bu modeller arasında özellikle transformer mimarisine dayanan BERT, NLP görevlerinde önemli bir dönüm noktası haline gelmiştir. BERT, Google araştırmacıları tarafından geliştirilen ve yayınlanan büyük bir dil modelidir [91]. Metin sınıflandırma, soru-cevap sistemleri, duygu analizi ve adlandırılmış varlık tanıma gibi birçok NLP görevinde üstün performans göstermektedir. BERT'in temel yeniliği, önceki dil modellerinin aksine, metin içerisindeki kelimelerin bağlamını-anlamını çift yönlü (bidirectional) olarak öğrenebilmesidir. Geleneksel dil modelleri genellikle tek yönlü (soldan sağa veya sağdan sola) bağlam bilgisi kullanırken, BERT modeli kelimelerin hem soldan hem de sağdan bağlamını eş zamanlı olarak dikkate almaktadır. BERT'in bu çift yönlü bağlam öğrenme yeteneği, özellikle kelimelerin anlamlarının bağlama göre değiştiği durumlarda (örneğin çok anlamlı kelimeler) büyük avantaj sağlamaktadır [91]. BERT dil modeli, transformer mimarisinin yalnızca kodlayıcı (encoder) kısmını kullanmaktadır. Transformer mimarisi, Vaswani ve arkadaşları [94] tarafından önerilen ve öz dikkat (self-attention) mekanizmasına dayanan bir derin öğrenme mimarisidir. Öz dikkat mekanizması, bir cümledeki her kelimenin diğer kelimelerle olan ilişkisini doğrudan modelleyerek, uzun mesafeli bağımlılıkları daha etkin bir şekilde yakalayabilmektedir. Bu sayede BERT, uzun ve karmaşık cümlelerdeki ilişkileri daha iyi öğrenebilmektedir. BERT'in eğitim süreci Şekil 3.6 :'da gösterildiği gibi iki temel aşamadan oluşmaktadır: ön eğitim (pre-training) ve ince ayar (fine-tuning). Ön eğitim aşamasında, model büyük ölçekli genel metin verileri (örneğin Wikipedia ve BookCorpus) üzerinde eğitilerek genel dil bilgisi ve bağlamsal ilişkileri öğrenmektedir. Bu aşamada kullanılan temel yöntemlerden biri maskeli dil modeli tekniğidir. Maskeli dil modeli yönteminde, giriş metnindeki kelimelerin yaklaşık %15'i rastgele olarak maskelenmektedir ve modelin görevi, maskelenmiş kelimeleri bağlamdan hareketle doğru şekilde tahmin etmektir. Bu yöntem, dil modelinin kelimelerin bağlamlarını derinlemesine kavramasını sağlamaktadır [91]. BERT modelinin genel mimarisi ve eğitim süreci Şekil 3.6'da gösterilmiştir. Şekilde, modelin ön eğitim aşamasında maskelenmiş kelimelerle bağlam öğrenme süreci ve ince ayar aşamasında ise önceden öğrenilen bilgilerin farklı NLP görevlerine uyarlanması detaylı olarak sunulmaktadır. Giriş verilerinin tokenizasyonu, token gömme, segment gömme ve pozisyon gömme katmanları ile transformer kodlayıcı katmanlarının işleyişi görsel olarak açıklanmaktadır. Ön eğitim aşamasından sonra, BERT modeline belirli NLP görevleri için ince ayar yapılabilmektedir.

İnce ayar aşamasında, ön eğitimli BERT modelinin üzerine görev odaklı ek katmanlar eklenebilir ve model, ilgili görev için optimize edilebilir. Bu süreç, transfer öğrenme (transfer learning) prensibine dayanmaktadır ve modelin farklı görevlerde yüksek performans göstermesine olanak tanımaktadır. Örneğin, sınıflandırma görevleri için basit bir doğrusal katman eklenirken, adlandırılmış varlık tanıma (NER) gibi dizisel etiketleme görevleri için genellikle CRF (Conditional Random Fields) veya doğrusal sınıflandırıcı katmanları eklenebilmektedir.



Şekil 3.6 : BERT dil modelinin genel mimarisi ve eğitim süreci.

BERT modeli, temel olarak BERTBase ve BERTLarge olarak iki farklı boyutta üretilmiştir. BERTBase modeli 12 transformer katmanı, 768 boyutlu gizli katman ve her katmanda 12 dikkat başlığı içerirken, BERTLarge modeli 24 transformer katmanı, 1024 boyutlu gizli katman ve her katmanda 16 dikkat başlığı içermektedir. Bu iki model arasındaki temel fark, katman sayısı ve parametre sayısıdır. BERTBase yaklaşık 110 milyon parametreye sahipken, BERTLarge yaklaşık 340 milyon parametreye sahiptir [91]. BERT'in başarısı, NLP alanında birçok yeni modelin geliştirilmesine de öncülük etmiştir. RoBERTa [95], DistilBERT [96], ALBERT [97] ve XLM-RoBERTa [98] gibi modeller, BERT'in temel prensiplerini korurken, eğitim verisi, eğitim süreci ve model mimarisinde yapılan iyileştirmelerle daha yüksek performans ve daha düşük hesaplama maliyeti sağlamayı amaçlamaktadır.

Bu çalışmada, BERT dil modeli, siber güvenlik alanında adlandırılmış varlık tanıma görevini gerçekleştirmek amacıyla kullanılmıştır. Siber tehdit istihbaratı metinlerinden tehdit aktörleri, zararlı yazılımlar, hedef sistemler ve saldırı kampanyaları gibi kritik varlıkların otomatik olarak çıkarılması için BERT tabanlı modellerin kullanımı, geleneksel yöntemlere kıyasla daha yüksek doğruluk ve etkinlik sağlamaktadır. Bu bağlamda, BERT'in çift yönlü bağlam öğrenme yeteneği ve transfer öğrenmesi yaklaşımı, siber güvenlik alanındaki karmaşık ve teknik metinlerin analizinde önemli avantajlar sunmaktadır.

3.3.4 Siber güvenlik alanında varlık çıkarımı

Siber güvenlik alanında varlık çıkarımı, tehdit istihbaratı raporları, güvenlik günlükleri, sosyal medya paylaşımları ve forumlar gibi çeşitli kaynaklardan kritik öneme sahip varlıkların otomatik olarak tanımlanması ve sınıflandırılması işlemidir. Bu varlıklar arasında IP adresleri, alan adları, zararlı yazılım isimleri, saldırı türleri, tehdit aktörleri, güvenlik açıkları (CVE numaraları), hash değerleri ve diğer teknik göstergeler bulunmaktadır [89]. Siber güvenlikte varlık çıkarımı, tehdit istihbaratı süreçlerinin otomatikleştirilmesi ve hızlandırılması açısından kritik öneme sahiptir. Siber tehdit istihbaratı analistleri, günlük olarak büyük miktarda metinsel veriyi incelemek ve bu veriler içerisinden tehditlerle ilgili anlamlı bilgileri çıkarmak zorundadır. Bu süreç manuel olarak gerçekleştirildiğinde zaman alıcı, maliyetli ve hata yapmaya açık hale gelmektedir [12]. Bu nedenle, otomatik siber varlık çıkarımı yöntemleri, güvenlik uzmanlarının iş yükünü azaltmakta ve tehditlere karşı daha hızlı ve etkin müdahale ve savunma mekanizmalarının geliştirilmesini sağlamaktadır.

Siber güvenlikte varlık çıkarımı için kullanılan yöntemler genel olarak kural tabanlı yöntemler, makine öğrenmesi tabanlı yöntemler ve derin öğrenme tabanlı yöntemler olarak sınıflandırılabilir [89, 99]. Kural tabanlı yöntemler, genellikle düzenli ifadeler (regular expressions) ve önceden tanımlanmış sözlükler kullanarak varlıkları tanımlar. Örneğin, IP adresleri, URL'ler ve hash değerleri gibi belirli formatlara sahip siber varlıklar, düzenli ifadeler yardımıyla metinlerden kolaylıkla çıkarılabilir. Ancak, kural tabanlı yöntemler, metinlerdeki karmaşık ve bağlama bağımlı varlıkları tanımlamakta yetersiz kalabilmektedir. Bu doğrultuda, makine öğrenmesi ve derin öğrenme tabanlı yaklaşımlar sayesinde metinlerdeki bağlam daha doğru şekilde analiz edilerek, siber varlıkların daha isabetli biçimde çıkarılması mümkün hale gelmektedir.

Makine öğrenmesi tabanlı yöntemler, etiketlenmiş veri kümelerinden öğrenilen modeller aracılığıyla varlıkları tanımlar. Bu yöntemler arasında Koşullu Rastgele Alanlar (Conditional Random Fields - CRF), Destek Vektör Makineleri (Support Vector Machines - SVM) ve Karar Ağaçları (Decision Trees) gibi algoritmalar sıklıkla kullanılmaktadır. Ancak son yıllarda, makine öğrenmesi tabanlı yaklaşımların yerini giderek daha fazla derin öğrenme tabanlı yöntemler almaktadır. Özellikle Transformer mimarisi üzerine inşa edilen BERT gibi büyük dil modelleri, bağlam bilgisini daha etkin şekilde yakalayarak varlık tanıma performansında kayda değer iyileştirmeler sağlamıştır [91].

Siber güvenlikte varlık çıkarımı sürecinde karşılaşılan temel zorluklardan biri, alanın kendine özgü terminolojisi ve standart dışı dil kullanımıdır. Siber güvenlik metinleri, teknik terimler, kısaltmalar, jargonlar ve standart dışı ifadeler içerebilir. Bu durum, genel amaçlı varlık çıkarımı modellerinin başarımını düşürmektedir [100]. Bu nedenle, siber güvenlik alanına özgü veri kümelerinin oluşturulması ve bu veri kümeleri üzerinde eğitilmiş modellerin kullanılması, varlık çıkarımı yöntemlerinin başarımını artırmak için kritik öneme sahiptir. Bir diğer zorluk ise, siber güvenlik metinlerinin sürekli olarak değişen ve gelişen tehdit ortamına bağlı olarak güncellenmesi gerekliliğidir. Yeni tehdit türleri, zararlı yazılımlar ve saldırı yöntemleri sürekli olarak ortaya çıkmakta ve bu durum varlık çıkarımı modellerinin güncellenmesini ve yeniden eğitilmesini gerektirmektedir [93]. Bu nedenle, varlık çıkarımı sistemlerinin sürekli öğrenme ve yeni varlık türlerini tanıma yeteneklerine sahip olması kritik şekilde önemlidir.

Sonuç olarak, siber güvenlikte varlık çıkarımı, tehdit istihbaratı süreçlerinin otomatikleştirilmesi ve hızlandırılması açısından kritik bir rol oynamaktadır. Kural tabanlı yöntemler, makine öğrenmesi ve derin öğrenme yöntemleri gibi farklı yaklaşımlar kullanılarak gerçekleştirilen geleneksel varlık çıkarımı yöntemleri, siber güvenlik profesyonellerinin iş yükünü azaltmakta ve tehditlere karşı daha hızlı aksiyon alma ve etkin bir şekilde müdahale edilmesini sağlamaktadır. Ancak, siber güvenlij alanının kendine özgü, yukarıda bahsedilen zorlukları sebebiyle, siber güvenlik alanına özgü veri kümelerinin oluşturulması ve sürekli öğrenme yeteneğine sahip modellerin geliştirilmesi kritik önem arz etmektedir.

3.4 Çizge Veri Tabanları

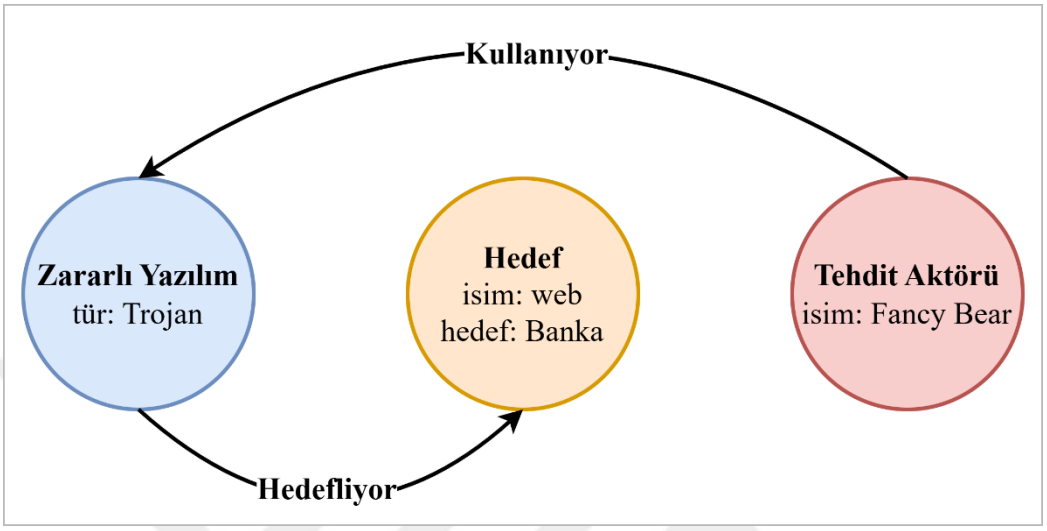
Son yıllarda büyük veri uygulamalarının yaygınlaşmasıyla birlikte, ilişkisel veri tabanlarının karmaşık ilişkileri yönetme ve sorgulama konusundaki sınırlılıkları daha

belirgin hale gelmiştir. Bu durum, çizge veri tabanlarının (Graph Databases) önemini artırmış ve birçok alanda tercih edilmesine neden olmuştur. Çizge veri tabanları, verileri düğümler (nodes) ve bu düğümler arasındaki ilişkiler (edges) şeklinde saklayan, sorgulayan ve yöneten veri tabanı sistemleridir [101]. Çizge veri tabanları, özellikle sosyal ağ analizi, tavsiye sistemleri, dolandırıcılık tespiti ve siber güvenlik gibi ilişkilerin yoğun olduğu alanlarda etkin çözümler sunmaktadır [102]. Çizge veri tabanlarının temel avantajları arasında ilişkisel veri tabanlarına göre daha hızlı sorgulama performansı, esnek veri modeli, karmaşık ilişkilerin kolayca modellenebilmesi ve görselleştirme kolaylığı bulunmaktadır [103]. İlişkisel veri tabanlarında karmaşık ilişkileri sorgulamak için çok sayıda tablo birleştirme (join) işlemi gerekirken, çizge veri tabanlarında ilişkiler doğrudan saklandığı için sorgulama performansı önemli ölçüde artmaktadır [102]. Çizge veri tabanları, NoSQL veri tabanları kategorisinde yer almakta olup, ilişkisel olmayan veri modellerini desteklemektedir. NoSQL veri tabanları, ilişkisel veri tabanlarının katı şema yapısına alternatif olarak ortaya çıkmış ve büyük veri uygulamalarında yaygın olarak kullanılmaktadır [104]. Çizge veri tabanları, NoSQL veri tabanları içerisinde özellikle ilişkisel verilerin modellenmesi ve sorgulanması açısından öne çıkmaktadır.

3.4.1 Çizge veri tabanı kavramları

Çizge veri tabanları, temel olarak üç ana bileşenden oluşmaktadırlar. Bunlar; düğümler (nodes), ilişkiler (edges) ve özelliklerdir (properties). Bu bileşenler, çizge veri tabanlarının temel yapı taşlarını oluşturmakta ve verilerin çizge yapısında saklanmasını sağlamaktadır. Çizge veri tabanlarında düğümler, varlıkları temsil edtmektedir. Örneğin, sosyal ağlarda kullanıcılar, siber güvenlikte IP adresleri, zararlı yazılımlar veya tehdit aktörleri gibi düğüm olarak modellenebilirler. Her düğüm, benzersiz bir kimliğe (Node ID) sahiptir ve çeşitli özellikler içerebilir [102]. İlişkiler, düğümler arasındaki bağlantıları temsil etmektedir. İlişkiler yönlü veya yönsüz olabilmekte ve genellikle düğümler arasındaki etkileşimleri, bağlantıları veya ilişkileri ifade etmektedir. Örneğin, sosyal ağlarda kullanıcılar arasındaki arkadaşlık ilişkileri, siber güvenlikte ise zararlı yazılımlar ve bu yazılımları kullanan tehdit aktörleri arasındaki bağlantılar ilişki olarak modellenebilir. Özellikler, düğümler ve ilişkiler hakkında ek bilgiler sağlamaktadır. Özellikler, veri tabanında anahtar-değer (key-value) çiftleri şeklinde saklanmaktadır ve düğümlerin veya ilişkilerin detaylarını ifade etmektedir. Örneğin, kullanıcı adında bir düğümde ad, yaş ve konum gibi bilgiler özellik olarak saklanabilir. Çizge veri tabanlarında sorgulama işlemleri, genellikle çizge sorgu dilleri kullanılarak gerçekleştirilmektedir. Çizge sorgu dilleri arasında

en yaygın kullanılanlardan biri Neo4j tarafından geliştirilen Cypher sorgu dilidir. Cypher, çizge veri tabanlarında düğümler ve ilişkiler üzerinde sorgulama yapmak için basit ve sezgisel bir sözdizimi sunmaktadır [105]. Çizge veri tabanlarının temel kavramlarını daha iyi anlaşılması için Şekil 3.7'de örnek bir çizge veri tabanının temel yapısı gösterilmektedir.



Şekil 3.7 : Çizge veri tabanı temel yapısı örneği.

Bu örnekte, Zararlı yazılım, Hedef ve Tehdit aktörü düğümleri bulunmaktadır. Zararlı yazılım düğümü, Tehdit aktörü düğümüyle "Kullanıyor" ilişkisi vasıtasıyla bağlantılıdır. Hedef düğümü ise Zararlı yazılım düğümüne "Hedefliyor" ilişkisiyle bağlanmaktadır. Her bir düğüm ve ilişki, ilgili varlıkları tanımlayan özelliklere sahiptir. Buna benzer çizge yapıları, özellikle siber güvenlikte tehdit aktörlerinin aktivitelerinin izlenmesi ve analiz edilmesi için faydalı bir görselleştirmeler sağlamaktadır. Çizge veri tabanlarının bir diğer önemli yanı ise kullanılan çizge algoritmalarıdır. Çizge algoritmaları, çizge veri tabanlarında saklanan veriler üzerinde analiz yapmak için kullanılmaktadır. Bu algoritmalar arasında en kısa yol algoritmaları (örneğin Dijkstra algoritması), merkezilik ölçümleri (örneğin PageRank algoritması) ve topluluk tespiti algoritmaları gibi (örneğin Louvain algoritması) siber güvenlik uzmanlarına yararlı bilgiler sağlayacak algoritmalar bulunmaktadır [106]. Çizge algoritmaları, özellikle siber güvenlik alanında tehditlerin tespiti ve analizinde etkin olarak kullanılmaktadır.

Sonuç olarak, çizge veri tabanları, ilişkisel veri tabanlarına göre karmaşık ilişkileri daha etkin bir şekilde modelleyebilme ve sorgulayabilme yeteneğine sahiptir. Bu özellikleri sayesinde çizge veri tabanları, siber güvenlik gibi ilişkilerin yoğun olduğu alanlarda önemli avantajlar sağlamaktadır. Çizge veri tabanlarının temel kavramları olan düğümler, ilişkiler

ve özellikler, verilerin çizge yapısında saklanmasını ve sorgulanmasını mümkün kılmaktadır.

3.4.2 Neo4j çizge veri tabanı

Neo4j, çizge veri tabanları arasında en yaygın kullanılan ve popüler olan açık kaynaklı bir çizge veri tabanı sistemidir. Neo4j, verileri düğümler (nodes), ilişkiler (edges) ve özellikler (properties) şeklinde saklayan, sorgulayan ve yöneten bir veri tabanı sistemidir [107]. Neo4j, özellikle karmaşık ilişkilerin yoğun olduğu sosyal ağ analizi, tavsiye sistemleri, dolandırıcılık tespiti, biyoinformatik ve siber güvenlik gibi alanlarda etkin çözümler sunmaktadır [102]. Neo4j, ilişkisel veri tabanlarına kıyasla karmaşık ilişkileri daha hızlı ve etkin bir şekilde sorgulayabilme yeteneğine sahiptir. Bu avantaj, Neo4j'nin ilişkileri doğrudan saklaması ve sorgulama sırasında tablo birleştirme işlemlerine ihtiyaç duymamasından kaynaklanmaktadır [108]. Neo4j, çizge veri tabanı sorgulama dili olarak Cypher dilini kullanmaktadır. Cypher, SQL benzeri basit ve sezgisel bir sözdizimine sahip olup, çizge verileri üzerinde sorgulama yapmak için geliştirilmiştir [105]. Neo4j'nin temel bileşenlerinden olan düğümler, varlıkları temsil eder ve benzersiz kimliklere sahiptir. İlişkiler ise düğümler arasındaki bağlantıları temsil etmektedir ve yönlüdür. Özellikler ise düğümler ve ilişkiler hakkında ek bilgiler sağlamaktadır ve anahtar-değer çiftleri şeklinde saklanmaktadır. Neo4j, çizge veri tabanı sistemleri arasında ölçeklenebilirlik, performans ve kullanım kolaylığı açısından öne çıkmaktadır. Neo4j, büyük ölçekli veri kümelerinde bile yüksek performanslı sorgulama ve analiz işlemlerini desteklemektedir [102]. Ayrıca Neo4j, çizge algoritmaları ve analitik işlemler için özel kütüphaneler ve araçlar sunmaktadır. Bu araçlar arasında Neo4j Graph Data Science (GDS) kütüphanesi bulunmaktadır. GDS kütüphanesi, merkezilik ölçümleri, topluluk tespiti ve en kısa yol algoritmaları gibi çeşitli çizge algoritmalarını içermektedir. Neo4j mimarisi, temel olarak üç bilşelenden oluşmaktadır. Bu bileşenlerden ilki depolama motorudur ve çizge verilerini disk üzerinde saklayıp ve yönetmekten sorumludur. Bir diğer bileşeni, sorgu motorudur ve Cypher sorgularını yorumlak ve çalıştırmaktan sorumludur. Son bileşen ise işlem motorudur. Bu bileşen, veri bütünlüğünün sağlanması görevini yerine getirmekte ve veri tutarlılığını sağlamaktadır. Neo4j, siber güvenlik alanında da yaygın olarak kullanılmaktadır. Özellikle tehdit istihbaratı, saldırı analizi ve zararlı yazılım tespiti gibi alanlarda Neo4j'nin çizge tabanlı analiz yetenekleri önemli avantajlar sağlamaktadır. Örneğin, Mahmoud ve arkadaşları [109], Neo4j tabanlı bir sistem geliştirerek, gelişmiş kalıcı tehditlerin tespitini

sağlamışlardır. Bu ve benzeri çalışmar, Neo4j ve benzeri çizge veri tabanlarının siber güvenlik alanındaki etkinliğini ve önemini ortaya koymaktadır.

Sonuç olarak, Neo4j çizge veri tabanı, karmaşık ilişkileri etkin bir şekilde modelleyebilme, sorgulayabilme ve analiz edebilme yeteneği sayesinde birçok alanda tercih edilmektedir. Neo4j'nin esnek veri modeli, yüksek performanslı sorgulama yetenekleri ve çizge algoritmaları desteği, özellikle siber güvenlik gibi ilişkilerin yoğun olduğu alanlarda önemli avantajlar sağlamaktadır.

3.4.3 Çizge veri tabanlarının siber güvenlikte kullanımı

Son yıllarda siber tehditlerin karmaşıklığı ve çeşitliliği artmış, geleneksel yöntemlerle bu tehditlerin tespiti ve analizi giderek daha zor bir hal almıştır. Siber tehditlerin karmaşık yapısı, saldırganların eylemlerinin genellikle birden fazla hedef ve aktör arasında karmaşık ilişkiler ağı oluşturmasına yol açmaktadır. Bu durum, tek bir noktaya odaklanmak yerine, eş zamanlı ve birbirine bağlı birçok saldırının gerçekleşebileceği anlamına gelmektedir. Bu durum, ilişkisel veri tabanları gibi geleneksel veri yönetim sistemlerinin yetersiz kalmasına yol açmıştır. Günümüz siber güvenliği, bu çok boyutlu ilişkileri anlamayı ve buna uygun kapsamlı savunma yöntemleri geliştirmeyi gerektirmektedir. Çizge veri tabanları, bu tür karmaşık ilişkileri etkin bir şekilde modelleyebilme ve sorgulayabilme yeteneği sayesinde siber güvenlik alanında önemli avantajlar sağlamaktadır [109, 110]. Çizge veri tabanları, siber güvenlikte tehdit istihbaratı, zararlı yazılım analizi, ağ güvenliği ve saldırı analizi gibi birçok alanda kullanılmaktadır. Bu kullanım alanları, çizge tabanlı veri analiz yöntemlerinin temelini oluşturmakta ve sonraki bölümlerde detaylandırılacak olan çizge analiz yöntemleri için veri altyapısını sağlamaktadır.

Tehdit istihbaratı, siber tehditlerin erken tespiti ve önlenmesi için kritik öneme sahiptir. Çizge veri tabanları, tehdit istihbaratı verilerinin toplanması, entegrasyonu ve analizinde yaygın olarak kullanılmaktadır. Tehdit istihbaratı verileri genellikle IP adresleri, alan adları, zararlı yazılım hash değerleri, tehdit aktörleri ve saldırı yöntemleri gibi birçok varlık ve ilişki içermektedir. Çizge veri tabanları, bu verileri düğümler ve ilişkiler şeklinde saklayarak tehditlerin hızlı ve etkin bir şekilde analiz edilmesini sağlamaktadır [111]. Zararlı yazılım analizinde de çizge veri tabanları sıklıkla kullanılmaktadır. Zararlı yazılımlar, sistem dosyaları, kayıt defteri girdileri, ağ bağlantıları ve süreçler gibi birçok varlıkla etkileşim halindedir. Çizge veri tabanları, bu etkileşimleri çizge yapısında modelleyerek zararlı yazılımların davranışlarını analiz etmeyi kolaylaştırabilmektedir. Çizge veri tabanları

sayesinde zararlı yazılımların davranışları ve yayılma yöntemleri daha net bir şekilde ortaya konabilmektedir. Çizge veri tabanları, ağ güvenliğinin sağlanmasında ve saldırı analizlerinde de kullanılmaktadır. Ağ güvenliği, siber güvenliğin temel bileşenlerinden biridir ve ağ trafiği verileri, IP adresleri, portlar, protokoller ve kullanıcılar gibi birçok varlık ve ilişki içermektedir. Çizge veri tabanları, bu verileri çizge yapısında saklayarak ağdaki anormal davranışları ve saldırıları tespit etmeyi kolaylaştırmaktadır. Çizge veri tabanları, ağ güvenliği verilerinin analizinde ilişkisel veri tabanlarına göre daha hızlı ve etkin sonuçlar sunmaktadır [110].

3.5 Çizge Tabanlı Veri Analizi

Çizge tabanlı veri analizi, çizge veri tabanlarında saklanan verilerin analiz edilmesi ve yorumlanması için kullanılan yöntemler bütünüdür. Çizge veri tabanları, verileri düğümler ve ilişkiler şeklinde saklayarak karmaşık ilişkileri etkin bir şekilde modelleyebilme yeteneğine sahiptir. Ancak bu verilerin anlamlı hale getirilmesi ve yorumlanması için çizge analiz yöntemlerinin kullanılması gerekmektedir [112]. Çizge tabanlı veri analizi, özellikle sosyal ağ analizi, biyoinformatik, finansal dolandırıcılık tespiti ve siber güvenlik gibi alanlarda yaygın olarak kullanılmaktadır. Çizge tabanlı veri analizinin temel amacı, çizge yapısındaki verilerden anlamlı bilgiler çıkarmak ve bu bilgileri karar alma süreçlerinde kullanmaktır. Çizge analiz yöntemleri, çizge yapısındaki verilerin özelliklerini ortaya çıkarmak, düğümler ve ilişkiler arasındaki bağlantıları analiz etmek ve çizge üzerindeki önemli düğümleri veya toplulukları belirlemek için kullanılmaktadır [112, 113].

3.5.1 Çizge analiz yöntemleri

Çizge analiz yöntemleri, karmaşık ilişkisel verilerin yapısal özelliklerini ortaya çıkarmak ve anlamlandırmak için kullanılan algoritmalar ve teknikler bütününü ifade etmektedir. Bu yöntemler, düğümler (varlıklar) ve kenarlar (ilişkiler) olarak temsil edilen çizge yapılarındaki verilerin sistematik olarak incelenmesini sağlamaktadır. Çizge analizi, özellikle geleneksel veri analiz yöntemlerinin yetersiz kaldığı, yüksek düzeyde bağlantılı veri kümelerinde gizli örüntüleri, toplulukları ve kritik bağlantı noktalarını tespit etmede kullanılmaktadır. Çizge analiz yöntemleri, matematiksel çizge teorisi temelinde geliştirilmiş olup, sosyal ağ analizi, biyoinformatik, telekomünikasyon ağları, ulaşım sistemleri, siber güvenlik ve finansal dolandırıcılık tespiti gibi çeşitli alanlarda yaygın olarak uygulanmaktadır [102]. Bu yöntemler, veri kümesinin topolojik özelliklerini inceleyerek,

düğümler arasındaki doğrudan ve dolaylı bağlantıları, etki alanlarını ve kritik yapısal özelliklerini belirlemektedir. Modern çizge analiz yöntemleri, büyük ölçekli ve dinamik çizgeleri işleyebilme kapasitesine sahiptir ve genellikle paralel hesaplama ve dağıtık sistemler kullanılarak uygulanmaktadır. Bu yöntemler, gerçek zamanlı analiz gerektiren uygulamalarda bile etkili sonuçlar üretebilmektedir. Ayrıca, makine öğrenimi ve yapay zeka teknikleriyle entegre edilerek, çizge yapılarındaki karmaşık desenlerin ve anormalliklerin otomatik olarak tespit edilmesini sağlamaktadır. Çizge analiz yöntemleri, uzmanlık alanı veri bilimi olan araştırmacılara, ilişkisel verilerin derinliklerine inme ve bu verilerdeki gizli bilgileri keşfetme imkanı sunarak, daha bilinçli kararlar alınmasına destek olmakta ve karmaşık sistemlerin daha iyi anlaşılmasına katkıda bulunmaktadır. Çizge analiz yöntemleri genel olarak aşağıdaki kategorilere ayrılabilir.

Merkeziyet analizi (Centrality analysis): Merkeziyet analizi, çizge üzerindeki düğümlerin önem derecesini belirlemek için kullanılan yöntemlerdir. Merkeziyet analizi yöntemleri arasında derece merkeziyeti (degree centrality), yakınlık merkeziyeti (closeness centrality), aradalık merkeziyeti (betweenness centrality) ve özvektör merkeziyeti (eigenvector centrality) bulunmaktadır [101, 114]. Bu yöntemler, özellikle siber güvenlikte kritik düğümlerin (örneğin, saldırganların kullandığı IP adresleri veya zararlı yazılımlar) belirlenmesinde etkin olarak kullanılmaktadır [113].

Derece merkeziyeti, bir düğümün doğrudan bağlantı kurduğu diğer düğümlerin sayısını ölçer ve ağdaki en aktif aktörleri belirlemeye yardımcı olur. Yönlü çizgelerde, gelen bağlantılar ve giden bağlantılar ayrı ayrı değerlendirilebilir. Örneğin, sosyal medya analizinde yüksek bağlantı sayısına sahip kullanıcılar popüler kişileri, yüksek giden bağlantı sayısına sahip kullanıcılar ise bilgi yayıcıları olarak temsil edilmektedir. Freeman [115] tarafından geliştirilen bu temel merkeziyet ölçütü, ağ yapısının ilk incelemelerinde sıklıkla kullanılmaktadır ancak dolaylı bağlantıları hesaba katmadığı için karmaşık ağlarda tek başına yetersiz kalabilmektedir.

Yakınlık merkeziyeti, bir düğümün ağdaki diğer tüm düğümlere olan ortalama mesafesini ölçmektedir ve bilginin ağ içinde ne kadar hızlı yayılabileceğini göstermektedir. Bu ölçüt, özellikle bilgi yayılımı, hastalık yayılımı ve etki analizi çalışmalarında kritik öneme sahiptir. Siber güvenlik bağlamında, yakınlık merkeziyeti yüksek olan düğümler, ağ içinde hızla yayılabilecek zararlı yazılımların potansiyel başlangıç noktalarını temsil edebilmtekdir.

Arasındalık/aradalık merkeziyeti, bir düğümün ağdaki en kısa yollar üzerinde bulunma sıklığını ölçer ve ağ içindeki bilgi akışını kontrol eden köprü düğümleri belirlemeye yardımcı olur. Siber güvenlik alanında, yüksek aradalık merkeziyetine sahip düğümler genellikle farklı alt ağlar arasında bağlantı kuran kritik noktaları temsil etmektedir ve bu düğümlerin hedef alınması veya izlenmesi, ağ savunmasında stratejik öneme sahiptir. Örneğin, botnet tespitinde, komuta kontrol sunucuları ile enfekte cihazlar arasındaki iletişimi sağlayan aracı düğümler, yüksek aradalık merkeziyeti değerleriyle tespit edilebilir [116].

Özvektör merkeziyeti ise, bir düğümün önemini sadece bağlantı sayısına göre değil, bağlantı kurduğu düğümlerin önemine göre de değerlendirir. Bonacich [117] tarafından geliştirilen bu ölçüt, Google'ın PageRank algoritmasının da temelini oluşturmaktadır ve özellikle etki analizi ve prestij değerlendirmesinde kullanılmaktadır. Siber güvenlik uygulamalarında, özvektör merkeziyeti, zararlı yazılım ağlarındaki kritik bileşenleri ve botnet komuta kontrol yapılarını belirlemede etkilidir. Ayrıca, finansal dolandırıcılık tespitinde, şüpheli işlem ağlarındaki kilit aktörleri tanımlamak için de kullanılmaktadır [118].

Topluluk tespiti (Community detection): Topluluk tespiti, çizge üzerindeki düğümlerin benzerliklerine göre gruplandırılması işlemine denmektedir. Bu yöntem, çizge üzerindeki düğümleri benzer özelliklere veya ilişkilere sahip topluluklar halinde gruplandırarak verilerin yapısal özelliklerini ortaya çıkarmaktadır. Topluluk tespiti yöntemleri arasında Louvain algoritması, Girvan-Newman algoritması ve Label Propagation algoritması bulunmaktadır [101]. Siber güvenlikte, tehdit aktörlerinin veya zararlı yazılımların oluşturduğu toplulukların belirlenmesinde bu yöntemler sıklıkla kullanılmaktadır.

Louvain algoritması, büyük ölçekli ağlarda topluluk yapılarını hızlı ve etkili bir şekilde tespit etmek için Blondel ve arkadaşları [119] tarafından geliştirilmiştir. Bu algoritma, ağdaki toplulukları, topluluk içi bağlantıların topluluklar arası bağlantılara göre yoğunluğunu maksimize edecek şekilde belirlemektedir. Siber güvenlik alanında, Louvain algoritması özellikle botnet tespitinde ve kötü amaçlı yazılım ailelerinin sınıflandırılmasında kullanılmaktadır. Varshini ve arkadaşları [120], zararlı yazılım sınıflandırması için çalışmalarında Louvain algoritmasını kullanarak sistem çağrısı çizgelerinden elde edilen topluluk yapılarını ortaya çıkarmış ve bu yapılar üzerinden elde edilen topluluk düzeyinde özellikleri sosyal ağ analizi metrikleriyle birleştirerek makine öğrenmesi modellerine girdi olarak kullanmışlardır.

Girvan-Newman algoritması, Newman ve Girvan [121] tarafından geliştirilen ve toplulukları belirlemek için ağdaki kenarların aradalık merkeziyetini kullanan hiyerarşik bir bölme yöntemidir. Bu algoritma, ağdaki en yüksek aradalık merkeziyetine sahip kenarları tekrarlı bir şekilde kaldırarak, ağı doğal topluluklara ayırmaktadır. Algoritma, özellikle topluluklar arası sınırları belirlemede etkilidir ancak hesaplama karmaşıklığının yüksek olması nedeniyle büyük ağlarda uygulanması zorlaşabilir [122]. Siber güvenlik uygulamalarında Girvan-Newman algoritması, APT gruplarının davranışlarına dayalı olarak oluşturulan bilgi ağlarında benzer özellikler gösteren grupları belirlemek ve bu gruplar arasındaki topluluk yapılarını ortaya çıkarmak amacıyla, Wenhao ve arkadaşları [123] tarafından önerilen benzeri çalışmalarda etkili bir şekilde kullanılmaktadır.

Label Propagation algoritması, Raghavan ve arkadaşları [124] tarafından önerilen, düşük hesaplama karmaşıklığına sahip ve büyük ağlarda etkili olan bir topluluk tespit yöntemidir. Algoritmada her düğüme başlangıçta benzersiz bir etiket atanır ve ardından düğümler, çevrelerindeki komşular arasında en sık görülen etiketi kendi etiketi olarak günceller; bu süreç, ağda benzer özelliklere sahip düğümlerin aynı topluluk altında toplanmasını sağlamaktadır. Süreç, etiketlerin dengeli bir duruma ulaşmasına kadar devam etmekte ve sonuçta aynı etikete sahip düğümler bir topluluk oluşturmaktadır. Siber güvenlik alanında, Label Propagation algoritması, büyük ve dinamik yapıya sahip, kısmen etiketlenmiş veri kümelerinde hızlı ve verimli etiket tahmini yapabilme kabiliyeti sayesinde tehdit sınıflandırma, anomali tespiti ve olay ilişkilendirme gibi görevlerde kullanılmaktadır. Shu ve arkadaşları [125], Label Propagation algoritması ile sürekli değişen verilerde her seferinde baştan hesaplama yapma gereksinimini ortadan kaldırmak ve etiketleme sürecinin verimliliğini artırmak amacıyla uygulamışlardır.

Topluluk tespiti yöntemleri, siber güvenlik alanında özellikle saldırı tespiti, anomali analizi ve zararlı yazılım sınıflandırması gibi kritik görevlerde önemli rol oynamaktadır. Bu yöntemler, ağ yapılarındaki benzer davranışları gruplandırarak tehditlerin daha hızlı ve doğru bir şekilde belirlenmesine olanak tanımaktadır.

3.5.2 Çizge temelli modellerle siber tehdit analizi

Son yıllarda siber tehditlerin karmaşıklığı, çeşitliliği ve karmaşıklığı önemli ölçüde artış göstermektedir. Günümüz siber saldırıları, yalnızca tekil hedeflere yönelik basit saldırılar olmaktan çıkıp, birden fazla varlık arasında karmaşık ilişkiler içeren, çok boyutlu, çok aşamalı ve dinamik süreçlere dönüşmüştür. Bu dönüşüm, geleneksel tehdir analiz

yöntemlerinin siber tehdit ortamını kavramada ve modellemede yetersiz kalmasına neden olmaktadır. Söz konusu yetersizlik, çizge teorisi temelli modellerin ve analiz yaklaşımlarının siber tehdit analizinde kullanımını zorunlu hale getirmiştir. Özellikle, gelişmiş kalıcı tehditler (APT) gibi karmaşık saldırıların tespiti ve analizi için çizge tabanlı yaklaşımlar, geleneksel yöntemlere göre daha etkili sonuçlar vermektedir [110].

Çizge temelli modeller, düğümler (varlıklar) ve kenarlar (ilişkiler) aracılığıyla karmaşık ilişkisel yapıları temsil etme kapasitesine sahiptir. Bu özellik, siber güvenlik bağlamında IP adresleri, domain adları, dosya hash değerleri, kullanıcı hesapları ve sistem bileşenleri gibi çeşitli varlıklar arasındaki ilişkilerin bütün bir şekilde modellenmesine olanak tanımaktadır. Böylece, geleneksel yöntemlerde gözden kaçabilecek karmaşık saldırı desenleri ve ilişki ağları daha görünür hale gelmektedir. Siber tehdit ortamının sürekli evrim geçiren doğası, tehdit aktörlerinin taktiklerini, tekniklerini ve prosedürlerini (Tactics, Techniques and Procedures - TTP) sürekli olarak geliştirmelerini ve değiştirmelerini içermektedir. Çizge temelli modeller, bu dinamik yapıyı yakalama ve temsil etme konusunda üstün yetenekler sergilemektedir. Özellikle, zamansal çizge modelleri, tehdit aktörlerinin davranışlarındaki değişimleri ve dönüşümü izleme ve analiz etme imkanı sunmaktadır. Bu sayede, siber güvenlik analistleri, tehdit aktörlerinin gelecekteki olası davranışlarını tahmin etme ve önleyici savunma stratejileri geliştirme konusunda daha donanımlı hale gelmektedir [6]. Çizge temelli modellerin siber güvenlik alanında sağladığı bir diğer önemli avantaj, büyük ölçekli ve heterojen veri kümelerini etkili bir şekilde işleyebilme kapasitesidir. Modern siber güvenlik ortamları, ağ trafiği kayıtları, sistem logları, tehdit istihbaratı verileri ve kullanıcı davranış analizleri gibi çeşitli kaynaklardan gelen devasa miktarda veriyi içermektedir. Çizge veri tabanları ve çizge işleme algoritmaları, bu heterojen ve büyük ölçekli verileri entegre etme, ilişkilendirme ve anlamlı içgörüler çıkarma konusunda üstün performans sergilemektedir. Ayrıca, çizge temelli modeller, siber güvenlik alanında anomali tespiti, tehdit istihbaratı, saldırı yüzeyi analizi ve adli bilişim gibi çeşitli uygulamalarda giderek daha fazla kullanılmaktadır. Özellikle, çizge tabanlı anomali tespiti algoritmaları, normal ağ davranışından sapmaları tespit etmede ve potansiyel saldırıları erken aşamada belirleme konusunda etkili sonuçlar vermektedir. Benzer şekilde, tehdit istihbaratı alanında, çizge modelleri, farklı tehdit aktörleri, kampanyaları ve taktikleri arasındaki ilişkileri ortaya çıkararak, daha kapsamlı ve bütünleşik bir tehdit manzarası sunmaktadır [27]. Çizge temelli modeller, siber tehdit analizinde farklı yöntem ve yaklaşımları içermektedir. Bu yöntemler arasında çizge benzerlik analizi [126], çizge gömme (graph embedding) [127], çizge sinir

ağları (graph neural networks - GNN) [128] ve çizge tabanlı anomali tespiti [129] gibi yaklaşımlar bulunmaktadır . Her bir yaklaşım, siber tehditlerin farklı yönlerini analiz ve tespit etmek için özel avantajlar sunmaktadır.

Çizge benzerlik analizi, farklı çizge yapılarının benzerliklerini ölçerek, bilinen tehdit desenlerinin yeni varyasyonlarını tespit etmeye olanak tanımaktadır. Bu yaklaşım, özellikle kötü amaçlı yazılım analizi ve sınıflandırması, zararlı yazılımların davranışlarının karşılaştırılması ve saldırıların ortak özelliklerinin belirlenmesi gibi alanlarda etkin çözümler sunmaktadır. Çizge benzerlik analizi, çizgelerin yapısal özelliklerini (düğüm sayısı, kenar yoğunluğu, derece dağılımı vb.) veya içerik özelliklerini (düğüm ve kenar etiketleri, öznitelikleri vb.) karşılaştırarak benzerlik skorları üretip bu skorlarla, bilinen tehdit desenlerinin yeni varyasyonlarını tespit etmek, kötü amaçlı yazılım ailelerini sınıflandırmak veya saldırı kampanyalarını ilişkilendirmek için kullanılmaktadır. Örneğin, bir kötü amaçlı yazılımın davranışsal çizgesi ile bilinen kötü amaçlı yazılım ailelerinin davranışsal çizgeleri arasındaki benzerlik ölçülerek, yeni bir kötü amaçlı yazılım örneğinin hangi aileye ait olduğu belirlenebilmektedir. Çizge benzerlik analizinde kullanılan temel yöntemler arasında çizge eşleştirme (graph matching), çizge düzenleme mesafesi (graph edit distance), çizge çekirdekleri (graph kernels) ve spektral yöntemler bulunmaktadır [130]. Bu yaklaşımların her biri, farklı çizge türleri ve uygulama senaryoları için çeşitli avantajlar sunmaktadır.

Çizge gömme teknikleri, karmaşık çizge yapılarını düşük boyutlu vektör uzaylarına dönüştürerek, çizge verilerinin daha verimli işlenmesini ve analiz edilmesini sağlayan yöntemlerdir. Bu yaklaşım, çizgelerin yapısal ve anlamsal özelliklerini koruyarak, makine öğrenmesi algoritmalarının bu vektörleri kolayca işleyebileceği formata dönüştürmektedir. Çizge gömme teknikleri, özellikle büyük ölçekli ağ verilerinin analizi, düğüm sınıflandırma, bağlantı tahmini ve topluluk tespiti gibi görevler için uygundur. Siber güvenlik alanında, çizge gömme teknikleri, ağ trafiği analizi, kötü amaçlı yazılım tespiti, kullanıcı davranış analizi ve tehdit istihbaratı gibi uygulamalarda kullanılmaktadır. Çizge gömme yöntemleri matris faktörizasyon tabanlı yöntemler, rastgele yürüyüş tabanlı yöntemler ve derin öğrenme tabanlı yöntemler olarak üç kategoriye ayrılamaktadır [131]. Bu yöntemler, çizge yapısındaki düğümlerin, kenarların veya alt çizgelerin vektör temsillerini oluşturarak, karmaşık çizge verilerinin daha etkili bir şekilde analiz edilmesini sağlamaktadır.

Çizge sinir ağları (GNN), çizge yapısındaki verileri işlemek ve analiz etmek için tasarlanmış özel bir derin öğrenme mimarileridir. Geleneksel sinir ağlarının aksine, GNN'ler

çizge yapısındaki düğümler ve kenarlar arasındaki ilişkileri doğrudan modelleyebilme yeteneğine sahiptir. Bu özellik, GNN'leri özellikle ilişkisel verilerin analizi için uygun hale getirmektedir. GNN'ler, her düğümün özelliklerini, komşu düğümlerin özelliklerini ve aralarındaki ilişkileri dikkate alarak, düğüm düzeyinde, kenar düzeyinde veya çizge düzeyinde temsiller oluşturmaktadır. Bu temsiller, düğüm sınıflandırma, bağlantı tahmini, çizge sınıflandırma ve çizge üretimi gibi çeşitli görevler için kullanılmaktadır. Siber güvenlik alanında, GNN'ler özellikle kötü amaçlı yazılım tespitinde, saldırı tespitinde, anomali tespitinde ve tehdit istihbaratı gibi uygulamalarda kullanılmaktadır. GNN'lerin başlıca türleri arasında çizge evrişimli ağlar (Graph Convolutional Network - GCN), çizge dikkat ağları (GAT), çizge LSTM (Long Short-Term Memory) ve çizge otokodlayıcılar bulunmaktadır [132]. Bu mimariler, farklı çizge türleri ve uygulama senaryoları için çeşitli avantajlar sunmaktadır.

Çizge tabanlı anomali tespiti, çizge yapısındaki normal davranış veya desenlerden farklı olanları belirleyerek potansiyel tehditleri tespit etmeye odaklanan bir yaklaşımdır. Bu yöntem, özellikle daha önce bilinmeyen saldırıların tespiti, içeriden tehdit tespiti ve gelişmiş kalıcı tehditlerin tespiti için etkili sonuçlar vermektedir. Çizge tabanlı anomali tespitinde kullanılan temel yöntemler arasında istatistiksel yöntemler, çizge madenciliği tabanlı yöntemler, topluluk tabanlı yöntemler ve derin öğrenme tabanlı yöntemler bulunmaktadır. Siber güvenlik alanında, çizge tabanlı anomali tespiti, ağ trafiği analizi, kullanıcı davranış analizi, sistem çağrı analizi ve tehdit istihbaratı gibi uygulamalarda kullanılmaktadır [133].

3.5.3 Çizge analizinde karşılaşılan zorluklar

Çizge analiz yöntemleri, siber güvenlik alanında karmaşık ilişkileri modelleme ve analiz etme konusunda önemli avantajlar sağlasa da uygulama aşamasında çeşitli zorluklarla karşılaşılmaktadır. Bu zorluklar, çizge tabanlı analiz yöntemlerinin etkinliğini ve doğruluğunu önemli ölçüde etkileyerek karar alma süreçlerini olumsuz yönde etkilemektedir. Çizge analizinde karşılaşılan temel zorluklar; büyük ölçekli çizgelerin yönetimi, veri kalitesi ve eksikliği, çizge algoritmalarının ölçeklenebilirliği, çizge verilerinin dinamik yapısı ve yorumlanabilirlik sorunları olarak sıralanabilir [134]. Çizge analizinde karşılaşılan en önemli zorluklardan biri, büyük ölçekli çizgelerin yönetimi ve analizidir. Siber güvenlik alanında kullanılan çizge verileri genellikle çok büyük boyutlarda olup çok sayıda düğüm ve ilişkiden oluşabilmektedir. Bu durum, çizge verilerinin depolanması, sorgulanması ve analiz edilmesi süreçlerinde ciddi performans sorunlarına yol

açabilmektedir [135]. Özellikle çizge algoritmalarının hesaplama karmaşıklığı, büyük çizgelerde işlem maliyetlerini önemli derecede artırmakta ve analiz süreçlerini yavaşlatmaktadır. Çizge analizinde karşılaşılan bir diğer önemli zorluk ise veri kalitesi ve veri eksikliğidir. Siber güvenlik alanında kullanılan çizge verileri genellikle farklı kaynaklardan elde edilmekte ve bu kaynaklar arasında tutarsızlıklar, eksik veriler ve hatalı bilgiler bulunabilmektedir [136]. Bu durum, çizge analiz yöntemlerinin doğruluğunu ve güvenilirliğini olumsuz yönde etkilemektedir. Bu bağlamda, Adarsh ve arkadaşları [136], çizge tabanlı analiz yöntemlerinin etkinliğinin büyük ölçüde veri kalitesine bağlı olduğunu vurgulamış ve eksik veya hatalı verilerin analiz sonuçlarını ciddi şekilde bozabileceğini belirtmiştir. Bu nedenle, büyük veriler içeren analizlerde uygulama öncesinde veri temizleme, veri doğrulama ve eksik verilerin tamamlanması gibi ön işlemlerin yapılması büyük önem arz etmektedir [137].

Siber güvenlik alanında kullanılan çizge verileri genellikle dinamik ve değişken bir yapıya sahiptir ve elde edilecek değer için verilerin sürekli olarak güncellenmesi gerekmektektedir [137]. Yeni tehditlerin ortaya çıkması, mevcut tehditlerin değişmesi ve ağ yapılarının sürekli olarak güncellenmesi, çizge verilerinin sürekli olarak güncel tutulmasını gerektirmektedir. Bu durum, çizge analiz yöntemlerinin sürekli olarak güncellenmesini ve dinamik verilere uyum sağlamasını zorunlu kılmaktadır. Çizge analiz yöntemlerinin bir diğer önemli zorluğu ise yorumlanabilirliktir. Özellikle çizge sinir ağları gibi derin öğrenme tabanlı çizge analiz yöntemlerinde, yüksek doğruluk oranları sağlamakla birlikte, sonuçların yorumlanması konusunda ciddi zorluklar yaşanmaktadır [138]. Bu durum, siber güvenlik uzmanlarının analiz sonuçlarını anlamasını ve karar alma süreçlerinde kullanmasını zorlaştırmaktadır. Yuan ve arkadaşları [138], çizge sinir ağlarının yorumlanabilirlik sorunlarını incelemiş ve bu modellerin genellikle kara kutu olarak değerlendirildiğini belirtmiştir. Bu nedenle, çizge analiz yöntemlerinin sonuçlarının yorumlanabilirliğini artırmak için açıklanabilir yapay zeka yöntemlerinin çizge analiz yöntemleriyle entegre edilmesi önerilmektedir.

Sonuç olarak, çizge analiz yöntemleri siber güvenlik alanında önemli avantajlar sağlamakla birlikte, yukarıda belirtilen zorluklar nedeniyle uygulama süreçlerinde çeşitli engellerle karşılaşılmaktadır. Bu zorlukların aşabilmek için daha fazla ölçeklenebilir çizge algoritmalarının geliştirilmesi, uygulama öncesinde veri kalitesinin artırılması ve dinamik çizge analiz yöntemlerinin olgunlaştırılması gibi alanlarda daha fazla araştırma yapılması gerekmektedir.

4. BERT DİL MODELİ VE BİLGİ ÇİZGELERİ KULLANILARAK SİBER TEHDİT İSTİHBARAT VERİLERİNDEN TEHDİT ANALİZİ YAPMAYA YÖNELİK YENİ BİR HİBRİT YAKLAŞIM

Çalışmanın bu bölümü, yapılandırılmamış metin verilerinden siber varlıkların çıkarımından başlayarak, bu varlıkların ayrıntılı analizine kadar uzanan süreci ele alan kapsamlı bir siber tehdit istihbaratı sistemini tanıtmaktadır.

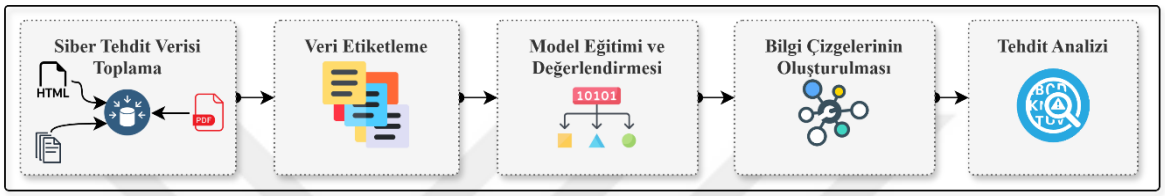
4.1 Giriş

Siber güvenlik alanı, dijital altyapıların karşı karşıya kaldığı çok yönlü tehditlerle birlikte sürekli dönüşmekte; saldırganların yöntemleri giderek daha karmaşık, hedefleri ise daha stratejik hâle gelmektedir. Bu durum, yalnızca meydana gelen saldırılara müdahale etmeyi değil, aynı zamanda siber tehditleri erken evrede tespit etmeyi ve olası saldırı senaryolarını önceden öngörmeyi gerekli kılmaktadır. Bu ihtiyacın karşılanmasında, tehdit istihbaratına dair yayımlanan metin tabanlı belgeler örneğin APT raporları, güvenlik açıklarına ilişkin teknik dökümanlar, üretici blogları ve çeşitli analiz içerikleri oldukça önemli bir rol oynamaktadır. Bu belgeler, saldırıların arkasında yatan motivasyonları, kullanılan teknikleri ve hedef alınan altyapıları anlamlandırmak için zengin birer veri kaynağıdır. Ancak bu veriler büyük ölçüde yapılandırılmamış ve biçimsel olarak oldukça çeşitlidir. Farklı kurumlardan, platformlardan ve kaynaklardan elde edilen belgeler arasında standart bir yapı bulunmaması, bu içeriklerin makine tarafından anlaşılabilir hâle getirilmesini zorlaştırmakta; bu da etkili analizlerin önünde önemli bir engel oluşturmaktadır. Bu nedenle, tehdit verilerinin yalnızca toplanması değil, çok kaynaklı ve biçimsel olarak uyumsuz bu belgelerin ortak bir yapıda merkezi olarak izlenmesi, anlamlandırılması ve bütüncül bir analiz sürecine dönüştürülmesi gerekmektedir.

Bu çalışma, yapay zekâ tabanlı doğal dil işleme yöntemlerini kullanarak siber güvenlik belgeleri üzerinde anlamlı bilgi çıkarımı gerçekleştiren ve çıkarılan bu bilgileri ilişkisel yapılarla analiz edebilen bir sistem önermektedir. Bu sistem, yalnızca verileri ayrıştırmakla kalmaz; aynı zamanda çıkarılan siber varlıkları bir bilgi çizgesi üzerinde yapılandırarak, tehdit aktörleri, saldırı teknikleri ve hedefler arasındaki bağlantıları görünür kılar. Böylece, saldırıların yapısal özelliklerinin, davranışsal örüntülerinin ve zamansal ilişkilerinin analiz edilmesi mümkün hâle gelir. Dağınık kaynaklardan elde edilen metinlerin merkezi olarak toplanması, ön işleme adımlarından geçirilmesi, anlamlı varlık ve ilişki

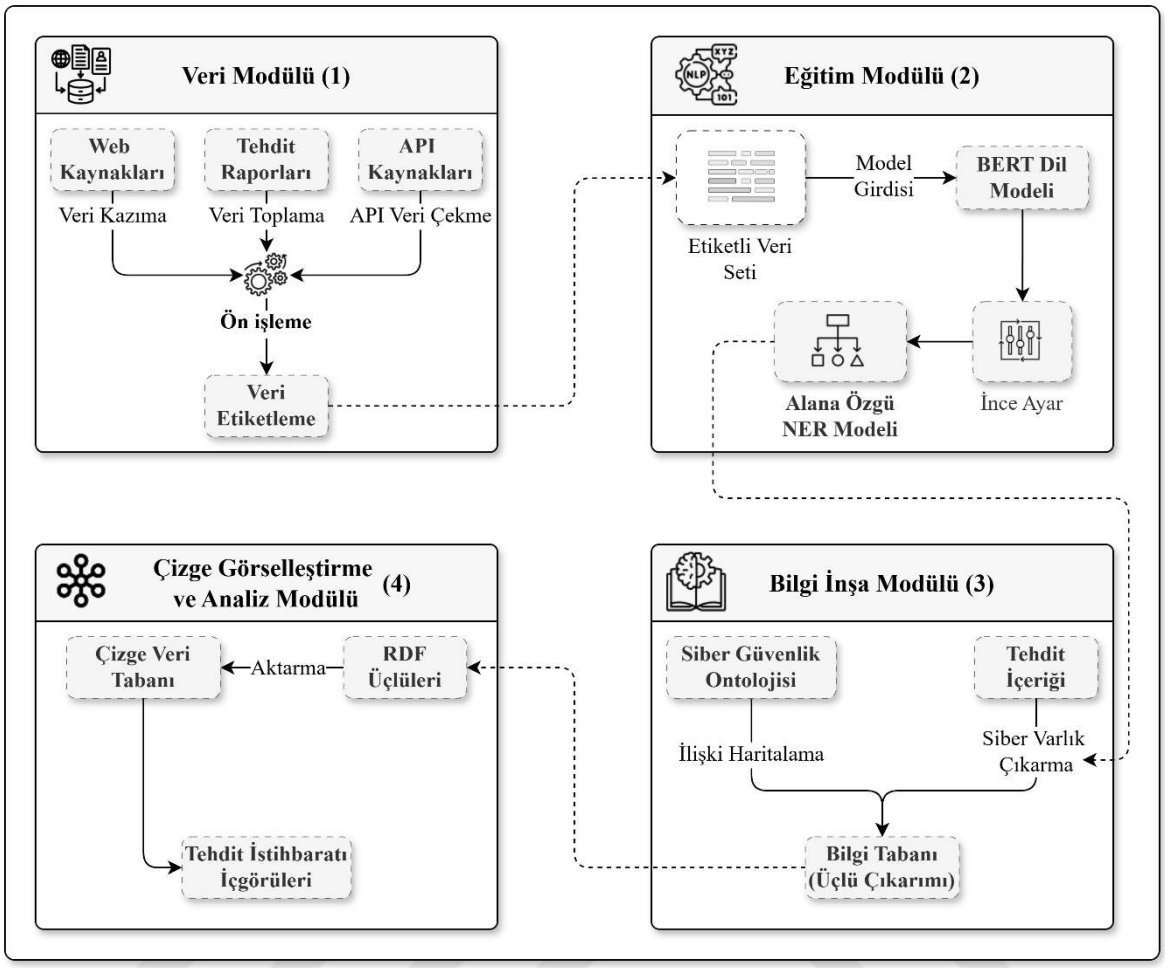
çıkarımı yapılması ve bu çıktılar üzerinden grafik tabanlı analiz gerçekleştirilmesi; çalışmanın genel yaklaşımını oluşturmaktadır.

Önerilen yaklaşımın temel amacı, güvenlik tehdit raporlarına ve ilgili içeriklere gömülü kritik bilgilerin işlenmesini, bu bilgilerin yapılandırılmış formata dönüştürülmesini ve bu yapı üzerinden eyleme geçirilebilir içgörüler elde edilmesini sağlamaktır. Böylece hem anlamlı analiz kolaylaşmakta hem de karar destek süreçleri güçlenmektedir. Sistemin genel mimarisi Şekil 4.1'de özetlenmiş olup, geliştirilen ve önerilen yaklaşım, akademik çıktı olarak uluslararası dergide yayınlanmıştır [148].



Şekil 4.1 : Önerilen sistemdeki temel adımlar.

Birinci adımda, web siteleri, uygulama programlama arayüzleri (Application Programming Interface-API) ve tehdit raporları dahil olmak üzere çeşitli çevrimiçi kaynaklardan metin tabanlı veriler toplanmaktadır. Çeşitli kaynaklardan gelen bu veriler, gereksiz ve ilgisiz öğelerden arındırılarak BERT modelinin eğitim aşaması için ön işlemden geçirilmtekdir. Ön işleme aşamasından sonra veriler, manuel olarak etiketlenmektedir ve eğitim için ön eğitimli BERT modeline girdi olarak verilmtekdir. Etiketlenmiş bu verilerle eğitilen model, tehdit metinlerinden tehdit aktörleri, kötü amaçlı yazılımlar, saldırı kampanyaları ve saldırı hedefleri gibi siber varlıkları çıkaran alana özgü bir model haline gelmektedir. Ham siber tehdit verilerinden çıkarılan siber varlıklar arasındaki ilişkiler Birleşik Siber Güvenlik Ontolojisi (Unified Cybersecurity Ontology -UCO) çerçevesi [139] kullanılarak düzenlenmektedir. Bu ontoloji tehdit aktörleri, kötü amaçlı yazılımlar, saldırı kampanyaları ve saldırı hedefleri arasındaki ilişkileri tanımlar. Bu aşamadan sonra veriler varlık-ilişki-varlık üçlülerine dönüştürülür ve bilgi çizgelerinin oluşturulması için bir bilgi tabanı formuna dönüştürülür. Oluşturulan üçlülerin bilgi çizgelerinde kullanılması, siber varlıklar arasındaki ilişkilerin görselleştirilmesi ve gelişmiş tehdit analizlerinin gerçekleştirilmesi için bir temel sağlamaktadır. Şekil 4.2'de gösterilen önerilen yaklaşım, siber tehdit istihbarat çalışmamızı temsil eden sıralı bir boru hattı olarak tanımlanmaktadır. Bu modüller veri modülü, eğitim modülü, bilgi inşa modülü ve çizge görselleştirme ve analiz modülüdür. Aşağıdaki alt bölümler her bir modülün ayrıntılı açıklamalarını sunmaktadır.



Şekil 4.2 : Önerilen yaklaşımın kapsamlı diyagramı.

4.2 Veri Modülü

Veri modülü önerilen sistemin temelini oluşturmaktadır. Bu aşamada veriler toplanır, işlenir, etiketleme işlemi yapılır ve etiketli veriler bir sonraki modül olan eğitim modülüne gönderilir. Bu modülün alt bileşenleri aşağıda detaylı olarak açıklanmıştır.

4.2.1 Veri toplama

Önerilen sistem için temel girdi olarak ham siber güvenlik verileri kullanılmaktadır. Veri toplama aşamasında bu veriler üç farklı kaynaktan elde edilmekte ve ön işleme aşamasında kullanılmaktadır. Çalışmamızda kullanılan veri kaynakları aşağıda açıklanmıştır.

Web Kaynakları: Web kaynaklarından toplanan veriler, teknik bilgiler içeren siber güvenlik makaleleri ve blog yazıları gibi metin tabanlı içeriklerdir. Bu içerikler web sayfalarından çekildiği için yapılandırılmamış bir formattadır ve söz konusu sayfaya özgü

web kazıma teknikleri kullanılarak çıkarılmaktadır [140]. HTML etiketleri gibi gereksiz öğeleri kaldırmak için web sayfalarından toplanan ham veriler üzerinde kapsamlı bir ön işleme aşaması gerçekleştirilmimştir. Ön işlemden geçirilen veriler sonraki adımlar için txt formatında saklanmıştır.

Tehdit Raporları: Tehdit raporları genel bir GitHub deposundan alınmıştır [141]. İlgili GitHub deposunda depolanan tehdit raporları PDF formatındadır ve depo düzenli olarak yeni raporlarla güncellenmtedir. Bu PDF dosyalarının metin içeriği Python kütüphanesi PyPDF2 [142] kullanılarak çıkarılmıştır. Bu içerikler daha sonra ön işleme tabi tutulmuş ve metinler uygun bir forma dönüştürülmüştür. Metinler daha sonra kullanılmak üzere txt formatında kaydedilmiştir.

API Kaynakları: MITRE ATT&CK [1] yapılandırılmış siber tehdit verileri sağlayan güvenilir bir çerçevedir. Bu kaynaktan indirilen verilere API veya Python paketleri ile erişilebilmektedir. Bu çalışmada, zararlı yazılım, tehdit aktörü ve kampanya listeleri mitreattack-python [143] paketi kullanılarak önerilen sistemde kullanılmak üzere MITRE ATT&CK'den alınmıştır.

Belirtilen kaynaklardan toplanan veriler, gereksiz içeriklerin çıkarılması ve düzenli bir yapıya dönüştürülmesi için ön işleme aşamasına girdi olarak verilmiştir.

4.2.2 Ön işleme

Önerilen sistemin bu aşamasında, çeşitli kaynaklardan toplanan ham siber metinlerin temizlenmesi ve belirli bir standarda dönüştürülmesi için detaylı bir ön işleme süreci gerçekleştirilmiştir. Bu işlem ham metin verilerini yapılandırılmış bir forma dönüştürerek bir sonraki aşama için hazır hale getirmiştir. Temizleme, belirteçleme (tokenizasyon), HTML etiketlerinin kaldırılması ve gereksiz karakterlerin çıkarılması gibi işlemler uygulanmıştır.

Metin Temizleme: Metin temizleme işlemi, ham verilerin netliğini ve tutarlılığını sağlamak için ilgisiz öğelerin çıkarılması işlemlerini içermektedir. Bu aşamada, sonraki modüllerde kullanılan veri setinin belirli bir kaliteye ulaşması için metinler küçük harfe dönüştürülmüş, unicode karakterler normalize edilmiş, fazladan boşluklar ve yeni satırlar kaldırılmış, analitik önemi olmayan “!”, “@” veya “#” gibi alfanümerik olmayan karakterler kaldırılmış ve web sayfalarından elde edilen metinlerden HTML ve script etiketleri kaldırılmıştır. Bu sayede verilerdeki gürültü en aza indirilmiş, verilerin kalitesi artırılmış ve etiketleme süreci için daha uygun hale getirilmiştir.

Belirteçleme: Belirteçleme, metni küçük ve yönetilebilir parçalara bölme işlemidir. Bu aşamada metin önce cümlelere sonra da kelimelere ayrılmakta ve analiz, kelime düzeyinde gerçekleştirilmektedir. Doğal dil işlemede, etkin bir adlandırılmış varlık tanıma işleminin gerçekleştirilebilmesi için belirteçleme aşamasının etkili bir şekilde uygulanması gerekmektedir. Çünkü NER sistemleri metinden varlıkları çıkarırken bu analizi token seviyesinde gerçekleştirmektedirler.

4.2.3 Veri etiketleme

Ön işleme aşamasından sonra, yapılandırılmış bir forma dönüştürülen CTI verileri, model eğitimi için yapılandırılmış bir veri kümesi oluşturmak üzere manuel ve kural tabanlı yöntemler kullanılarak etiketlenmiştir. Etiketleme sürecinde tehdit aktörleri, kötü amaçlı yazılımlar, saldırı kampanyaları ve saldırı hedefleri gibi varlıklar Çizelge 4.1'de gösterildiği gibi etiketlenmiştir. Bu adımda gerçekleştirilen etiketleme işleminin doğruluğu, eğitilecek modelin doğru tahminler yapabilmesi için kritik öneme sahiptir. Bu aşamada MITRE ATT&CK'den alınan zararlı yazılım, saldırı kampanya isimleri ve tehdit aktörü listeleri kural tabanlı yöntemler kullanılarak veri setinin oluşturulmasında kullanılmıştır. Bu sayede eğitilen modelin daha fazla varlık çıkarması hedeflenmiştir.

Bu çalışmada etiketleme için kullanılan veri kümesi birden fazla kaynaktan elde edilmiştir. WeLiveSecurity [140] ve Fortinet [144] adresinden toplam 540 haber makalesi toplanmıştır. Ayrıca, seçilen 100 tehdit raporu halka açık APT Siber Suç Kampanyası Koleksiyonları GitHub deposundan indirilmiştir [141] ve kötü amaçlı yazılım, tehdit aktörleri ve saldırı kampanyaları listeleri MITRE ATT&CK çerçevesinden elde edilmiştir. Bu kaynaklardan 640 belge toplanmış ve 257 adet belge adlandırılmış varlık tanıma modelini eğitmek için elle etiketlenmiştir. Ayrıca [145] çalışmasındaki veriseti de mevcut verisetine eklenerek eğitim sürecinde kullanılılmıştır. Varlıkları etiketlemek için İç-Dış-Başlangıç (Inside-Outside-Beginning - IOB) etiketleme şeması kullanılmıştır. IOB etiketleme şeması NER görevlerinde yaygın olarak kullanılmaktadır. Cümlelerin sınırlarını belirlemek ve metin içindeki varlık kategorilerini temsil etmek için kullanışlı bir yapıya sahiptir.

Çizelge 4.1 : Siber güvenlik için IOB etiketleme şeması örneği.

Kelime	Etiket
The	O
APT29	B-ThreatActor
group	I-ThreatActor
used	O
the	O
malware	O
SUNBURST	B-Malware
in	O
a	O
campaign	B-Campaign
targeting	O
U.S.	B-Target
government	I-Target
agencies	I-Target
.	O

Çalışmada, ThreatActors, Targets, Malware ve Campaigns gibi varlıkların etiketlenmesi işlemi bu modülde gerçekleştirilmiş ve varlıkların dağılımı Çizelge 4.2'de gösterilmiştir. Etiketlenmiş veri kümesi daha sonra eğitim, doğrulama ve test gibi alt kümelere ayrılmıştır. Eğitim setinde 6.340, doğrulama setinde 793 ve test setinde 793 varlık bulunmaktadır.

Çizelge 4.2 : Veriseti etiket dağılımı.

Varlık Tipi	Adet
O (Outside)	7926
B-ThreatActor	2848
I-ThreatActor	1196
B-Target	1746
I-Target	1220
B-Malware	2204
I-Malware	516
B-Campaign	760
I-Campaign	644

IOB etiketleme şemasının yapısı aşağıda açıklanmıştır:

B-Tag (Başlangıç): Bir etiket dizisindeki bir varlığın başlangıcını belirtir (örneğin, bir ThreatActor varlığının başlangıç etiketi için B-ThreatActor).

I-Tag (İç): Aynı varlık içindeki sonraki belirteçleri tanımlar (örneğin, bir tehdit aktörü varlığındaki ikinci belirteç için I-ThreatActor).

O-Tag (Outside): Herhangi bir varlığa ait olmayan belirteçleri tanımlar.

Bu sistematik etiketleme, siber güvenlik metnindeki kilit varlıkları net bir şekilde tanımlayan ve ayıran yapılandırılmış bir verisetinin oluşturulmasını sağlamaktadır. Veri kümesinin etiketlenmeyen belgelerden oluşan kalan kısmı, eğitilmiş NER modelini kullanarak varlıkları çıkarmak için Bilgi Oluşturma Modülü sırasında kullanılmıştır. Bu tür etiketli veri kümeleri, yapılandırılmamış tehdit raporlarındaki varlıkların ve ilişkilerin tanınmasını sağlamak için dil modellerini eğitmek için gereklidir.

4.3 Eğitim Modülü

Eğitim modülü, siber güvenlik alanındaki adlandırılmış varlık tanıma görevi için ön eğitimli bir BERT modeline ince ayar (Fine-tuning) yapmak üzere tasarlanmıştır [91]. BERT'in geleneksel modellere göre en üstün özelliği çift yönlü dönüştürücü yaklaşımı kullanmasıdır. BERT, metindeki her bir kelimenin hem sol hem de sağ bağlamını aynı anda işleyebilmekte ve bu yaklaşımla metinleri bağlamsal olarak daha çok anlamakta ve sayısal gösterimler üretmektedir. Ön eğitimli BERT modeli, ince ayar yapılmadan kullanıldığında metinlerden isimler, yerler, tarihler vb. gibi genel varlıkları çıkarabilmektedir. Metinlerden alana özgü varlıkları çıkarabilmek için, özel olarak tasarlanmış etiketli bir veri kümesi kullanarak BERT modeline ince ayar yapmak gerekmektedir.

4.3.1 BERT modeline ince ayar yapılması

BERT'in ince ayar (fine-tuning) işlemi, modelin ön eğitimi sırasında kazandığı dil anlayışını koruyarak etiketlenmiş siber güvenlik veriseti üzerinde yeniden eğitilmesini içermektedir. Adlandırılmış varlık tanıma için ince ayar yapılmış BERT modeli, BERT mimarisindeki birkaç önemli katman ve parametreyi kullanmaktadır. Gömme katmanı (embeddings layer), 28.996 kelimelik bir sözlük boyutuna ve 768 gizli boyuta sahip kelime gömmelerini, 512 belirtece kadar olan diziler için konum gömmelerini ve cümle segmentlerini ayırt etmek için token türü gömmelerini içermektedir. Kodlayıcı (encoder), her biri boyutu 768 olan gizli çok başlı öz-dikkat(self-attention) mekanizmalarını ve ileri beslemeli katmanlar için 3072 ara boyutuna sahip olan 12 transformer katmanından (BertLayer) oluşmaktadır. Düzenleme (dropout) regularizasyonu, öz-dikkat katmanında 0,4 oranında, gömme katmanları ve çıktı katmanları gibi ağın diğer bölümlerinde ise 0,1 oranında uygulanmaktadır. Son olarak, modelin üstüne 768 giriş özelliği ve NER görevindeki varlık etiketlerine karşılık gelen 10 çıktı sınıfından oluşan doğrusal bir

sınıflandırma başlığı eklenmiştir. Bu sınıflardan biri, dizileri sabit uzunlukta tutmak için kullanılan dolgu (PAD-padding) etiketine karşılık gelmektedir.

İnce ayar süreci, modelin ağırlıklarını alanına özgü görevlerdeki performansını optimize edecek şekilde ayarlarlanmıştır. Adlandırılmış varlık tanıma için BERT'in ince ayarının yapılmasındaki kritik zorluklardan biri, NER görevlerindeki sınıf dengesizliğini ele almaktır. NER görevlerindeki çoğu versetinde, herhangi bir varlığa ait olmayan belirteçleri temsil eden "O" sınıfı, çoğunlukla diğer sınıflara göre daha fazla bulunmaktadır. Bu da veri dengesizliğine yol açmaktadır. Bu dengesizliği gidermek için, eğitim sırasında kayıp fonksiyonu olarak focal loss [146] kullanılmıştır. Ayrıca basit dengeleme kullanılarak etiket dağılımında denge oluşturulmaya çalışılmıştır. Focal loss, standart çapraz entropi kaybının formülizasyonuna modülasyon faktörü ekleyerek sınıflandırılması daha zor örneklere odaklanmaktadır.

Hiperparametreler, eğitim süreci sırasında modelin çeşitli yönlerini kontrol eden ve modelin performansını etkileyen elle ayarlanabilen değerlerdir. Bu çalışmada kullanılan hiperparametreler sonuç bölümünde paylaşılmıştır. Aşağıda, eğitim modülünde BERT modeli için kullanılan hiperparametrelerin kısa açıklamaları yer verilmiştir.

Maksimum uzunluk (maxLength): Bir girdi dizesinin (örneğin bir cümlenin) olabileceği maksimum uzunluğu belirtir. Bu parametre, modelin girdi için işleyeceği belirteç sayısını belirtmektedir. Daha büyük bir uzunluk ayarı modelin daha fazla bağlam yakalamasını sağlar, ancak hesaplama maliyetini artırmaktadır. Öte yandan, daha küçük bir uzunluk ayarı ise bilgi kaybına yol açabilmektedir.

Yığın boyutu (batchSize): Yığın boyutu parametresi, modelin her eğitim adımı sırasında bir seferde öğrendiği örnek sayısını belirtmektedir. Yüksek yığın boyutu değerine sahip bir model daha istikrarlı bir öğrenme süreci gerçekleştirir, ancak aynı zamanda daha fazla bellek kullanmaktadır. Öte yandan, bu parametrenin düşük bir değere ayarlanması bellek kullanımını azaltabilir ancak modelin öğrenme sürecini daha kararsız hale getirmektedir.

Test Boyutu (testSize): Test için kullanılacak veri setinin oranını belirtmektedir. Örneğin, 0,2 değeri verisetinin %20'sinin test için kullanılacağını göstermektedir.

Gizli Katman Seyreltmesi (hiddenDropout): Her eğitim adımında gizli katmanlardaki nöronları belirli bir olasılıkla geçici olarak devre dışı bırakarak modelin eğitim verilerine aşırı uyum sağlamasını önleyen parametredir.

Dikkat Seyreltmesi (attDropout): Her eğitim adımında transformer katmanlardaki dikkat mekanizmalarını belirli bir olasılıkla geçici olarak devre dışı bırakarak modelin eğitim verilerine aşırı uyum sağlamasını önleyen parametredir.

Öğrenme Oranı (LearningRate): Bu, modelin parametrelerinin her eğitim adımında ne kadar güncelleneceğini kontrol eden parametre ayarıdır. Bu oran daha düşük ayarlanırsa, model daha yavaş ancak daha istikrarlı bir öğrenme gerçekleştirmektedir. Öte yandan, bu oranın daha yüksek ayarlanması, eğitim sürecini hızlandırırken modelin optimum ağırlıkları aşma riskini artırmaktadır.

Adam Epsilon (epsilon): Sıfıra bölünmeyi önlemek için Adam optimize edicide paydaya eklenen küçük bir sabittir. Bu, özellikle gradyanlar seyrek veya küçük olduğunda güncellemeleri dengelemektedir.

Dönem (epochs): Bu parametre, tüm eğitim veri setinin model tarafından işlenme sayısını belirtir. Dönem sayısını artırmak, modelin veri kümesinden daha fazla bilgi edinmesine olanak tanır, ancak aynı zamanda eğitim verilerine aşırı uyum riskini de artırabilmektedir.

Maksimum Gradyan Normu (maxGradNorm): Geriye yayılma sırasında gradyanların büyüklüğünü sınırlayan parametredir. Gradyan kırpma, gradyanların patlamasını önler ve özellikle BERT gibi derin modellerde kararlı eğitimi sağlamaktadır.

Isınma Oranı (warmupRatio): Öğrenme oranını sıfırdan maksimum değere kademeli olarak artırmak için kullanılan eğitim adımlarının oranını belirtir. Bu parametre, eğitim sürecinin başında modelin kararsız hale gelmesini önler ve öğrenme oranının stabil bir şekilde hedeflenen değere ulaşmasını sağlar.

Ağırlık Sönümleme (weightDecay): Ağırlık sönümleme parametresi, ağırlıkların çok büyük değerler almasını önleyerek ve modelin yeni durumlara daha iyi uyum sağlayabilmesi için eğitimin her adımında ağırlıkları kademeli olarak azaltarak modelin aşırı uyum sağlamasını önleyen parametredir.

Erken Durdurma (earlyStoppingPatience): Modelin doğrulama performansı değerlendirilirken belirtilen dönem sayısı boyunca herhangi bir gelişme olmazsa eğitim sürecini durduran bir parametredir. Bu parametre, kaynakların verimli kullanılmasını sağlamakta ve modelin aşırı uydurulmasını önlemektedir.

Gradyan Biriktirme Adımları (gradientAccumulationSteps): Büyük verisetleriyle eğitim yaparken yığın boyutunu (batch size) etkili bir şekilde artırmak için kullanılan bir

parametredir. Küçük yığın (mini-batch) gradyanları birkaç adım boyunca biriktirilir ve belirli bir adım sayısına ulaşıldığında tek seferde ağırlık güncellemesi yapılır. Bu yöntem, özellikle bellek sınırlamalarını aşarak daha büyük toplu işlem boyutlarıyla eğitimi mümkün kılmaktadır.

Focal Kaybı - Gama (gamma): Focal kaybı formülündeki odaklanma faktörünü kontrol eden değişkendir. Daha yüksek bir gamma değeri, iyi sınıflandırılmış örnekler için kaybı azaltarak modelin sınıflandırılması zor sınıflara daha fazla odaklanmasını sağlamaktadır.

Focal Kaybı - Alfa (alpha): Farklı sınıfların önemini dengeleyen parametredir. Daha yüksek bir alfa değeri azınlık sınıflarını vurgulayarak eğitim sırasında sınıf dengesizliğinin giderilmesine yardımcı olmaktadır.

4.3.2 Eğitim modülü sonuçları

Modelin performansını değerlendirmek için kesinlik (precision), duyarlılık (recall) ve F1-Skoru gibi temel metrikler kullanılmıştır. Bu metrikler, makine öğrenmesi ve doğal dil işleme çalışmalarında yaygın olarak kullanılmaktadır. Kesinlik, bir modelin doğru tahmin ettiği pozitif örneklerin toplam tahmin edilen pozitif örneklere oranı şeklinde Denklem 4.1'de şu şekilde tanımlanır:

$$kesinlik = \frac{doğru\ tahmin\ sayısı}{doğru\ tahmin\ sayısı + yanlış\ tahmin\ sayısı} \tag{4.1}$$

Duyarlılık ise Denklem 4.2'de ifade edildiği gibi verideki pozitif örneklerden model tarafından doğru tahmin edilenlerin oranını ifade etmektedir.

$$duyarlılık = \frac{doğru\ tahmin\ sayısı}{toplam\ gerçek\ varlık\ sayısı} \tag{4.2}$$

F1-Skoru ise kesinlik ile duyarlılık arasındaki dengeyi ölçer ve bu iki değerin harmonik ortalaması ile hesaplanmaktadır. Modelin genel performansını değerlendirmek için kullanılır ve Denklem 4.3 ile ifade edilir.

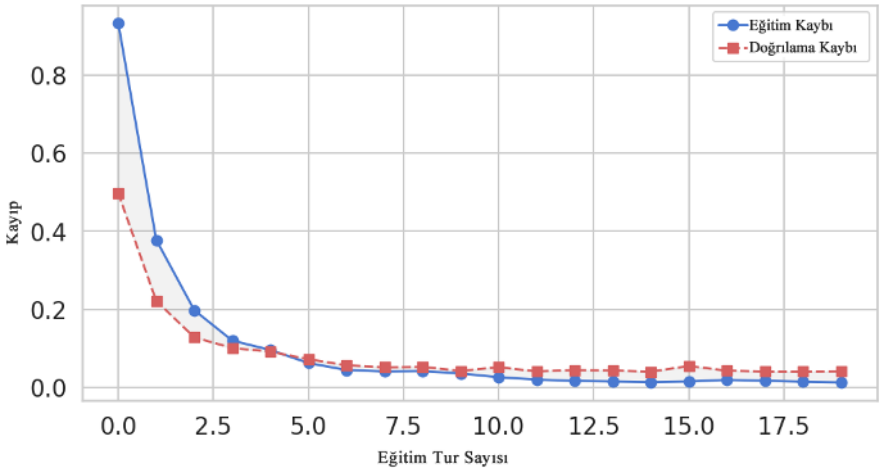
$$f1\ skoru = 2 \times \frac{kesinlik \times duyarlılık}{kesinlik + duyarlılık} \tag{4.3}$$

Eğitim modülü ilgili tüm kodlar ve sonuçlar, verimli bir eğitim ve değerlendirmesi için GPU kaynaklarından yararlanılarak Google Colab üzerinde gerçekleştirilmiştir. Çizelge 4.3'te ince ayar sırasında kullanılan hiperparametreler ve değerleri listelenmektedir.

Çizelge 4.3 : BERT modelinin ince ayar görevinde kullanılan parametreler.

Hiperparametreler	Değer
Maksimum uzunluk (maxLength)	128
Yığın boyutu (batchSize)	32
Test Boyutu (testSize)	0,2
Gizli Katman Seyreltmesi (hiddenDropout)	0,1
Dikkat Seyreltmesi (attDropout)	0,4
Öğrenme Oranı (LearningRate)	3e-5
Adam Epsilon (epsilon)	1e-8
Dönem (epochs)	20
Maksimum Gradyan Normu (maxGradNorm)	1,0
Isınma Oranı (warmupRatio)	0,1
Ağırlık Sönümleme (weightDecay)	0,01
Erken Durdurma (earlyStoppingPatience)	5
Gradyan Biriktirme (gradientAccumulationSteps)	3
Focal Kaybı - Gama (gamma)	2,0
Focal Kaybı - Alfa (alpha)	0,25

Şekil 4.3'te gösterilen öğrenme eğrisi, eğitim ve doğrulama kayıplarının dönemler boyunca ilerleyişini göstermektedir. Eğitim kaybı ilk dönemler sırasında hızla azalmaktadır ve bu da modelin eğitim verilerinin altında yatan örüntüleri etkili bir şekilde öğrendiğini göstermektedir. Benzer şekilde, doğrulama kaybı da önemli ölçüde azalarak 10. dönem civarında sabitlenmektedir ve bu da modelin görülmeyen verilere iyi genelleme yapabileceğini göstermektedir. Eğitim ve doğrulama kaybı eğrileri arasındaki nispeten küçük boşluk, 13. dönemde erken durdurma uygulamasıyla desteklenen minimum aşırı uyumu vurgulamaktadır.



Şekil 4.3 : Model öğrenme sürecinde eğitim ve doğrulama kaybı değişim grafiği.

Eğitim sürecinde, Çizelge 4.4'te görüldüğü üzere 13. dönem sonunda 0,9897'ye ulaşan doğrulama doğruluğu ile 0,9357'lik en yüksek doğrulama F1-skoruna ulaşılmıştır.

Çizelge 4.4 : Dönem bazında model performans tablosu.

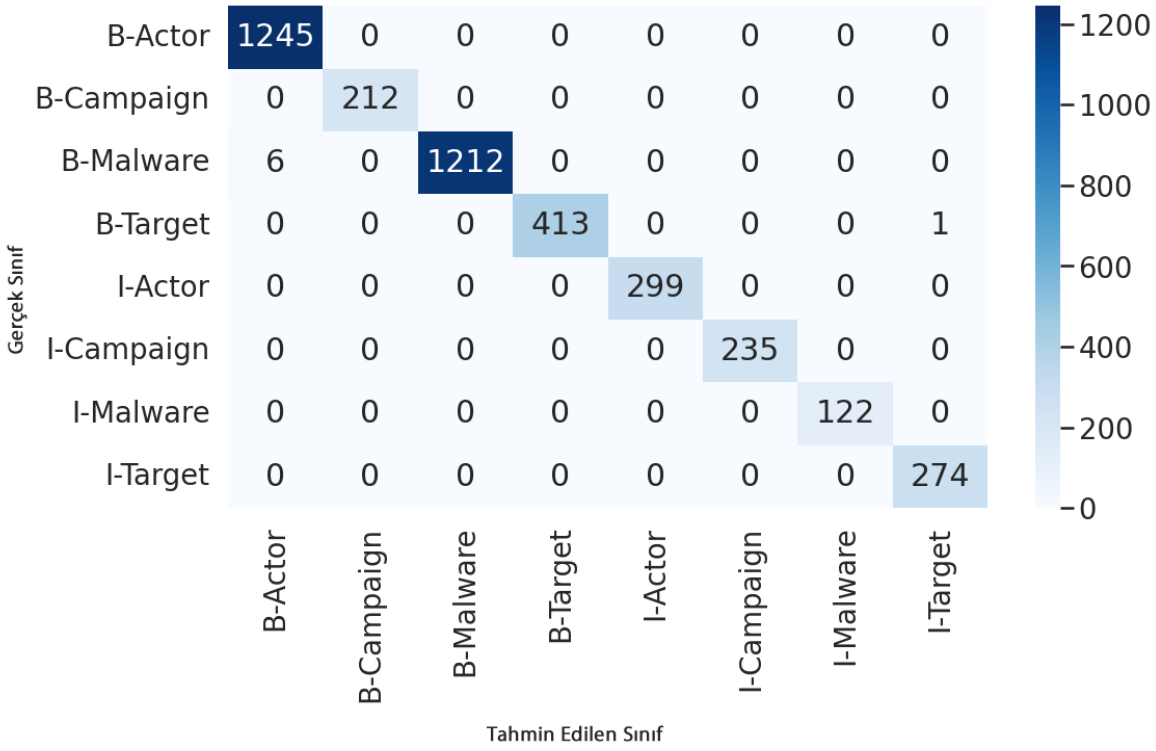
Dönem	Eğitim Kaybı	Doğrulama Kaybı	Doğrulama Doğ.	F1 Skoru
1	0.8944	0.4358	0.9266	0.5982
2	0.3599	0.1967	0.9535	0.7316
3	0.1886	0.1301	0.9590	0.7969
4	0.1189	0.1046	0.9647	0.8166
5	0.0960	0.0968	0.9686	0.8385
6	0.0621	0.0674	0.9787	0.8800
7	0.0443	0.0689	0.9783	0.8779
8	0.0410	0.0611	0.9811	0.8896
9	0.0406	0.0617	0.9823	0.9038
10	0.0346	0.0604	0.9830	0.9006
11	0.0261	0.0510	0.9852	0.9152
12	0.0208	0.0487	0.9868	0.9238
13	**0.0174**	**0.0430**	**0.9897**	**0.9357**
14	0.0152	0.0443	0.9892	0.9333
15	0.0142	0.0425	0.9895	0.9351
16	0.0164	0.0494	0.9866	0.9212
17	0.0204	0.0595	0.9861	0.9227
18	0.0161	0.0505	0.9886	0.9332

Çizelge 4.5, modelin varlık türlerindeki performansına ilişkin değerlendirme metriklerini göstermektedir. Tehdit aktörleri ve saldırı kampanyaları için %97 ve kötü amaçlı yazılımlar için %98 gibi çoğu sınıf için yüksek F1 puanları modelin etkinliğini göstermektedir. Bununla birlikte, hedef sınıfı, öncelikle daha düşük hassasiyet nedeniyle %83 gibi nispeten daha düşük bir F1 puanı sergilemiştir. Bu durum, bu sınıfın diğerlerinden ayırt edilmesinde potansiyel bir zorluk olduğunu göstermektedir ve bu sınıf için daha fazla eğitim verisi gerektirmektedir.

Çizelge 4.5 : Adlandırılmış varlık tanıma için test seti performans ölçütleri.

Varlık	Kesinlik (%)	Duyarlılık (%)	F1-Skoru (%)	Örnek Sayısı
Actor	96	99	97	1245
Campaign	95	100	97	212
Malware	96	99	98	1218
Target	73	95	83	414
Micro Avg	92	99	95	3089
Macro Avg	90	98	94	3089
Weighted Avg	93	99	96	3089

Şekil 4.4'teki karışıklık matrisi, çoğu tahminin doğru olarak gerçekleştiği ve çok az hatanın gözlemlendiği modelin güçlü performansının bir göstergesidir. B-Malware için altı ve B-Target için bir olan minimum yanlış sınıflandırma, Çizelge 4.5'teki yüksek F1 puanlarıyla tutarlı olarak modelin sağlamlığını yansıtmaktadır. Bu da modelin siber güvenlik varlıklarını etkin bir şekilde tanıma konusundaki güvenilirliğini pekiştirmektedir.



Şekil 4.4 : Varlık düzeyinde karışıklık matrisi.

Ayrıntılı değerlendirme ölçütleriyle birlikte öğrenme eğrisi, ince ayarlı BERT modelinin siber güvenliğe özgü varlıkları yüksek doğruluk ve genellenebilirlikle başarılı bir şekilde yakaladığını göstermektedir. Bu sonuçlar, eğitim modülünün yapılandırılmamış tehdit istihbaratını daha sonraki analizler için yapılandırılmış verilere dönüştürmedeki etkinliğini doğrulamaktadır.

4.4 Bilgi İnşa Modülü

Bilgi Oluşturma Modülü, varlıkları bilgi tabanında, alana özgü ontoloji ve ince ayarlı adlandırılmış varlık tanıma modeli kullanarak yapılandırılmış bir biçimde depolamaktadır. Modül, varlıklar arasındaki ilişkileri belirlenen ontoloji kurallarına göre tanımlamakta ve tehdit istihbarat verilerini depolamak ve sorgulamak için temel oluşturan üçlüler

üretmektedir. Bu modül üç bileşenden oluşur. İlk olarak, işlenmiş tehdit raporlarından siber güvenlikle ilgili varlıkları çıkarma süreci, eğitilmiş BERT modeli kullanılarak gerçekleştirilmektedir. Ardından, varlıklar arasındaki ilişkiler tanımlanmış ontolojiye göre kurulmakta ve son olarak varlıklar ile ilişkiler üçlüler olarak gösterilmektedir. Bu üçlüler, tehdit istihbarat bilgilerinin sorgulanmasını ve analiz edilmesini sağlayan yapılandırılmış bir bilgi tabanında depolanmaktadır.

4.4.1 İnce ayarlı model kullanılarak varlık çıkarımı

Bilgi oluşturma sürecinin ilk adımı, eğitim modülünde geliştirilen ince ayarlı NER modeli kullanılarak önceden işlenmiş tehdit raporlarından varlıkları çıkarmayı içermektedir. Veri modülündeki ön işleme adımlarıyla hazırlanan bu tehdit raporları, anlamlı varlıkları belirlemek için model tarafından işlenmektedir. Model, bilgi tabanının temelini oluşturan aşağıdaki temel varlık türlerini ham metinlerden çıkarmaktadır:

Tehdit Aktörü(ThreatActor): Kötü amaçlı faaliyetlerde bulunan bireyleri, grupları veya kuruluşları temsil eder (örn. APT28).

Kötü Amaçlı Yazılım(Malware): Fidye yazılımı, truva atları veya casus yazılım (örn. Emotet) gibi saldırılarda kullanılan kötü amaçlı yazılımları ifade eder.

Saldırı Kampanyaları(Campaign): Tehdit aktörlerinin belirli hedeflere ulaşmak için organize çabalarını ifade etmektedir (örn. Aurora Operasyonu).

Saldırı Hedefleri(Target): Saldırı kampanyaları veya kötü amaçlı yazılımlar tarafından hedeflenen sektörleri, kuruluşları veya sistemleri (örneğin, finansal kuruluşlar) tanımlamaktadır.

Çalışmada model ile varlık çıkarımı yapılırken, varlıklar arasındaki ilişkiler de önceden tanımlanmış bir şemaya göre anlamsal bağlantıları tanımlayan ve doğrulayan aşağıda açıklanan ontoloji kullanılarak kurulmuştur.

4.4.2 Ontoloji tabanlı ilişki kurulumu

Bu çalışmada, siber güvenlik analizinin belirli gereksinimlerini ele alan bir ontoloji tasarlamak için birleşik siber güvenlik ontolojisinden (UCO) yararlanılmıştır. Bu ontoloji, belirli bir kötü amaçlı yazılımın bir tehdit aktörü tarafından nasıl kullanıldığı veya bir saldırı kampanyasının belirli bir sektörü nasıl hedef aldığı gibi varlıklar arasındaki etkileşimleri modellemek için standartlaştırılmış bir şema sağlamaktadır. Çizelge 4.6, ontolojide

tanımlanan varlık türlerine ve aralarındaki ilişkilere genel bir bakış sağlamaktadır. Söz konusu varlıklar arasındaki ilişkilerin örnek bir çizge gösterimi Şekil 4.5'te gösterilmiştir. Kırmızı düğümler tehdit aktörlerini, mavi düğümler ise kötü amaçlı yazılımları göstermektedir, sarı düğümler hedefleri, yeşil düğümler ise kampanyaları göstermektedir.

Çizelge 4.6 : Siber güvenlik ontolojisindeki varlıklar ve ilişkiler.

Kaynak Varlık	İlişki	Hedef Varlık
ThreatActor	hasAssociatedCampaign	Campaign
ThreatActor	hasAssociatedMalware	Malware
ThreatActor	hasTargetedField	Target
ThreatActor	hasAlias	ThreatActor
Malware	isUsedBy	ThreatActor
Campaign	isLaunchedBy	ThreatActor
Campaign	usesMalware	Malware
Campaign	hasTargetedField	Target

Bu ilişkiler, ThreatActors, Malware, Campaigns ve Targets gibi varlıkları ilişkilendirerek düzenli bir bilgi tabanı yapısı sağlamaktadır. Örneğin, *hasAssociatedCampaign* ilişkisi bir ThreatActor'ı katıldığı belirli kampanyalarla ilişkilendirirken, *hasAssociatedMalware* ilişkisi bir ThreatActor'ı kullandığı kötü amaçlı yazılımla ilişkilendirmektedir. Benzer şekilde, *hasTargetedField* ilişkisi bir ThreatActor'ı hedeflediği etki alanlarıyla ilişkilendirirken ve *hasAlias* ilişkisi bir ThreatActor'ın takma adlarını tanımlamaktadır.

Şekil 4.5 : Siber tehdit ilişkilerinin çizge gösterimi.

Campaigns ve Malware'in rollerini daha ayrıntılı analiz etmek için ilişkiler diğer varlıklarla tanımlanmaktadır. Örneğin, *isUsedBy* ilişkisi bir kötü amaçlı yazılımı onu kullanan tehdit aktörüne bağlarken, *isLaunchedBy* ilişkisi bir kampanyayı onu başlatan tehdit aktörüne bağlamaktadır. Kampanyalar ise *usesMalware* ilişkisiyle tanımlanmaktadır. Bu ilişki, bir kampanyanın hangi kötü amaçlı yazılımı kullandığını belirtirmektedir.

4.4.3 Üçlülerin temsili

Siber tehdit verilerinden siber varlıklar çıkarılıp aralarındaki ilişkiler kurulduktan sonra veriler Neo4j çizge veri tabanına üçlüler(triple) halinde aktarılmaktadır. Bu üçlüler, çizge analizi modülünde tehdit aktörleri, kötü amaçlı yazılımlar, kampanyalar ve hedefler arasındaki ilişkilerin ayrıntılı ve görsel olarak incelenmesine olanak tanımaktadır. Neo4j'in çizge görselleştirme yeteneklerinden yararlanılarak önerilen yaklaşım, kötü amaçlı yazılım kullanımındaki kalıpları keşfetme ve tehdit aktörlerinin potansiyel hedeflerini tahmin etme gibi analizleri mümkün kılmıştır.

4.5 Çizge Görselleştirme ve Analiz Modülü

Çizge görselleştirme ve analiz modülü, önceki modülde oluşturulan RDF üçlülerini depolamak ve analiz etmek için kullanılmaktadır. Çıkarılan varlıkları ve ilişkilerini tanımlayan bu üçlüler, Neo4j kullanılarak bir çizge veri tabanına aktarılmaktadır. Neo4j'in çizge depolama ve sorgulama yeteneklerinden yararlanılarak yapılandırılmış veriler, gelişmiş analitik görevleri destekleyen esnek bir biçime dönüştürülmüştür. Üçlüler çizge veri tabanına aktarıldıktan sonra Neo4j, ThreatActors, Campaigns, Malware, Target ve bu varlıkların arasında olan ilişkilerin verimli bir şekilde sorgulanmasını ve görselleştirilmesini sağlamaktadır. Analistler, sıklıkla hedef alınan sektörleri belirlemek, belirli tehdit aktörlerinin operasyonel sahasını analiz etmek veya kötü amaçlı yazılım kullanım stratejilerindeki örüntüleri keşfetmek gibi değerli içgörüleri ortaya çıkarmak için bu bağlantıları incelemektedirler. Bu yapılandırılmış ve görsel temsil, yalnızca karmaşık ilişkilerin keşfedilmesini kolaylaştırmakla kalmamakla beraber aynı zamanda temel örüntülerin ve eğilimlerin daha hızlı ve daha doğru bir şekilde belirlenmesini sağlayarak tehdit istihbaratında karar alma süreçlerini de desteklemektedir.

Bu çalışmada, tehdit aktörleri, zararlı yazılımlar, kampanyalar ve hedefler gibi varlıkları çıkarmak için toplam 540 siber tehdit istihbarat belgesi analiz edilmiştir. Bu varlıklar, çıkarılan verilerin temsilini standartlaştırmak için tasarlanmış özel bir siber tehdit

ontolojisine dayalı olarak yapılandırılmış ve RDF dosyalarında saklanmıştır. Her belge önce ayrı bir RDF dosyasına dönüştürülmüş ve ardından tek bir veri kümesinde birleştirilmiştir. Birleştirilmiş RDF verileri, belgelerden çıkarılan varlıklar arasındaki ilişkilerin ve özelliklerin sorgulanmasını ve çizge tabanlı analizini sağlamak için Neo4j çizge veri tabanına aktarılmıştır.

Bu bölümde, çizge veri tabanının genel yapısı ve içeriği incelenmiş; düğüm tipleri, ilişki türleri ve toplam sayılar üzerine temel tanımlayıcı analizler gerçekleştirilmiştir. Bu analizler, veri modelinin anlaşılması ve ileri düzey analizlere temel oluşturulması amacıyla yapılmıştır. Çizge yapısı, toplamda 3924 düğümden ve 5917 ilişkiden oluşmaktdır. Tanımlanan ilişkiler Çizelge 4.6'da ve düğümlere ait sayılar Çizelge 4.7'de gösterilmiştir. Bu yapılandırılmış temsil, siber tehdit varlıkları arasındaki ilişkilerin ve etkileşimlerin verimli bir şekilde sorgulanmasına ve gelişmiş analizine olanak tanıyarak tehdit aktörü davranışları, kötü amaçlı yazılım ilişkileri ve hedeflenen alanlar hakkında değerli bilgiler sağlamaktadır.

Çizelge 4.7 : Çizge veri tabanındaki düğümler ve sayıları.

Düğüm	Sayı
ns0__ThreatActor	416
ns0__Malware	3,049
ns0__Target	378
ns0__Campaign	81

Belirtilen sayılar, düğümler ile arasındaki bağlantıların yoğunluğunu ve modelin karmaşıklığını ortaya koymaktadır. Genel olarak, yapılan tanımlayıcı istatistikler sonucunda, çizge veri tabanının geniş ve zengin bir veri modeline sahip olduğu, toplam düğüm ve ilişki sayıları, modelin büyük ölçekli analizler ve çıkarımlar için uygun olduğunu göstermektedir. Ayrıca, düğüm ve ilişki türleri arasındaki çeşitlilik te farklı siber güvenlik perspektiflerinden analizler yapılabilmesine olanak tanımaktadır.

4.5.1 Temel analizler

Bu bölümde, veri tabanında yer alan zararlı yazılımlar, tehdit aktörleri ve hedef alanlar arasındaki ilişkileri anlamaya yönelik temel analizler gerçekleştirilmiştir. Bu analizler, zararlı yazılımların hangi tehdit aktörleri tarafından kullanıldığını, hedeflerin yoğunlaştığı alanları ve bu zararlı yazılımların saldırdığı hedeflerin genel dağılımını ortaya koymaktadır. Analiz sonuçları, siber tehdit ekosistemindeki etkileşimlerin kapsamını ve

stratejik tercihlerini anlamak için önemli bir temel sunmaktadır. Her analiz, ilgili sorgular ve tablolarla detaylandırılarak aşağıda sunulmuştur.

Tehdit aktörlerinin kullandığı zararlı yazılım sayısının analizi, siber tehdit gruplarının operasyonel kapasitelerini ve karmaşıklık düzeyini anlamak açısından önemli içgörüler sağlamaktadır. Gerçekleştirilen çizge veri tabanı sorgusu, her bir tehdit aktörünün kullandığı benzersiz zararlı yazılım sayısını ortaya koymaktadır. Çizelge 4.8'de sunulan analiz sonuçlarına göre, düğüm kimliği 5141 olan tehdit aktörünün 347 farklı zararlı yazılım kullanarak en geniş zararlı yazılım çeşitliliğine sahip aktör olarak öne çıkmaktadır. Buna karşılık, düğüm kimliği 2398 olan ikinci sıradaki tehdit aktörünün 55 zararlı yazılım kullanması, düğüm kimliği 5141 olan aktörün diğer aktörlere kıyasla ne kadar ön planda olduğunu ortaya koymaktadır. İlk 10'da yer alan diğer aktörlerin kullandıkları zararlı yazılım sayıları 16 ile 34 arasında değiştiği gözlemlenmiştir.

Çizelge 4.8 : Tehdit aktörlerinin zararlı yazılım kullanım dağılımı.

Tehdit Aktörü	Zararlı Yazılım Sayısı
5141	347
2398	55
261	34
2261	25
687	21
254	20
1248	19
17	18
933	17
4515	16

Bu veriler ışığında önemli çıkarımlar yapılabilir. Düğüm kimliği 5141 olan aktörün bu denli geniş bir zararlı yazılım portföyüne sahip olması, grubun yüksek teknik kapasitesini ve kaynak zenginliğini göstermektedir. Grubun bu kadar çeşitli zararlı yazılım kullanması, farklı hedef sistemlere ve güvenlik önlemlerine karşı geniş bir saldırı yelpazesine sahip olduğunu da ortaya koymaktadır. Gerçekleştirilen bir diğer analizde, her bir zararlı yazılımın kaç farklı tehdit aktörü tarafından kullanıldığının belirlenmesi hedeflenmiştir. Bu kapsamda, çizge veri tabanında Malware düğümleri ile bu düğümlere bağlı olan ThreatActor düğümleri arasındaki *isUsedBy* ilişkisi kullanılarak, zararlı yazılımlar ile bu yazılımları kullanan aktörler arasındaki bağlantılar incelenmiştir. Analiz sürecinde, her bir zararlı yazılım düğümüne bağlı benzersiz tehdit aktörü düğümleri sayılmış ve bu sayılar sıralanarak en çok kullanılan zararlı yazılımlar belirlenmiş, elde edilen sonuçlar Çizelge 4.9'da gösterilmiştir.

Bu yöntem, zararlı yazılımların siber tehdit ekosistemindeki yaygınlığını ve farklı tehdit aktörleri tarafından kullanım oranlarını ortaya koymaktadır.

Çizelge 4.9 : En çok kullanılan zararlı yazılımlar ve aktör sayıları.

Zararlı Yazılım	Kullanan Aktör Sayısı
617	5
1570	3
318	2
1465	2
1140	2
2005	2
2276	1
4643	1
5410	1
4309	1

Analiz sonuçları, düğüm kimliği 617 olan zararlı yazılımının beş farklı tehdit aktörü tarafından kullanıldığını ve siber tehdit ekosisteminde en yaygın kullanılan zararlı yazılımlardan biri olduğunu göstermektedir. Diğer zararlı yazılımlar ise daha az sayıda tehdit aktörü tarafından kullanılmakla birlikte, özellikle düğüm kimlikleri 318, 1465, 1140 ve 2005 olan zararlı yazılımlar iki farklı tehdir aktörü tarafından kullanıldığı tespit edilmiştir. Bu durum, bazı zararlı yazılımların çok yönlü kullanım potansiyeline sahip olduğunu ve farklı tehdit aktörleri tarafından tercih edildiğini göstermektedir. Öte yandan, düğüm kimlikleri 2276, 4643, 5410 ve 4309 olan zararlı yazılımlar ise yalnızca bir tehdit aktörü tarafından kullanıldığı belirlenmiştir. Gerçekleştirilen bir diğer analizde, tehdit aktörlerinin hedef aldıkları alanların dağılımları incelenmiştir. Bu kapsamda, çizge veri tabanında ThreatActor düğümleri ile bu düğümlere bağlı olan Target düğümleri arasındaki *hasTargetedField* ilişkisi kullanılarak, tehdit aktörlerinin hangi alanlara yoğunlaştığı analiz edilmiştir. Söz konusu analiz, siber tehdit aktörlerinin hedef tercihlerini ve saldırı eğilimlerini ortaya koymaktadır. Elde edilen analiz sonuçları Çizelge 4.10'da sunulmuştur. Bu bulgular, özellikle bazı sektörlerin tehdit aktörleri tarafından daha sık hedef alındığını göstermektedir. Sektörel hedeflerin analizi, siber güvenlik politikalarının şekillendirilmesinde ve kaynakların önceliklendirilmesinde önemli bir rol oynamaktadır. Böylelikle, tehdit aktörlerinin operasyonel yönelimleri daha net bir şekilde değerlendirilebilmektedir.

Çizelge 4.10 : En çok hedeflenen alanlar ve hedefleyen aktör sayıları.

Hedef Alan	Hedefleyen Aktör Sayısı
Google	27
Infrastructure	7
Government	5
Banking	5
Government Agencies	4
Banks	4
Political	4
Information	3
Law	3
Data	3

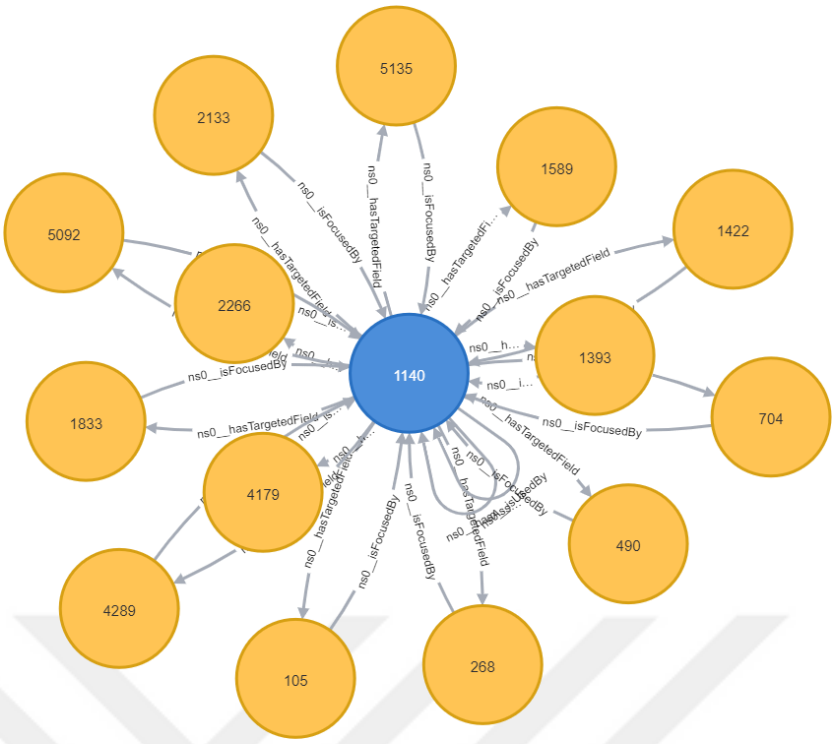
Analiz sonuçları, düğüm kimliği 942 olan hedefin 27 farklı tehdit aktörü tarafından hedef alındığını ve siber saldırılarda en sık tercih edilen hedef olduğunu ortaya koymaktadır. Düğüm kimliği 2266 olan altyapı hedefi, 7 tehdit aktörü tarafından hedeflenmiş, düğüm kimlikleri 1459 (hükümet) ve 315(bankacılık sistemleri) olan hedefler ise 5'er tehdit aktörü tarafından hedef alınmıştır. Ayrıca, düğüm kimlikleri 1893 (kamu kurumları), 4433 (bankalar) ve 4701 (politik hedefler) olan hedeflerin her biri 4 tehdit aktörü tarafından hedeflenmiştir. Düğüm kimlikleri 4733, 5283 ve 817 olan hedefler ise 3'er tehdit aktörü tarafından hedef alınmıştır. Bu dağılım, tehdit aktörlerinin özellikle kritik altyapılar ve finansal kurumlara yönelik saldırılara odaklandığını göstermektedir. Bunun yanı sıra, devlet kurumlarının ve politik hedeflerin de önemli ölçüde saldırı hedefi olduğu görülmektedir.

Zararlı yazılımların hedef aldıkları alanların dağılımını incelemeye yönelik gerçekleştirilen son analizde, her bir kötü amaçlı yazılımın kaç farklı hedef alanına saldırı gerçekleştirdiği belirlenmiştir. Bu analiz, çizge veri tabanında yer alan Malware düğümleri ile bu düğümlere bağlı Target düğümleri arasındaki *hasTargetedField* ilişkisi kullanılarak gerçekleştirilmiştir. Böylece her bir zararlı yazılımın etki alanı genişliği ve potansiyel saldırı kapsamı analiz edilmiştir. Elde edilen bulgular, siber tehditlerin ne derece yaygın ve çok yönlü olduğunu göstermesi açısından önemli veriler sunmaktadır. İlgili sorgu sonucu oluşan sonuçlar Çizelge 4.11'de, düğüm kimliği 1140 olan kötü amaçlı yazılım için oluşturulan örnek çizge Şekil 4.6'da gösterilmektedir. Bu çizge, zararlı yazılımın hedef alanlarla olan ilişkilerini açık biçimde göstermekte ve saldırı örüntülerine dair görsel bir perspektif sağlamaktadır.

Çizelge 4.11 : En çok hedef alanına sahip zararlı yazılımlar.

Zararlı Yazılım	Hedef Alan Sayısı
2168	19
1140	14
617	11
1046	9
4206	7
2230	6
5575	4
2432	4
4441	3
1766	3

Çizelge 4.11'de, yer alan sonuçlara göre, düğüm kimliği 2168 olan zararlı yazılım toplam 19 farklı hedef alanına saldırı gerçekleştirmiştir ve analiz kapsamında en geniş saldırı yelpazesine sahip kötü amaçlı yazılım olarak öne çıkmaktadır. Bunu, 14 farklı hedef alanı ile düğüm kimliği 1140 olan yazılım ve 11 hedef alanıyla düğüm kimliği 617 olan yazılım takip etmektedir. İlk beşte yer alan diğer zararlı yazılımlar ise sırasıyla 9 hedef alanı (1046) ve 7 hedef alanı (4206) ile dikkat çekmektedir. Çizelgede yer alan tüm zararlı yazılımlar en az üç farklı hedefe yönelik saldırı gerçekleştirmiş olup, bu durum söz konusu zararlı yazılımların çok yönlü kullanım potansiyeline sahip olduğunu göstermektedir. Ayrıca, bu analiz yalnızca saldırı yoğunluğunu değil, aynı zamanda zararlı yazılımların hangi alanlara yönelme eğiliminde olduğunu da değerlendirme imkânı sunmaktadır. Düğümler arasındaki ilişkileri görselleştirmek amacıyla oluşturulan örnek çizge, düğüm kimliği 1140 olan zararlı yazılım için Şekil 4.6'da sunulmuştur. Bu çizge, zararlı yazılımın hedef alanlarla olan ilişkilerini açık biçimde göstermekte ve saldırı örüntülerine dair görsel bir perspektif sağlamaktadır. Bu analiz, her bir zararlı yazılımın kaç farklı hedef alanına saldırı gerçekleştirdiğini belirlemek amacıyla yapılmıştır. Çizge veri tabanında Malware ve Target düğümleri arasındaki ilişkiler incelenerek, zararlı yazılımların hedef çeşitliliği tespit edilmiştir. Analiz sırasında her bir zararlı yazılımın bağlantılı olduğu hedef alanları sayılmış ve bu sayıların sıralanmasıyla en geniş etki alanına sahip zararlı yazılımlar belirlenmiştir.



Şekil 4.6 : Düğüm kimliği 1140 olan zararlı yazılımın ve hedeflerinin çizge gösterimi.

Bu dağılım, bazı zararlı yazılımların (2168, 1140 ve 617) çok sayıda hedef alanına saldırı düzenleyebilen çok yönlü tehditler olduğunu gösterirken, diğer zararlı yazılımların (4441, 1766) daha spesifik hedeflere odaklandığını düşündürmektedir. Bu durum, zararlı yazılımların farklı tasarım amaçlarını ve saldırı stratejilerini yansıtmaktadır.

4.5.2 Merkezilik analizleri

Siber tehdit ekosistemindeki aktörler, zararlı yazılımlar ve hedefler arasındaki ilişkilerin karmaşık yapısını daha iyi kavrayabilmek amacıyla merkezilik analizleri gerçekleştirilmiştir. Merkezilik analizleri, ağ içerisindeki varlıkların önem derecelerini ve etkileşim güçlerini ortaya koyarak, kritik aktörlerin ve bağlantıların belirlenmesine olanak sağlamaktadır. Bu analizler sayesinde, tehdit ekosisteminde merkezi konumda bulunan ve ağın genel yapısını şekillendiren unsurların tespit edilmesi mümkün olmaktadır. Bu kapsamda ilk olarak, ağdaki varlıkların etki düzeylerini ve önem sıralamalarını belirlemek üzere PageRank algoritması uygulanmıştır. Ardından, ağdaki bilgi akışını ve bağlantı noktalarını daha ayrıntılı biçimde incelemek amacıyla arasındalık merkeziliği analizi gerçekleştirilmiştir. Son olarak, tehdit aktörlerinin kendi aralarındaki ilişkileri ve hedeflerle olan bağlantılarını daha kapsamlı biçimde ortaya koymak üzere ilişki ve hedef analizleri

yapılmıştır. Bu analizlerin sonuçları, siber tehditlerin önlenmesi ve müdahale stratejilerinin geliştirilmesi açısından önemli bulgular sunmaktadır.

4.5.2.1 PageRank analizi

PageRank [147], Google tarafından web sayfalarının önemini ve sıralamasını belirlemek amacıyla geliştirilen ilk ve en yaygın kullanılan algoritmalardan biridir. Stanford Üniversitesi'nde Sergey Brin ve Lawrence Page tarafından ortaya konulan bu algoritma, ağ yapılarındaki varlıkların önem derecelerini belirlemek için bağlantı analizi yöntemini temel almaktadır. PageRank algoritmasının temel prensibi, bir web sayfasının önemini, o sayfaya yönlendirilmiş bağlantıların sayısına ve bu bağlantıların kalitesine göre değerlendirmektir. Bu bağlamda, her bağlantı bir tür oy olarak kabul edilmektedir ve yüksek PageRank değerine sahip sayfalardan gelen bağlantılar, düşük PageRank değerine sahip sayfalardan gelen bağlantılara kıyasla daha fazla ağırlığa sahip olmaktadır. Böylece, algoritma yalnızca bağlantıların niceliğini değil, aynı zamanda niteliğini de dikkate alarak daha doğru ve güvenilir bir önem sıralaması oluşturmaktadır. Matematiksel olarak ifade edildiğinde, bir sayfanın PageRank değeri, kendisine yönelen tüm bağlantıların ağırlıklı toplamı olarak hesaplanmaktadır ve bu değer, ağ içerisindeki diğer sayfaların değerleriyle etkileşimli olarak belirlenmektedir. Siber güvenlik bağlamında bu algoritma, tehdit aktörlerinin, zararlı yazılımların ve hedeflerin ağ üzerindeki önem derecelerini ve merkeziyetlerini belirlemek için uygulanmıştır. Bu analiz kapsamında, çizge veri tabanında ns0__ThreatActor, ns0__Malware ve ns0__Target türündeki düğümler seçilmiş ve bu düğümler arasındaki ilişkileri temsil eden *isUsedBy* ve *hasTargetedField* ilişkileri kullanılarak kapsamlı bir ağ modeli oluşturulmuştur. Oluşturulan bu ağ modeli üzerinde PageRank algoritması çalıştırılarak, ağ içerisindeki varlıkların önem dereceleri hesaplanmış ve elde edilen sonuçlar önem sırasına göre sıralanmıştır. Analiz sonucunda en yüksek PageRank skoruna sahip varlıklar Çizelge 4.12'de sunulmuştur. Bu sonuçlara göre, tehdit aktörleri arasında düğüm kimliği 5141 olan aktör en yüksek PageRank değerine sahiptir. Bu durum, söz konusu tehdit aktörünün analiz edilen raporlarda Şekil 4.7'de görüldüğü üzere merkezi ve kritik bir konuma sahip olduğunu göstermektedir. Düğüm kimliği 5141 olan tehdit aktörünün yüksek PageRank skoruna sahip olması, bu aktörün siber tehdit ekosisteminde diğer varlıklarla yoğun ve etkili bir etkileşim içerisinde olduğunu ve ağın genel yapısı üzerinde önemli bir etkiye sahip olduğunu ortaya koymaktadır. Bu bulgu, ilgili tehdit aktörünün potansiyel risk seviyesinin yüksek olduğunu ve siber güvenlik açısından öncelikli olarak ele alınması gerektiğini vurgulamaktadır.

Çizelge 4.12 : PageRank analiz sonuçları.

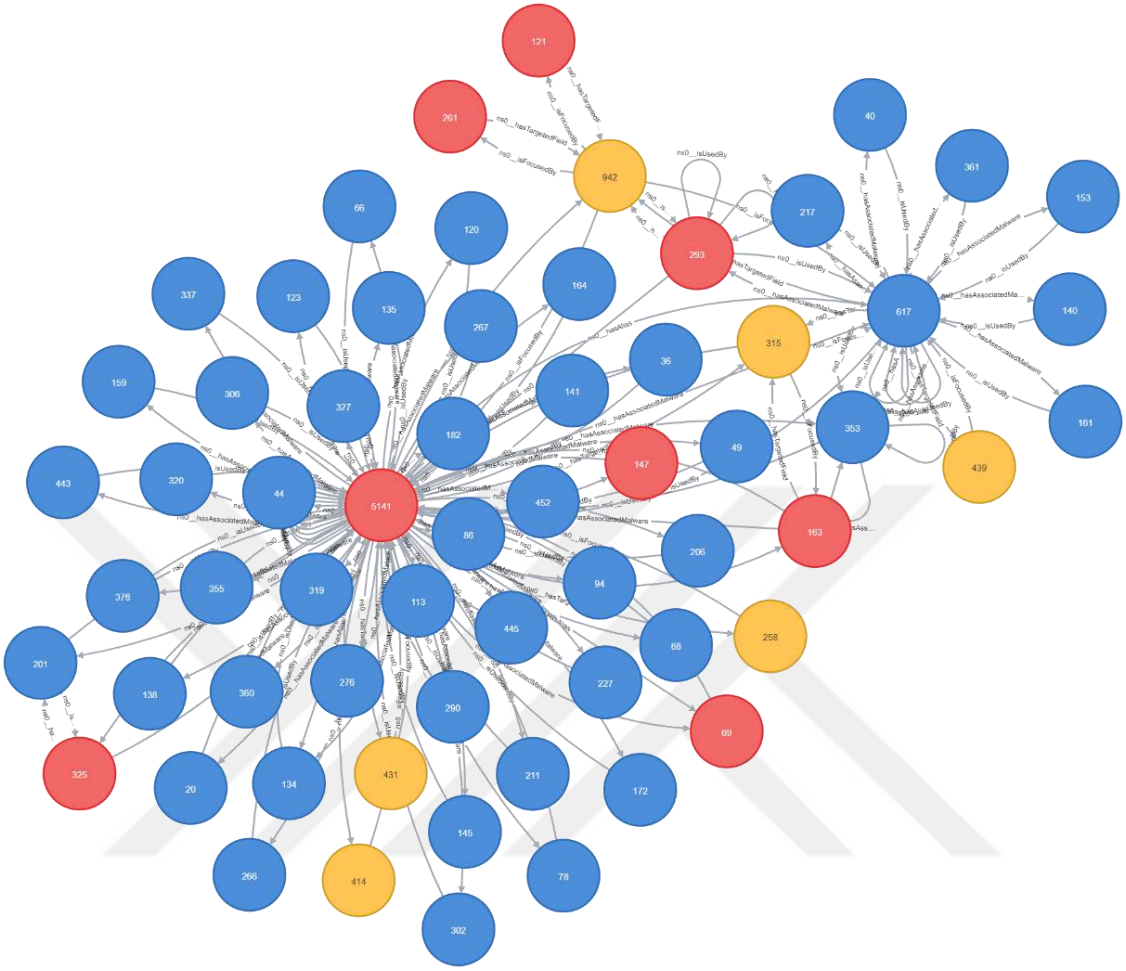
Varlık	Varlık Türü	PageRank Puanı
5141	ThreatActor	47.48703526528706
318	Malware	40.94367740432827
942	Target	30.737362379102464
2005	Malware	25.2273782163075
617	Malware	20.765970624242765
5019	ThreatActor	16.89351018603959
1570	Malware	10.03687849699953
4310	Malware	9.110095624092462
5287	ThreatActor	8.60296731273154
1181	ThreatActor	8.330008061564474
1140	Malware	7.899375172058753
4301	Malware	7.899047188839869
2261	ThreatActor	7.841911916028978
2358	Malware	7.451164220239371
333	ThreatActor	7.256816322613145

Zararlı yazılımlar açısından değerlendirildiğinde, düğüm kimlikleri 318, 2005 ve 617 olan zararlı yazılımlar, elde edilen yüksek PageRank skorları ile dikkat çekmektedir. Özellikle düğüm kimliği 318 olan zararlı yazılımın ikinci en yüksek PageRank skoruna sahip olması, bu zararlı yazılımın siber saldırılarda yaygın biçimde kullanıldığını ve birden fazla tehdit aktörü tarafından sıklıkla tercih edildiğini ortaya koymaktadır. Bu durum, söz konusu zararlı yazılımın tehdit ekosisteminde merkezi bir rol üstlendiğini ve saldırı kampanyalarında kritik bir araç olarak kullanıldığını göstermektedir.

Hedefler kategorisi incelendiğinde ise, düğüm kimliği 942 ile temsil edilen Google platformunun en yüksek PageRank skoruna sahip olduğu görülmektedir. Bu bulgu, Google'ın siber saldırganlar açısından öncelikli ve stratejik bir hedef olduğunu doğrulamaktadır. Ayrıca, bu sonuç büyük teknoloji şirketlerinin ve yaygın kullanılan dijital platformların siber tehdit aktörlerinin ana hedefleri arasında yer aldığını ve saldırganların bu tür platformları hedefleyerek daha geniş çaplı etki yaratmayı amaçladıklarını göstermektedir.

PageRank analizi sonuçları, siber güvenlik stratejilerinin geliştirilmesi ve tehdit aktörlerinin davranış modellerinin daha iyi anlaşılması açısından önemli içgörüler sağlamaktadır. Ağ içerisinde yüksek PageRank skoruna sahip varlıkların belirlenmesi, potansiyel siber tehditlerin erken aşamada tespit edilmesi ve önlenmesi açısından kritik önem taşımaktadır. Bu bağlamda, analiz sonuçları, güvenlik önlemlerinin önceliklendirilmesi, tehditlere karşı müdahale süreçlerinin etkinleştirilmesi ve kaynakların daha verimli kullanılması konusunda güvenlik uzmanlarına yol gösterici nitelikte bilgiler

sunmaktadır. Böylece, siber tehditlere karşı daha önleyici ve etkili savunma stratejilerinin oluşturulmasına katkı sağlanmaktadır.



Şekil 4.7 : Düğüm kimliği 5141 olan tehdit aktörünün PageRank'e dayalı merkezi konumu.

4.5.2.2 Arasındalık merkeziliği analizi

Arasındalık Merkeziliği [115] (Betweenness Centrality), bir ağdaki düğümlerin diğer düğümler arasındaki bilgi akışını ne ölçüde kontrol ettiğini veya etkilediğini ölçen önemli bir merkezilik metriğidir. Bu metrik ilk olarak 1977'de Linton C. Freeman tarafından geliştirilmiştir ve özellikle sosyal ağ analizlerinde yaygın olarak kullanılmakta ve ağ içerisindeki kritik düğümlerin belirlenmesinde önemli rol oynamaktadır. Bir düğümün arasındalık merkeziliği değeri, ağdaki diğer düğümler arasındaki en kısa yolların kaçının o düğüm üzerinden geçtiğini ölçmektedir. Bu bağlamda, yüksek arasındalık merkeziliği değerine sahip düğümler, ağdaki bilgi akışını kontrol eden, farklı gruplar arasında bağlantı sağlayan ve ağın bütünlüğünü koruyan köprü veya aracı konumundaki düğümler olarak

değerlendirilmektedir. Bu düğümlerin ağdan çıkarılması, ağın genel bağ yapısını önemli ölçüde etkileyebilmekte ve ağın parçalanmasına neden olabilmektedir.

Siber güvenlik bağlamında arasındalık merkeziliği analizi, tehdit aktörleri, zararlı yazılımlar ve hedefler arasındaki ilişkilerde kritik rol oynayan varlıkları belirlemek amacıyla uygulanmıştır. Bu analiz kapsamında, çizge veri tabanında bulunan ns0__ThreatActor, ns0__Malware ve ns0__Target türündeki düğümler seçilmiş, bu düğümler arasındaki ilişkileri temsil eden *isUsedBy* ve *hasTargetedField* bağlantıları kullanılarak kapsamlı bir ağ modeli oluşturulmuştur. Oluşturulan ağ modeli üzerinde arasındalık merkeziliği algoritması çalıştırılmış ve elde edilen sonuçlar normalize edilerek önem sırasına göre sıralanmıştır. Analiz sonucunda en yüksek arasındalık merkeziliği puanına sahip varlıklar Çizelge 4.13'te sunulmuştur.

Çizelge 4.13 : Varlık türlerine göre arasındalık merkeziliği analiz sonuçları.

Varlık	Varlık Türü	Arasındalık Merkeziliği Puanı
5141	ThreatActor	1.0
617	Malware	0.3901142392838114
4227	Malware	0.1642686192779392
4301	Malware	0.1503561307294298
2230	Malware	0.14525843771446814
4310	Malware	0.1292375926298783
2398	ThreatActor	0.11130906791489355
353	Malware	0.11094518489504086
1140	Malware	0.0987546873878478
5450	Malware	0.09272272263770384

Siber tehdit aktörleri, zararlı yazılımlar ve hedefler arasındaki ilişkilerin incelendiği arasındalık merkeziliği analizi, ağdaki kritik aracı düğümleri ortaya çıkarmıştır. Bu sonuçlar, siber tehdit ekosistemindeki bilgi ve saldırı akışının yapısı hakkında önemli bulgular sunmaktadır.

Tehdit Aktörlerinin Rolü: Düğüm kimliği 5141 olan tehdit aktörünün en yüksek arasındalık merkeziliği değerine sahip olması, bu aktörün ağda kritik bir köprü görevi üstlendiğini göstermektedir. Bu durum, söz konusu tehdit aktörünün diğer tehdit aktörleri, zararlı yazılımlar ve hedefler arasındaki bağlantıları kontrol eden merkezi bir konumda olduğunu ortaya koymaktadır. Bu aktörün devre dışı bırakılması veya etkisiz hale getirilmesi, ağdaki saldırı ve bilgi akışını önemli ölçüde sekteye uğratabilecek potansiyele sahiptir.

Zararlı Yazılımların Önemi: İlk beş içerisinde dört zararlı yazılımın (617, 4227, 4301 ve 2230) yer alması, bu yazılımların siber saldırılarda önemli bir aracı rol oynadığını göstermektedir. Özellikle düğüm kimliği 617 olan zararlı yazılımın yüksek arasındalık merkeziliği değeri (0.39), farklı tehdit aktörleri tarafından yaygın olarak kullanıldığını ve çeşitli hedeflere yönelik saldırılarda aracı görevi gördüğünü işaret etmektedir. Bu zararlı yazılımların tespiti ve engellenmesi, saldırı zincirlerinin kırılması açısından büyük önem taşımaktadır. Bu analiz sonuçları, siber güvenlik uzmanlarına ve kuruluşlara, tehdit aktörlerinin ve zararlı yazılımların ağ içindeki stratejik konumlarını anlama ve buna göre savunma mekanizmalarını geliştirme konusunda değerli içgörüler sağlayabilmektedir.

4.5.2.3 Tehdit aktörlerinin ilişki ve hedef analizleri

Siber tehdit ekosisteminde faaliyet gösteren tehdit aktörlerinin davranış örüntülerini ve aralarındaki olası bağlantıları anlamak amacıyla, tehdit aktörlerinin ortak hedef seçimlerine dayalı bir ilişki analizi gerçekleştirilmiştir. Bu analiz kapsamında, çizge veri tabanı üzerinde geliştirilen özel sorgular aracılığıyla tehdit aktörlerinin hedef kümeleri incelenmiş ve bu kümeler arasındaki kesişimlerin yoğunluğu tespit belirlenmiştir. Böylece, farklı tehdit aktörlerinin benzer hedeflere yönelik saldırı eğilimlerive ortak ilgi alanları ortaya çıkarılmıştır. Analiz sonucunda elde edilen bulgular Çizelge 4.14'te gösterilmiştir.

Çizelge 4.14 : Benzer hedeflere saldıran tehdit aktörleri analiz sonuçları.

Aktör 1	Aktör 2	Ortak Hedefler	Ortak Hedef Sayısı
5141	2082	banking, organizations, infrastructure, energy	4
5141	2398	bank, Google, banking	3
1125	5141	infrastructure, financial, information	3
2285	51411	payment, infrastructure, banks	3
4515	1581	Google, government agencies	2
5141	463	healthcare, law	2

Analiz sonuçları incelendiğinde, özellikle düğüm kimliği 5141 olan tehdit aktörünün diğer tehdit aktörleriyle önemli ölçüde hedef paylaşımı dikkat çekmektedir. Örneğin, düğüm kimliği 5141 ve 2082 olan tehdit aktörleri arasında dört ortak hedef (bankacılık, organizasyonlar, altyapı ve enerji sektörleri), her iki grubun da kritik altyapı ve finansal sistemlere yönelik benzer saldırı hedefleri seçtiğini göstermektedir. Bu yoğun hedef kesişimi, söz konusu tehdit aktörlerinin hedef seçimlerinde sistematik ve stratejik bir yaklaşım izlediklerini ve potansiyel olarak benzer motivasyonlara veya yöntemlere sahip olduklarını düşündürmektedir. Benzer şekilde, düğüm kimlikleri 2398 olan tehdit aktörü ile 5141 olan tehdit aktörü arasındaki üç ortak hedef (Google, bank, banking), bu iki grubun da

büyük teknoloji şirketleri ve finans sektörüne yönelik sistematik bir ilgi gösterdiğini ortaya çıkarmaktadır. Ayrıca, düğüm kimliği 1125 ve 2285 olan tehdit aktörlerinin, düğüm kimliği 5141 olan tehdit aktörü ile paylaştığı ortak hedefler (altyapı sistemleri, finansal kurumlar ve bilgi sistemleri), bu aktörlerin kritik altyapı ve finansal sistemlere yönelik ortak bir yönelim içinde olduklarını göstermektedir. Bu analizden elde edilen bulgular, tehdit aktörlerinin hedef seçimlerinde belirgin bir sektörel yoğunlaşma olduğunu göstermektedir. Özellikle finansal kurumlar, kritik altyapı sistemleri, büyük teknoloji şirketleri ve önemli organizasyonlar, birden fazla tehdit aktörünün ortak hedefleri arasında yer almaktadır. Bu durum, söz konusu sektörlerin siber tehditlere karşı savunma stratejilerini güçlendirmeleri ve tehdit istihbaratı paylaşımı gibi kurumlar arası işbirliği mekanizmalarını geliştirmeleri gerektiğini vurgulamaktadır.

Sonuç olarak, tehdit aktörleri arasındaki bu hedef benzerlikleri, grupların olası işbirliklerini veya benzer motivasyonlarla hareket ettiklerini düşündürmektedir. Bu analiz sonuçları, siber güvenlik uzmanlarına ve ilgili kuruluşlara, tehdit aktörlerinin davranış modellerini daha iyi anlama, saldırıların önlenmesi için gerekli önlemleri alma ve kurumlar arası koordinasyonu artırma konusunda değerli içgörüler sağlamaktadır. Böylece, siber tehditlere karşı daha etkin ve önleyici savunma stratejilerinin geliştirilmesine katkıda bulunulmaktadır.

4.5.3 Siber tehdit aktörleri arasındaki yol analizleri

Yol analizi (Path Analysis), bir ağ yapısında bulunan düğümler arasındaki bağlantı rotalarını ve bu rotaların özelliklerini inceleyen önemli bir çizge teorisi tekniğidir. Bu analiz yöntemi, ağ içerisindeki düğümler arasında var olan bağlantıların yapısını, yönünü ve derinliğini ortaya koymak amacıyla en kısa yol algoritmaları ve çeşitli yol keşfi yöntemlerini kullanmaktadır. Yol analizi sayesinde, ağdaki varlıklar arasındaki doğrudan ve dolaylı ilişkiler belirlenebilmekte, bu ilişkilerin yoğunluğu ve kritik bağlantı noktaları tespit edilebilmektedir.

Siber güvenlik bağlamında yol analizi teknikleri, tehdit aktörleri, zararlı yazılımlar ve hedefler arasındaki ilişki zincirlerini ortaya çıkarmak ve potansiyel saldırı vektörlerini belirlemek için yaygın olarak kullanılmaktadır. Bu analizler, tehdit aktörlerinin saldırılarını gerçekleştirirken izledikleri olası yolları ve yöntemleri anlamaya yardımcı olmakta, böylece saldırıların erken aşamada tespit edilmesi ve önlenmesi için önemli bilgiler sağlamaktadır. Ayrıca yol analizi, tehdit aktörlerinin kullandığı zararlı yazılımların yayılma rotalarını ve

hedeflere ulaşma süreçlerini ortaya koyarak, saldırı zincirlerinin kırılması ve savunma mekanizmalarının güçlendirilmesi açısından kritik içgörüler sunmaktadır. Bu bağlamda gerçekleştirilen yol analizleri, siber tehdit ekosistemindeki karmaşık ilişkilerin daha iyi anlaşılmasına ve tehditlere karşı daha etkin savunma stratejilerinin geliştirilmesine önemli katkılar sağlamaktadır.

4.5.3.1 Aktörler arası işbirliği yol analizi

Bu analiz kapsamında, tehdit aktörlerinin zararlı yazılım kullanım dağılımını gösteren Çizelge 4.8'deki bulgular temel alınarak, en aktif olarak tespit edilen iki tehdit aktörü arasındaki potansiyel işbirliği yolları incelenmiştir. Bu aktörlerden düğüm kimliği 5141 olan tehdit aktörü toplamda 347 farklı zararlı yazılım kullanımıyla en aktif aktör olarak öne çıkarken, düğüm kimliği 2398 olan tehdit aktörü ise 55 farklı zararlı yazılım kullanımıyla ikinci sırada yer almaktadır. Bu iki tehdit aktörü arasındaki olası bağlantıların ve işbirliği yollarının belirlenmesi amacıyla çizge veri tabanı üzerinde iki farklı Cypher sorgusu çalıştırılmıştır. İlk sorguda, iki aktör arasındaki maksimum dört adımlı tüm olası bağlantı yolları araştırılmıştır. İkinci sorguda ise, ortak hedeflere sahip olan tehdit aktörleri arasındaki en kısa bağlantı yolları tespit edilmiştir. Gerçekleştirilen analiz sonucunda, düğüm kimlikleri 5141 ve 2398 olan tehdit aktörleri arasında Şekil 4.8'de görüldüğü üzere düğüm kimliği 617 olan zararlı yazılım üzerinden doğrudan bir bağlantı olduğu belirlenmiştir. Bu bulgu, daha önce gerçekleştirilen zararlı yazılım kullanım analizinde Çizelge 4.9'da görüldüğü gibi düğüm kimliği 617 olan zararlı yazılımın toplamda beş farklı tehdit aktörü tarafından kullanıldığı tespitiyle tutarlılık göstermektedir. Bu durum, söz konusu zararlı yazılımın farklı tehdit aktörleri arasında ortak bir araç olarak kullanıldığını ve aktörler arasında potansiyel bir işbirliği veya ortak operasyonel yöntemlerin varlığını işaret etmektedir.

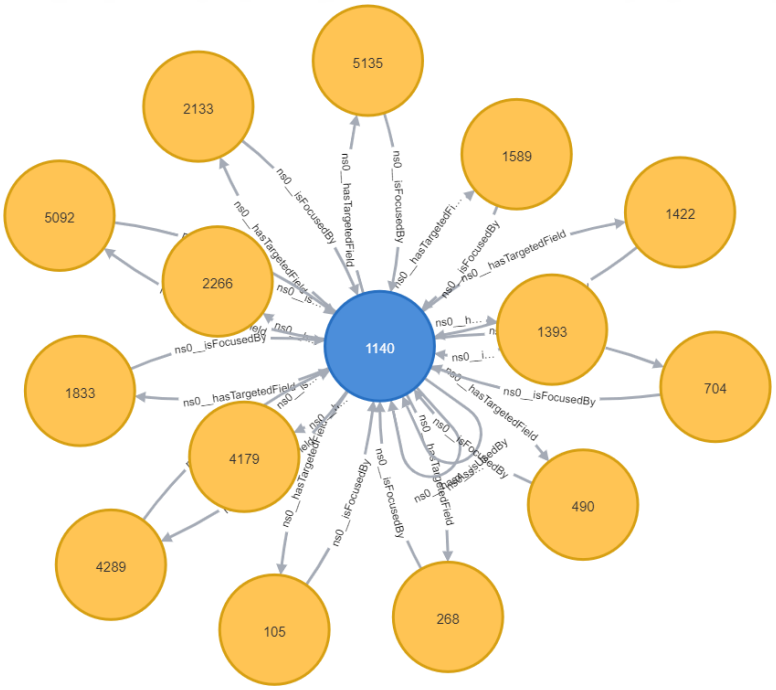
Şekil 4.8 : En aktif iki tehdit aktörü arasındaki olası yollar.

Ayrıca, düğüm kimlikleri 5141 ve 2398 olan tehdit aktörlerinin üç ortak hedefe (Bank, Google ve Banking sektörü) sahip olduğu tespit edilmiştir. Bu ortak hedefler, daha önce gerçekleştirilen ve Çizelge 4.10'da sunulan en çok hedeflenen alanlar analizindeki bulgularla da örtüşmektedir. Bu analizde, Google platformunun toplamda 27 farklı tehdit aktörü tarafından hedeflendiği ve bankacılık sektörünün de tehdit aktörleri için öncelikli hedefler arasında yer aldığı belirlenmiştir. Dolayısıyla, düğüm kimlikleri 5141 ve 2398 olan tehdit aktörlerinin ortak hedef seçimleri, tehdit aktörlerinin hedef belirleme süreçlerinde sektörel yoğunlaşma ve stratejik benzerlikler olduğunu göstermektedir.

Sonuç olarak, bu analizden elde edilen bulgular, tehdit aktörleri arasındaki potansiyel işbirliği yollarını ve ortak yöntemleri ortaya koymaktadır. Özellikle ortak zararlı yazılım kullanımı ve benzer hedef seçimleri, tehdit aktörlerinin saldırı stratejilerinde sistematik ve koordineli bir yaklaşım izleyebileceğini düşündürmektedir. Bu durum, siber güvenlik uzmanlarının tehdit aktörlerinin davranış modellerini daha iyi anlaması ve bu doğrultuda savunma stratejilerini güçlendirmesi açısından önemli içgörüler sağlamaktadır. Ayrıca, tehdit aktörleri arasındaki bağlantıların ve ortak hedeflerin belirlenmesi, kurumlar arası tehdit istihbaratı paylaşımının önemini vurgulamakta ve siber tehditlere karşı daha etkin ve önleyici savunma mekanizmalarının geliştirilmesine katkıda bulunmaktadır.

4.5.3.2 Zararlı yazılımlar ile hedefler arasındaki kritik yollar

Zararlı yazılımlar ile hedefler arasındaki kritik bağlantı yollarını ve ilişki örüntülerini belirlemek amacıyla iki farklı sorgu uygulanmıştır. İlk sorguda, üçten fazla hedefe yönelik saldırı gerçekleştiren zararlı yazılımların tüm hedef bağlantıları detaylı olarak incelenmiştir. Bu sorgu, zararlı yazılımların hedef seçimindeki çeşitliliği ve saldırı kapsamını ortaya koymayı amaçlamaktadır. İkinci sorguda ise, ortak hedeflere sahip olan zararlı yazılımlar arasındaki ilişkiler araştırılarak, zararlı yazılımlar arasındaki potansiyel bağlantılar ve ortak saldırı stratejileri belirlenmeye çalışılmıştır. Bu analizler, önceki bölümlerde gerçekleştirilen ve Çizelge 4.11'de sunulan analiz sonuçlarında tespit edilen, düğüm kimlikleri 2168 (19 hedef), 1140 (14 hedef) ve 617 (11 hedef) olan zararlı yazılımlar gibi çok hedefli zararlı yazılımların hedef seçim örüntülerini daha ayrıntılı biçimde anlamamıza olanak sağlamıştır. Özellikle düğüm kimliği 2168 olan zararlı yazılımın toplamda 19 farklı hedefe yönelik saldırı gerçekleştirmesi, bu zararlı yazılımın geniş bir saldırı yüzeyine sahip olduğunu ve farklı sektörlerdeki hedeflere yönelik sistematik bir saldırı stratejisi izlediğini göstermektedir. Benzer şekilde, düğüm kimliği 1140 ve 617 olan zararlı yazılımların da sırasıyla 14 ve 11 farklı hedefe Şekil 4.9'da görüldüğü üzere saldırması, bu zararlı yazılımların da çoklu hedeflere yönelik saldırılarda kritik rol oynadığını ortaya koymaktadır.



Şekil 4.9 : Zararlı yazılım 1140 ile ilişkili hedef düğümler.

Bu analizlerin sonuçları, zararlı yazılımların hedef seçiminde belirgin örüntüler ve stratejik tercihler olduğunu göstermektedir. Çok hedefli zararlı yazılımların ortak hedefleri paylaşması, bu zararlı yazılımların benzer motivasyonlarla hareket ettiğini veya ortak saldırı kampanyalarında kullanıldığını düşündürmektedir. Ayrıca, ortak hedeflere yönelik saldırılar gerçekleştiren zararlı yazılımlar arasındaki bağlantıların belirlenmesi, siber tehdit ekosistemindeki saldırı zincirlerinin ve potansiyel işbirliklerinin anlaşılması açısından önemli içgörüler sunmaktadır.

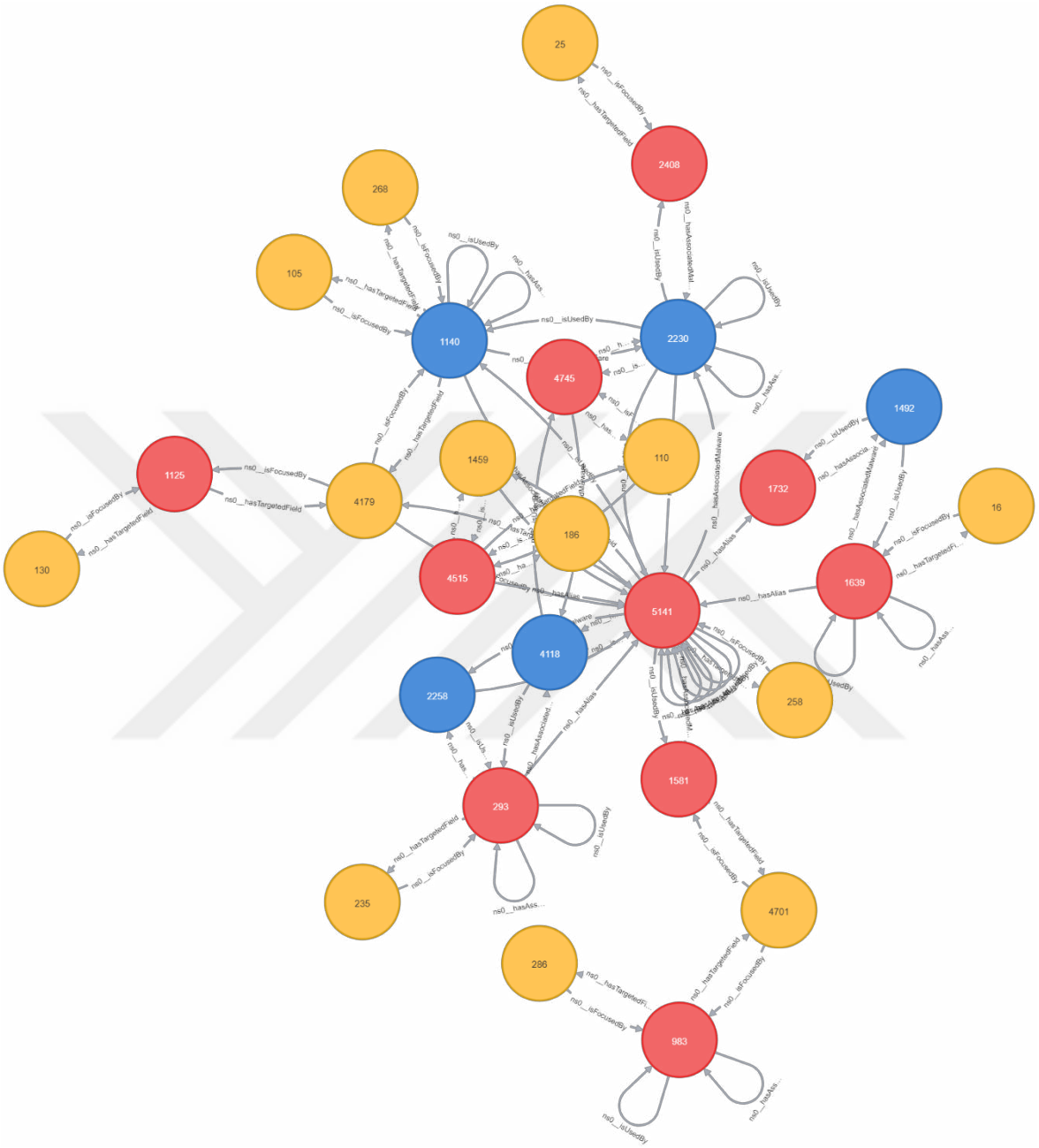
Sonuç olarak, zararlı yazılımlar ile hedefler arasındaki kritik yolların belirlenmesi, siber güvenlik uzmanlarına ve kuruluşlara, zararlı yazılımların saldırı stratejilerini daha iyi anlama ve bu doğrultuda savunma mekanizmalarını güçlendirme konusunda değerli bilgiler sağlamaktadır. Bu analizler sayesinde, zararlı yazılımların yayılma rotaları ve hedeflere ulaşma süreçleri daha net biçimde ortaya konulmakta, böylece siber tehditlere karşı daha etkin ve proaktif savunma stratejilerinin geliştirilmesine katkıda bulunulmaktadır.

4.5.3.3 Saldırı zincirleri analizi

Bu analiz kapsamında, tehdit aktörlerinden hedeflere uzanan saldırı zincirleri iki farklı perspektiften incelenmiştir. İlk olarak, önceki merkezilik analizlerinde yüksek PageRank (47.48) ve maksimum arasındalık merkeziliği (1.0) değerleriyle ağ içerisinde baskın ve merkezi bir konuma sahip olduğu belirlenen düğüm kimliği 5141 olan tehdit aktörünün hedeflere ulaşmak için kullandığı en kısa yollar detaylı olarak araştırılmıştır. Bu sorgu, söz konusu tehdit aktörünün saldırılarını gerçekleştirirken izlediği doğrudan ve dolaylı bağlantı rotalarını ortaya koymayı amaçlamaktadır. İkinci olarak ise, tehdit aktörlerinin zararlı yazılımlar aracılığıyla hedeflerle kurdukları bağlantılar analiz edilmiştir. Bu ikinci sorgu, zararlı yazılımların tehdit aktörleri ile hedefler arasındaki ilişki zincirlerinde oynadığı kritik rolü ve saldırıların gerçekleşme süreçlerini daha net biçimde ortaya koymaktadır.

Gerçekleştirilen yol analizleri sonucunda elde edilen bulgular, siber tehdit ekosisteminin karmaşık, çok katmanlı ve dinamik yapısını açıkça ortaya koymaktadır. Özellikle düğüm kimliği 5141 olan tehdit aktörünün ağ içerisindeki Şekil 4.10'da görüldüğü üzere merkezi konumu, diğer tehdit aktörleri ve zararlı yazılımlarla olan yoğun etkileşimi ve hedeflere yönelik saldırı stratejileri, bu analizlerle daha kapsamlı ve derinlemesine anlaşılmıştır. Bu bulgular, önceki bölümlerde gerçekleştirilen PageRank ve arasındalık

merkeziliği analizlerinin sonuçlarıyla tutarlılık göstermekte ve düğüm kimliği 5141 olan tehdit aktörünün ağdaki kritik rolünü destekleyen yeni kanıtlar sunmaktadır.



Şekil 4.10 : Düğüm kimliği 5141 olan tehdit aktöründen hedeflere giden yollar.

Bu analiz çalışması, tehdit aktörlerinin davranış modellerinin ve saldırı stratejilerinin anlaşılması açısından kritik öneme sahiptir. Özellikle zararlı yazılımlar üzerinden kurulan aktör-hedef bağlantılarının ve ortak hedef seçimlerinin belirlenmesi, siber güvenlik uzmanlarına ve kuruluşlara proaktif güvenlik önlemleri geliştirme konusunda önemli içgörüler sağlamaktadır. Saldırı zincirlerinin detaylı olarak ortaya konulması, tehditlerin

erken aşamada tespit edilmesi, saldırıların gerçekleşmeden önce engellenmesi ve savunma mekanizmalarının güçlendirilmesi açısından büyük önem taşımaktadır. Bu bağlamda, gerçekleştirilen saldırı zincirleri analizi, siber tehditlere karşı daha etkin, önleyici ve stratejik savunma yaklaşımlarının geliştirilmesine önemli katkılar sağlamaktadır.

4.5.4 Siber tehdit ağlarında tahmine yönelik analizler

Siber tehdit aktörlerinin gelecekteki potansiyel hedeflerini öngörebilmek, proaktif ve etkili siber güvenlik stratejilerinin geliştirilmesi açısından kritik bir öneme sahiptir. Bu bağlamda, çizge teorisi tabanlı tahmin yöntemleri, tehdit aktörlerinin davranış kalıplarını ve saldırı eğilimlerini analiz etmek için güçlü ve sistematik bir çerçeve sunmaktadır. Bu yöntemler, tehdit aktörlerinin mevcut bağlantılarını, ilişki örüntülerini ve ağ yapısını detaylı olarak inceleyerek, gelecekteki olası saldırı hedeflerini belirlemeye ve tahmin etmeye olanak sağlamaktadır. Bu çalışmada, tehdit aktörlerinin gelecekteki potansiyel hedeflerini tahmin etmek amacıyla iki farklı çizge teorisi tabanlı analiz yöntemi uygulanmıştır. İlk olarak, ortak komşular (Common Neighbors) analizi gerçekleştirilmiştir. Bu analiz yöntemi, tehdit aktörleri ile hedefler arasındaki mevcut bağlantıları ve ortak bağlantı noktalarını inceleyerek, henüz doğrudan bağlantısı olmayan ancak gelecekte bağlantı kurulma olasılığı yüksek olan düğümleri belirlemeyi amaçlamaktadır. İkinci olarak ise Adamic/Adar İndeksi analizi uygulanmıştır. Adamic/Adar İndeksi, ortak komşular analizine benzer şekilde çalışmakla birlikte, ortak bağlantıların nadirliği ve önem derecesini dikkate alarak daha hassas ve güvenilir tahminler yapılmasına olanak tanımaktadır. Bu tahmine yönelik analizler sayesinde, tehdit aktörlerinin gelecekteki saldırı hedefleri hakkında önemli içgörüler elde edilmiştir. Bu içgörüler, siber güvenlik uzmanlarına ve kuruluşlara, tehdit aktörlerinin saldırı stratejilerini önceden tahmin etme ve bu doğrultuda proaktif güvenlik önlemleri geliştirme konusunda değerli bilgiler sağlamaktadır. Böylece, tehditlerin gerçekleşmeden önce tespit edilmesi ve önlenmesi mümkün hale gelmekte, siber güvenlik savunma mekanizmalarının etkinliği önemli ölçüde artırılmaktadır. Sonuç olarak, gerçekleştirilen tahmine yönelik analizler, siber tehditlere karşı daha stratejik, proaktif ve etkili savunma yaklaşımlarının oluşturulmasına önemli katkılar sunmaktadır.

4.5.4.1 Ortak komşuluk analizi

Ortak komşuluk (Common Neighbors) analizi, iki düğüm arasındaki potansiyel bağlantıyı tahmin etmek amacıyla kullanılan temel ve etkili bir çizge teorisi yöntemidir. Bu yaklaşım, iki düğümün paylaştığı ortak komşu düğümlerin sayısını temel alarak, ortak

bağlantı sayısı arttıkça iki düğüm arasında gelecekte bağlantı kurulma olasılığının da artacağını varsayar. Siber güvenlik bağlamında ortak komşuluk analizi, tehdit aktörlerinin hedef seçim kalıplarını ve gelecekteki potansiyel saldırı hedeflerini anlamak açısından önemli bir araç olarak kullanılmaktadır. Bu çalışmada gerçekleştirilen ortak komşuluk analizi sonucunda, düğüm kimliği 5141 olan tehdit aktörünün özellikle devlet kurumlarına yönelik güçlü ve sistematik bir ilgi gösterdiği ortaya çıkmıştır. Analiz sonuçları, düğüm kimliği 5141 olan tehdit aktörünün, düğüm kimlikleri 933, 1581 ve 4515 olan diğer tehdit aktörleri ile birlikte devlet kurumları (Government agencies) hedefine yönelik üç ortak bağlantıya sahip olduğunu göstermektedir. Ayrıca, düğüm kimliği 5141 olan tehdit aktörünün "Crypto" ve "Government agency" hedeflerine yönelik de ikişer ortak bağlantıya sahip olduğu belirlenmiştir. Bu analiz sonuçları Çizelge 4.15'te detaylı olarak sunulmuştur.

Çizelge 4.15 : Ortak komşuluk analiz sonuçları.

Tehdit Aktörü	Potansiyel Hedef	Ortak Bağlantılar	Ortak Düğümler
5141	Goverment agencies	3	933, 1581, 4515
5141	Crypto	2	687, 2261
5141	Goverment agency	2	254, 2168

Bu bulgular, düğüm kimliği 5141 olan tehdit aktörünün çok yönlü ve stratejik bir saldırı yaklaşımı benimsediğini göstermektedir. Özellikle devlet kurumlarına yönelik yüksek ilgi ve farklı sektörlere yayılan hedef çeşitliliği, bu tehdit aktörünün kapsamlı bir istihbarat toplama, etki yaratma ve stratejik hedeflere ulaşma amacı taşıdığını düşündürmektedir. Diğer tehdit aktörleriyle paylaşılan ortak hedeflerin varlığı ise, bu hedef seçimlerinin rastgele olmadığını, aksine bilinçli ve koordineli bir stratejik yaklaşımın ürünü olduğunu ortaya koymaktadır.

4.5.4.2 Adamic/Adar İndeks Analizi

Adamic/Adar indeksi, ortak komşuluk (Common Neighbors) analizinin daha gelişmiş ve karmaşık bir versiyonudur. Bu yöntem, iki düğüm arasındaki ortak bağlantıların popülerliğini logaritmik bir ölçekte ters orantılı olarak değerlendirerek, daha nadir ve özgün bağlantılara daha yüksek ağırlık vermektedir. Böylece, sık rastlanan bağlantılardan ziyade, daha spesifik ve anlamlı bağlantılar ön plana çıkarılarak tahminlerin doğruluğu artırılmaktadır. Gerçekleştirilen Adamic/Adar indeksi analizi sonucunda, düğüm kimliği 5141 olan tehdit aktörünün potansiyel hedefleri için tutarlı bir şekilde 2,71'lik Adamic/Adar indeksi puanı ve 10 ortak bağlantı tespit edilmiştir. Analiz sonuçlarında, kritik altyapı hedefleri arasında özellikle SSH servisleri gibi operasyonel sistemler öne çıkarken, oyun

sektörü, ve kafeler gibi sivil hedefler de dikkat çekici biçimde yer almaktadır. Bu analiz sonuçları Çizelge 4.16'da detaylı olarak sunulmuştur.

Çizelge 4.16 : Adamic/Adar indeks analizi sonuçları.

Tehdit Aktörü	Potansiyel Hedef	Adamic/Adar indeks	Ortak Bağlantılar
5141	SSH service	2,71	10
5141	Gaming	2,71	10
5141	Coffee shops	2,71	10

Bu bulgular, düğüm kimliği 5141 olan tehdit aktörünün altyapı sistemlerine yönelik sistematik ve stratejik bir ilgi gösterdiğini ortaya koymaktadır. Aynı zamanda, tehdit aktörünün sivil hedef yelpazesine de sahip olması, grubun farklı amaçlar doğrultusunda farklı sektörleri hedeflediğini ve kapsamlı, çok yönlü bir siber saldırı stratejisi izlediğini düşündürmektedir. Bu çeşitlilik, tehdit aktörünün hem kritik altyapı sistemlerini hedefleyerek etki yaratmayı, hem de sivil alanlara yönelik saldırılarla dikkat çekmeyi amaçladığını göstermektedir. Her iki analiz yöntemi (Ortak Komşuluk ve Adamic/Adar indeksi) birbirini tamamlayıcı nitelikte sonuçlar ortaya koymuştur. Ortak komşuluk analizi, tehdit aktörünün temel hedef profilini ve genel eğilimlerini belirlerken, Adamic/Adar indeksi analizi bu profili daha incelikli ve detaylı bir şekilde değerlendirme imkânı sunmuştur. Ayrıca, düğüm kimliği 5141 olan tehdit aktörünün önceki analizlerde elde edilen yüksek PageRank skoru (47.48) ve maksimum arasındalık merkeziliği (betweenness) değeri (1.0), bu aktörün siber tehdit ekosistemindeki merkezi ve kritik konumunu doğrulamakta ve desteklemektedir.

Sonuç olarak, Adamic/Adar indeksi analizi sayesinde elde edilen bu bulgular, tehdit aktörlerinin gelecekteki saldırı hedeflerini daha hassas ve güvenilir biçimde öngörmek açısından önemli katkılar sağlamaktadır. Bu analizler, siber güvenlik uzmanlarına tehdit aktörlerinin davranış modellerini daha iyi anlama, saldırıların gerçekleşmeden önce önlenmesi için gerekli stratejileri oluşturma ve önleyici güvenlik önlemleri geliştirme konusunda değerli içgörüler sunmaktadır.

5. SONUÇ

Bu çalışmada, siber tehdit istihbarat verilerinin analizinde doğal dil işleme teknikleri ile bilgi çizgelerini bir araya getiren yenilikçi bir yaklaşım geliştirilmiş ve önerilmiştir. Önerilen yaklaşım, firma ve kurumlar tarafından yayınlanan güvenlik raporları, APT saldırı raporları ve API kaynaklarından elde edilen yapılandırılmamış tehdit verilerini işleyerek tehdit aktörleri, kötü amaçlı yazılımlar, saldırı kampanyaları ve saldırı hedefleri gibi kritik siber varlıkları başarıyla tanımlamış ve aralarındaki ilişkileri analiz edilebilir hâle getirmiştir. Bu analizler, tehdit aktörlerinin belirli hedef alanlarla olan ilişkilerini görünür kılarak, daha önce tespit edilemeyen bağlantıların ortaya çıkarılmasına olanak sağlamıştır. Süreç boyunca, yapay zeka tabanlı BERT dil modeli kullanılarak, ham metinlerden cümle yapılarının çıkarımı ve temizlenmiş metinlerden adlandırılmış varlıkların tespiti gerçekleştirilmiştir. Dil modelleri ile bilgi çizgelerinin kombinasyonu, dinamik tehdit ortamında anlamlı içgörüler sunan etkili bir analiz yapısı ortaya koymuştur. Çalışma kapsamında kullanılan yaklaşımların performansı, elde edilen doğruluk oranları ve çizge tabanlı analizlerin tutarlılığı ile teyit edilmiştir. Elde edilen siber varlıklar arasındaki ilişkiler, ontoloji tabanlı yaklaşımlarla yapılandırılarak neo4j ortamında çizge algoritmalarıyla analiz edilmiş ve tehditlerin yapısal örüntüleri derinlemesine incelenmiştir.

Tez çalışması kapsamında geliştirilen ve önerilen uygulama, yapay zeka tabanlı BERT dil modeli ve bilgi çizgesi temelli hibrit yaklaşımın siber tehdit istihbaratında etkili bir yaklaşım olduğu göstermektedir. Geleneksel yöntemlerle tespit edilmesi zor olan tehdit aktörleri ve hedefler arasındaki ilişkiler daha açık ve yorumlanabilir hale gelmiştir. Ayrıca, sistemin sunduğu analizler, mevcut tehditleri daha iyi anlamanın yanı sıra gelecekteki potansiyel hedefleri öngörme yeteneğine de sahiptir. Böylece, siber güvenlik stratejilerinin daha sağlam temellere dayanarak geliştirilmesi ve tehditlere karşı önlemler alınması mümkün hale gelecektir. Geliştirilen yaklaşım, finans, nesnelerin interneti, enerji ve akıllı şebekeler gibi kritik altyapılar için yayınlanan metin tabanlı raporlar için de kullanılabilir. Bu yönüyle, sektörel bazda siber güvenlik stratejilerinin iyileştirilmesine katkı sağlamaktadır.

Gelecekteki çalışmalarda, tehdit aktörleri, kötü amaçlı yazılımlar ve davranış kalıplarını etiketleyen daha geniş bir veri seti oluşturulabilir. Böylece daha detaylı analizlerin yapılabilmesine olanak sağlayabilir. Sistemin analiz kapasitesi, yalnızca siber varlık çıkarımıyla kalmayıp, tehditlere karşı geliştirilen savunma yöntemlerinin de eğitim

verilerinde etiketlenerek daha kapsamlı bir mimariye genişletilebilir. Bu sayede siber tehditlere karşı, güvenlik uygulayıcılarının savunma stratejileri geliştirmesi daha basit indirgenmiş olabilir ve daha etkin çözümler geliştirilebilir. Önerilen sisteme özelleştirilebilir varlık ilişkisi özelliği de eklenebilir. Bu özellik sayesinde güvenlik uzmanlarının analiz ihtiyaçlarına uygun olarak varlıklar arasındaki ilişkileri seçmelerine olanak sağlanabilir. Güvenlik uzmanları, tehdit aktörleri, hedefler veya kötü amaçlı yazılımlar arasındaki belirli ilişkileri önceliklendirebilir. Böylece sistem güvenlik uzmanlarının ihtiyaçlarına özel, daha etkili analizler ve daha esnek bir çözüm sunabilir. Ayrıca, BERT modeli, ham ve etiketlenmemiş metinlerle MLM görevi için yeniden eğitilebilir. Bu sayede model, alana özgü kavramları ve bağlamsal ilişkileri daha iyi öğrenebilir. Ancak bu görev için çok fazla siber güvenlik metin girdisi gerekmektedir. Siber güvenlik alanına özgü bir dil modeli geliştirmek için, güvenlik raporları, günlük kayıtları ve yapılandırılmamış metinler kullanılarak bu model eğitilebilir. Bu tür bir model, siber tehdit istihbaratında kullanılan terimleri ve bağlamları daha doğru bir şekilde anlamlandırabilir. Örneğin, tehdit aktörleri, zararlı yazılımlar, saldırı kampanyaları, saldırı davranışları, hedefler ve daha fazla siber güvenlik objesi ilgili metinlerden çıkarılabilir. Ayrıca, teknik belgelerdeki karmaşık cümle yapılarını çözümlemek ve bağlamsal ilişkileri analiz etmek için etkili bir araç olabilir. Özellikle Türkçe içerikli siber güvenlik metinlerinde, dilsel farklılıkların üstesinden gelmek için böyle bir modelin geliştirilmesi önemlidir. RAG tabanlı bir sistemle entegre edildiğinde, model, sorgulara daha hassas ve alana özgü yanıtlar sunabilir. Bu tür bir dil modeli, siber tehditlerin analizi ve önlenmesinde kritik bir rol oynayabilir.

KAYNAKLAR

[1] **MITRE.** (2024). MITRE ATT&CK Framework. [Çevrimiçi]. Erişim adresi: https://attack.mitre.org

[2] **Meta.** (2024). ThreatExchange: A Threat Intelligence Sharing Platform. [Çevrimiçi]. Erişim adresi: https://developers.facebook.com/products/threat-exchange/

[3] **Symantec.** (t.y.). Symantec Enterprise Blogs-Threat Intelligence". [Çevrimiçi]. Erişim adresi: https://symantec-enterprise-blogs.security.com/blogs/threat-intelligence.

[4] **Barnum, S.** (2014). Standardizing cyber threat intelligence information with the structured threat information expression (stix). *Mitre Corporation, 11,* 1-22.

[5] **MITRE.** (2022). MAEC-Malware Attribute Enumeration and Characterization, Ağustos 2022. [Çevrimiçi]. Erişim adresi: https://maecproject.github.io/

[6] **Husari, G., Al-Shaer, E., Ahmed, M., Chu, B., & Niu, X.** (2017, December). Ttpdrill: Automatic and accurate extraction of threat actions from unstructured text of cti sources. In *Proceedings of the 33rd annual computer security applications conference* (pp. 103-115). https://doi.org/10.1145/3134600.3134646

[7] **Zhou, Y., Tang, Y., Yi, M., Xi, C., & Lu, H.** (2022). CTI view: APT threat intelligence analysis system. *Security and Communication Networks, 2022*(1), 9875199. https://doi.org/10.1155/2022/9875199

[8] **Jones, C. L., Bridges, R. A., Huffer, K. M., & Goodall, J. R.** (2015, April). Towards a relation extraction framework for cyber-security concepts. In *Proceedings of the 10th Annual Cyber and Information Security Research Conference* (pp. 1-4). https://doi.org/10.1145/2746266.2746277

[9] **Alves, F., Bettini, A., Ferreira, P. M., & Bessani, A.** (2021). Processing tweets for cybersecurity threat awareness. *Information Systems, 95*, 101586. https://doi.org/10.1016/j.is.2020.101586

[10] **Kim, E., Kim, K., Shin, D., Jin, B., & Kim, H.** (2018, June). CyTIME: Cyber Threat Intelligence ManagEment framework for automatically generating security rules. In *Proceedings of the 13th International Conference on Future Internet Technologies* (pp. 1-5). https://doi.org/10.1145/3226052.3226056

[11] **Zhang, H., Shen, G., Guo, C., Cui, Y., & Jiang, C.** (2021). EX-Action: Automatically extracting threat actions from cyber threat intelligence report based on multimodal learning. *Security and Communication Networks, 2021*(1), 5586335. https://doi.org/10.1155/2021/5586335

[12] **Piplai, A., Mittal, S., Joshi, A., Finin, T., Holt, J., & Zak, R.** (2020). Creating cybersecurity knowledge graphs from malware after action reports. *IEEE Access, 8, 211691-211703.* https://doi.org/10.1109/ACCESS.2020.3039234

[13] **Sarhan, I., & Spruit, M.** (2021). Open-cykg: An open cyber threat intelligence knowledge graph. *Knowledge-Based Systems, 233,* 107524. https://doi.org/10.1016/j.knosys.2021.107524

[14] **Alam, M. T., Bhusal, D., Park, Y., & Rastogi, N.** (2022). CyNER: A python library for cybersecurity named entity recognition. *arXiv* preprint arXiv:2204.05754. https://doi.org/10.48550/arXiv.2204.05754

[15] **Zhu, Z., & Dumitras, T.** (2018, April). Chainsmith: Automatically learning the semantics of malicious campaigns by mining threat intelligence reports. In *2018 IEEE European symposium on security and privacy (EuroS&P).* London, UK. *(*pp. 458-472). IEEE. https://doi.org/10.1109/EuroSP.2018.00039

[16] **Noor, U., Anwar, Z., Amjad, T., & Choo, K. K. R.** (2019). A machine learning-based FinTech cyber threat attribution framework using high-level indicators of compromise. *Future Generation Computer Systems, 96,* 227-242. https://doi.org/10.1016/j.future.2019.02.013

[17] **Qin, Y., Shen, G. W., Zhao, W. B., Chen, Y. P., Yu, M., & Jin, X.** (2019). A network security entity recognition method based on feature template and CNN-BiLSTM-CRF. *Frontiers of Information Technology & Electronic Engineering, 20(*6), 872-884. https://doi.org/10.1631/FITEE.1800520

[18] **Bose, A., Behzadan, V., Aguirre, C., & Hsu, W. H.** (2019). A novel approach for detection and ranking of trendy and emerging cyber threat events in Twitter streams. In *2019 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)* (pp. 871–878). IEEE. https://doi.org/10.1145/3341161.3344379

[19] **Ma, P., Jiang, B., Lu, Z., Li, N., & Jiang, Z.** (2020). Cybersecurity named entity recognition using bidirectional long short-term memory with conditional random fields. *Tsinghua Science and Technology, 26*(3), 259-265. https://doi.org/10.26599/TST.2019.9010033

[20] **Kim, D., & Kim, H. K.** (2019). Automated dataset generation system for collaborative research of cyber threat analysis. *Security and Communication Networks, 2019(*1), 6268476. https://doi.org/10.1155/2019/6268476

[21] **Zhang, H., Guo, Y., & Li, T.** (2019). Multifeature named entity recognition in information security based on adversarial learning. *Security and Communication Networks, 2019(*1), 6417407. https://doi.org/10.1155/2019/6417407

[22] **Georgescu, T. M., Iancu, B., & Zurini, M.** (2019). Named-entity-recognition-based automated system for diagnosing cybersecurity situations in IoT networks. *Sensors, 19(*15), 3380. https://doi.org/10.3390/s19153380.

[23] **Sun, T., Yang, P., Li, M., & Liao, S.** (2021). An automatic generation approach of the cyber threat intelligence records based on multi-source information fusion. *Future Internet, 13*(2), 40. https://doi.org/10.3390/fi13020040

[24] **Wu, H., Li, X., & Gao, Y.** (2020). An effective approach of named entity recognition for cyber threat intelligence. In *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)* (Vol. 1, pp. 1370-1374). IEEE. https://doi.org/10.1109/ITNEC48623.2020.9085102

[25] **Simran, K., Sriram, S., Vinayakumar, R., & Soman, K.P.** (2020). *Deep learning approach for ıntelligent named entity recognition of cyber security.* In: Thampi, S., et al. Advances in Signal Processing and Intelligent Recognition Systems. SIRS 2019. Communications in Computer and Information Science, vol 1209. Springer, Singapore. https://doi.org/10.1007/978-981-15-4828-4_14

[26] **Kim, G., Lee, C., Jo, J., & Lim, H.** (2020). Automatic extraction of named entities of cyber threats using a deep Bi-LSTM-CRF network. *International Journal of Machine Learning and Cybernetics, 11,* 2341-2355. https://doi.org/10.1007/s13042-020-01122-6

[27] **Jia, J., Yang, L., Wang, Y., & Sang, A.** (2025). Hyper attack graph: Constructing a hypergraph for cyber threat intelligence analysis. *Computers & Security, 149,* 104194. https://doi.org/10.1016/j.cose.2024.104194

[28] **Srivastava, S., Paul, B., & Gupta, D.** (2023). Study of word embeddings for enhanced cyber security named entity recognition. *Procedia Computer Science, 218,* 449-460. https://doi.org/10.1016/j.procs.2023.01.027

[29] **Ahmed, K., Khurshid, S. K., & Hina, S.** (2024). CyberEntRel: Joint extraction of cyber entities and relations using deep learning. *Computers & Security, 136*, 103579. https://doi.org/10.1016/j.cose.2023.103579

[30] **Satvat, K., Gjomemo, R., & Venkatakrishnan, V. N.** (2021, September). Extractor: Extracting attack behavior from threat reports. In *2021 IEEE European Symposium on Security and Privacy (EuroS&P)* (pp. 598-615). IEEE. https://doi.org/10.1109/EuroSP51992.2021.00046

[31] **Guo, Y., Liu, Z., Huang, C., Wang, N., Min, H., Guo, W., & Liu, J.** (2023). A framework for threat intelligence extraction and fusion. *Computers & Security, 132,* 103371. https://doi.org/10.1016/j.cose.2023.103371

[32] **Jo, H., Lee, Y., & Shin, S.** (2022). Vulcan: Automatic extraction and analysis of cyber threat intelligence from unstructured text. *Computers & Security, 120,* 102763. https://doi.org/10.1016/j.cose.2022.102763

[33] **Wang, G., Liu, P., Huang, J., Bin, H., Wang, X., & Zhu, H.** (2024). KnowCTI: Knowledge-based cyber threat intelligence entity and relation extraction. *Computers & Security, 141,* 103824. https://doi.org/10.1016/j.cose.2024.103824

[34] **DeWitt, D., & Gray, J.** (1992). Parallel database systems: The future of high performance database systems. *Communications of the ACM, 35*(6), 85-98. https://doi.org/10.1145/129888.129894

[35] **Walter, T.** (2009). Teradata past, present, and future. *UCI ISG lecture series on scalable data management, 1*(1), 44-48.

[36] **Tansley, S., & Tolle, K. M. (2009).** *The fourth paradigm: data-intensive scientific discovery* (Vol. 1). T. Hey (Ed.). Redmond, WA: Microsoft research.

[37] **Ghemawat, S., Gobioff, H., & Leung, S. T.** (2003, October). The Google file system. In *Proceedings of the nineteenth ACM symposium on Operating systems principles* (pp. 29-43). https://doi.org/10.1145/1165389.945450

[38] **Dean, J., & Ghemawat, S.** (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM, 51*(1), 107-113. https://doi.org/10.1145/1327452.1327492

[39] **Erl, T., Khattak, W., & Buhler, P.** (2016). *Big data fundamentals: concepts, drivers & techniques.* Prentice Hall Press.

[40] **Laney, D.** (2001). 3D data management: controlling data volume, velocity and variety. *META Group Research Note 6* (70), 1.

[41] **Lee, I.** (2017). Big data: Dimensions, evolution, impacts, and challenges. *Business horizons, 60(*3), 293-303. https://doi.org/10.1016/j.bushor.2017.01.004

[42] **Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V.** (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research, 70,* 263-286. https://doi.org/10.1016/j.jbusres.2016.08.001

[43] **Chen, C. P., & Zhang, C. Y.** (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences, 275,* 314-347. https://doi.org/10.1016/j.ins.2014.01.015

[44] **Wang, Y., & Wiebe, V. J.** (2014). Big Data Analytics on the characteristic equilibrium of collective opinions in social networks. *In International Journal of Cognitive Informatics and Natural Intelligence (IJCINI). 8*(3), 29–44. https://doi.org/10.4018/IJCINI.2014070103

[45] **Zhang, F., Liu, M., Gui, F., Shen, W., Shami, A., & Ma, Y.** (2015). A distributed frequent itemset mining algorithm using Spark for Big Data analytics. *Cluster Computing, 18,* 1493-1501. https://doi.org/10.1007/s10586-015-0477-1

[46] **Chen, H., Chiang, R. H., & Storey, V. C.** (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly,36,*(4), 1165-1188. https://doi.org/10.2307/41703503

[47] **Edwards, R., & Fenwick, T.** (2016). Digital analytics in professional work and learning. *Studies in Continuing Education, 38*(2), 213-227. https://doi.org/10.1080/0158037X.2015.1074894

[48] **Kong, W., Wu, Q., Li, L., & Qiao, F.** (2014, July). Intelligent Data Analysis and its challenges in big data environment. In *2014 IEEE International Conference on System Science and Engineering (ICSSE)* (pp. 108-113). IEEE. https://doi.org/10.1109/ICSSE.2014.6887915

[49] **Gantz J, & Reinsel D.** (2011). Extracting value from chaos. IDC iView,1-12.

[50] **Deshpande, A., & Kumar, M.** (2018). *Artificial intelligence for big data: complete guide to automating big data solutions using artificial intelligence techniques.* Packt Publishing Ltd.

[51] **Hilbert, M., & López, P.** (2011). The world's technological capacity to store, communicate, and compute information. *Science, 332*(6025), 60-65. https://doi.org/10.1126/science.1200970

[52] **Cavanillas, J. M., Curry, E., & Wahlster, W.** (2016). *New horizons for a data-driven economy: a roadmap for usage and exploitation of big data in Europe.* Springer Nature.

[53] **Mittal, M., Balas, V. E., Goyal, L. M., & Kumar, R. (Eds.).** (2019). *Big data processing using spark in cloud* (Vol. 43). Berlin, Germany: Springer. https://doi.org/10.1007/978-981-13-0550-4

[54] **Hu, H., Wen, Y., Chua, T. S., & Li, X.** (2014). Toward scalable systems for big data analytics: A technology tutorial. *IEEE access, 2*, 652-687**.** https://doi.org/10.1109/ACCESS.2014.2332453

[55] **Sruthika, S., & Tajunisha, N.** (2015, March). A study on evolution of data analytics to big data analytics and its research scope. In *2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)* (pp. 1-6). IEEE. https://doi.org/10.1109/ICIIECS.2015.7193065

[56] **Loshin, D.** (2012). *Business intelligence: The savvy manager's guide*. Morgan Kaufmann Pub.

[57] **Adams, N. M.** (2010). Perspectives on data mining. *International Journal of Market Research, 52*(1), 11-19. https://doi.org/10.2501/S147078531020103X

[58] **Talia, D.** (2013). Clouds for scalable big data analytics. *Computer, 46*(5), 98-101. https://doi.org/10.1109/MC.2013.162

[59] **Simoff, S., Böhlen, M. H., & Mazeika, A. (Eds.).** (2008). *Visual data mining: theory, techniques and tools for visual analytics* (Vol. 4404). Springer Science & Business Media.

[60] **Chen, F., Deng, P., Wan, J., Zhang, D., Vasilakos, A. V., & Rong, X.** (2015). Data mining for the internet of things: literature review and challenges. *International Journal of Distributed Sensor Networks, 11*(8), 431047. https://doi.org/10.1155/2015/431047

[61] **L'heureux, A., Grolinger, K., Elyamany, H. F., & Capretz, M. A.** (2017). Machine learning with big data: Challenges and approaches. *IEEE Access, 5*, 7776-7797. https://doi.org/10.1109/ACCESS.2017.2696365

[62] **Taylor, R. C.** (2010). An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics. *BMC Bioinformatics, 11*(12), 1-6. https://doi.org/10.1186/1471-2105-11-S12-S1

[63] **Lane, N. D., Xu, Y., Lu, H., Campbell, A. T., Choudhury, T., & Eisenman, S. B.** (2011). Exploiting social networks for large-scale human behavior modeling. *IEEE Pervasive Computing, 10*(4), 45-53. https://doi.org/10.1109/MPRV.2011.70

[64] **Shen, Z., Ma, K. L., & Eliassi-Rad, T.** (2006). Visual analysis of large heterogeneous social networks by semantic and structural abstraction. *IEEE Transactions on Visualization and Computer Graphics, 12*(6), 1427-1439. https://doi.org/10.1109/TVCG.2006.107

[65] **Ma, H., King, I., & Lyu, M. R.** (2012). Mining web graphs for recommendations. *IEEE Transactions on Knowledge and Data Engineering, 24*(6), 1051-1064**.** https://doi.org/10.1109/TKDE.2011.18

[66] **Xiaohui, Z., & Xianghua, M.** (2021). A reputation-based approach using consortium blockchain for cyber threat intelligence sharing. *arXiv preprint* arXiv:2107.06662.

[67] **Christian D.** (2018). Cyber threat intelligence standards–a high level overview, In *2018 CTI-EU | Bonding EU Cyber Threat Intelligence Workshop*, The European Union Agency for Cybersecurity (ENISA), 2018.

[68] **Tounsi, W., & Rais, H.** (2018). A survey on technical threat intelligence in the age of sophisticated cyber attacks. *Computers & Security, 72*, 212-233. https://doi.org/10.1016/j.cose.2017.09.001

[69] **Alaeifar, P., Pal, S., Jadidi, Z., Hussain, M., & Foo, E.** (2024). Current approaches and future directions for cyber threat intelligence sharing: A survey. *Journal of Information Security and Applications, 83,* 103786. https://doi.org/10.1016/j.jisa.2024.103786

[70] **Connolly, J., Davidson, M., Richard, M., Skorupka,C.** (2012). The trusted automated exchange of indicator information (taxii), 2014, *The MITRE Corporation*.

[71] **Buttner, A., & Ziring, N.** (2009). Common platform enumeration (cpe)-specification version 2.2. The MITRE Corporation-National Security Agency, 3.

[72] **Barnum, S.** (2008). Common attack pattern enumeration and classification (CAPEC) schema. Department of Homeland Security.

[73] **ThreatConnect.** (2023). ThreatConnect Intelligence Platform (TIP). Erişim: 18 Haziran 2023. [Çevrimiçi]. Erişim adresi: https://threatconnect.com/threat-intelligence-platform/

[74] **Recorded Future.** (2023). Recorded Future Intelligence Platform. Erişim: 18 Haziran 2023. [Çevrimiçi]. Erişim adresi: https://www.recordedfuture.com/platform/intelligence-cloud

[75] **IBM Secuirty.** (2023). IBM X-Force Exchange. Erişim: 18 Haziran 2023. [Çevrimiçi]. Erişim adresi: https://exchange.xforce.ibmcloud.com/

[76] **Inc. Cisco Systems.** (2023). Cisco Talos Intelligence Group. Erişim: 18 Haziran 2023. [Çevrimiçi]. Erişim adresi: https://talosintelligence.com/

[77] **AO Kaspersky Lab.** (2023). Kaspersky Threat Intelligence Services. Erişim: 18 Haziran 2023. [Çevrimiçi]. Erişim adresi: https://www.kaspersky.com/enterprise-security/threat-intelligence

[78] **Inc. CrowdStrike Holdings.** (2023). The CrowdStrike Falcon. Erişim: 18 Haziran 2023. [Çevrimiçi]. Erişim adresi: https://www.crowdstrike.com/platform/

[79] **Z. Pokorny.** (2019). *The threat intelligence handbook: Moving toward a security intelligence program*. Annapolis, MD, USA: CyberEdge Group, 2019.

[80] **Janiszewski, M., Felkner, A., & Lewandowski, P.** (2019). A novel approach to national-level cyber risk assessment based on vulnerability management and threat intelligence. *Journal of Telecommunications and Information Technology, (2),* 2019. https://doi.org/10.26636/jtit.2019.130919

[81] **Riesco, R., & Villagrá, V. A.** (2019). Leveraging cyber threat intelligence for a dynamic risk framework: Automation by using a semantic reasoner and a new combination of standards (STIX™, SWRL and OWL). *International Journal of Information Security, 18*(6), 715-739. https://doi.org/10.1007/s10207-019-00433-2

[82] **Kayode-Ajala, O.** (2023). Applications of Cyber Threat Intelligence (CTI) in financial institutions and challenges in its adoption. *Applied Research in Artificial Intelligence and Cloud Computing, 6(*1), 1-21.

[83] **Kusal, S., Patil, S., Choudrie, J., Kotecha, K., Vora, D., & Pappas, I.** (2023). A systematic review of applications of natural language processing and future challenges with special emphasis in text-based emotion detection. *Artificial Intelligence Review, 56*(12), 15129-15215. https://doi.org/10.1007/s10462-023-10509-0

[84] **Mimura, M., & Ito, R.** (2022). Applying NLP techniques to malware detection in a practical environment. *International Journal of Information Security, 21*(2), 279-291. https://doi.org/10.1007/s10207-021-00553-8

[85] **Lansley, M., Polatidis, N., & Kapetanakis, S.** (2019). Seader: A social engineering attack detection method based on natural language processing and artificial neural networks. In *Computational Collective Intelligence: 11th International Conference, ICCCI 2019, Hendaye, France, September 4–6, 2019, Proceedings, Part I 11* (pp. 686-696). Springer International Publishing.

[86] **Gao, C., Zhang, X., Han, M., & Liu, H.** (2021). A review on cyber security named entity recognition. *Frontiers of Information Technology & Electronic Engineering, 22*(9), 1153-1168. https://doi.org/10.1631/FITEE.2000286

[87] **Martin, J. H., & Jurafsky, D.** (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (Vol. 23). Upper Saddle River: Pearson/Prentice Hall. https://web.stanford.edu/~jurafsky/slp3/, 2025

[88] **Nadeau, D., & Sekine, S.** (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes, 30*(1), 3-26. https://doi.org/10.1075/li.30.1.03nad

[89] **Bridges, R. A., Jones, C. L., Iannacone, M. D., Testa, K. M., & Goodall, J. R.** (2013). Automatic labeling for entity extraction in cyber security. *arXiv preprint arXiv*:1308.4941**.** https://doi.org/10.48550/arXiv.1308.4941

[90] **Yadav, V., & Bethard, S.** (2019). A survey on recent advances in named entity recognition from deep learning models. *arXiv preprint arXiv*:1910.11470.

[91] **Devlin, J., Chang, M. W., Lee, K., & Toutanova, K.** (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171-4186). https://doi.org/10.18653/V1/N19-1423

[92] **Lafferty, J., McCallum, A., & Pereira, F.** (2001, June). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Icml (Vol. 1, No. 2, p. 3).

[93] **Joshi, A., Lal, R., Finin, T., & Joshi, A.** (2013, September). Extracting cybersecurity related linked data from text. In 2013 *IEEE seventh international conference on semantic computing* (pp. 252-259). IEEE. https://doi.org/10.1109/ICSC.2013.50

[94] **Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser L., & Polosukhin, I.** (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

[95] **Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Zettlemoyer, L., & Stoyanov, V.** (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:*1907.11692.

[96] **Sanh, V., Debut, L., Chaumond, J., & Wolf, T.** (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:*1910.01108.

[97] **Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R.** (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv*:1909.11942.

[98] **Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Myle Ott, M., Zettlemoyer, L., & Stoyanov, V.** (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv*:1911.02116.

[99] **Dionísio, N., Alves, F., Ferreira, P. M., & Bessani, A.** (2019, July). Cyberthreat detection from twitter using deep neural networks. In *2019 international joint conference on neural networks (IJCNN)* (pp. 1-8). IEEE. https://doi.org/10.1109/IJCNN.2019.8852475

[100] **Mittal, S., Das, P. K., Mulwad, V., Joshi, A., & Finin, T.** (2016, August). Cybertwitter: Using twitter to generate alerts for cybersecurity threats and vulnerabilities. In 2016 IEEE/ACM

International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 860-867). IEEE. https://doi.org/10.1109/ASONAM.2016.7752338

[101] **Das, R., & Soylu, M.** (2023). A key review on graph data science: The power of graphs in scientific studies. *Chemometrics and Intelligent Laboratory Systems, 240*, 104896. https://doi.org/10.1016/j.chemolab.2023.104896

[102] **Fernandes, D., & Bernardino, J.** (2018). Graph Databases Comparison: AllegroGraph, ArangoDB, InfiniteGraph, Neo4J, and OrientDB. In *Proceedings of the 7th International Conference on Data Science, Technology and Applications - DATA;* ISBN 978-989-758-318-6; ISSN 2184-285X, SciTePress, pages 373-380**.** https://doi.org/10.5220/0006910203730380

[103] **Yoon, B. H., Kim, S. K., & Kim, S. Y.** (2017). Use of graph database for the integration of heterogeneous biological data. *Genomics & informatics, 15*(1), 19-27. https://doi.org/10.5808/GI.2017.15.1.19

[104] **Davoudian, A., Chen, L., & Liu, M.** (2018). A survey on NoSQL stores. *ACM Computing Surveys (CSUR), 51*(2), 1-43. https://doi.org/10.1145/3158661

[105] **Francis, N., Green, A., Guagliardo, P., Libkin, L., Lindaaker, T., Marsault, V., Plantikow, S., Rydberg, M., Selmer, P., & Taylor, A.** (2018, May). Cypher: An evolving query language for property graphs. In *Proceedings of the 2018 international conference on management of data* (pp. 1433-1445). https://doi.org/10.1145/3183713.3190657

[106] **M. Needham ve A. E. Hodler.** (2019). *Graph Algorithms: Practical Examples in Apache Spark & Neo4j*, First Edition. Sebastopol, CA: O'Reilly Media, Inc.

[107] **Robinson, I., Webber, J., & Eifrem, E.** (2015). Graph databases: new opportunities for connected data. O'Reilly Media, Inc.

[108] **Vukotic, A., Watt, N., Abedrabbo, T., Fox, D., & Partner, J.** (2014). *Neo4j in action.* Manning Publications Co.

[109] **M. Mahmoud, M. Mannan, ve A. Youssef.** (2023). APTHunter: Detecting Advanced Persistent Threats in Early Stages", *Digital Threats: Research and Practice*, *4*(1), 1-31, https://doi.org/10.1145/3559768

[110] **Noel, S., Harley, E., Tam, K. H., Limiero, M., & Share, M.** (2016). CyGraph: graph-based analytics and visualization for cybersecurity. In *Handbook of statistics* (Vol. 35, pp. 117-167). Elsevier. https://doi.org/10.1016/bs.host.2016.07.001

[111] **Husák, M., Komárková, J., Bou-Harb, E., & Čeleda, P.** (2018). Survey of attack projection, prediction, and forecasting in cyber security. *IEEE Communications Surveys & Tutorials, 21*(1), 640-660. https://doi.org/10.1109/COMST.2018.2871866

[112] **Cai, H., Zheng, V. W., & Chang, K. C. C.** (2018). A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE Transactions on Knowledge and Data Engineering, 30*(9), 1616-1637. https://doi.org/10.1109/TKDE.2018.2807452

[113] **Pourhabibi, T., Ong, K. L., Kam, B. H., & Boo, Y. L.** (2020). Fraud detection: A systematic literature review of graph-based anomaly detection approaches. *Decision Support Systems, 133*, 113303. https://doi.org/10.1016/j.dss.2020.113303

[114] **Das, K., Samanta, S., & Pal, M.** (2018). Study on centrality measures in social networks: a survey. *Social Network Analysis and Mining, 8*, 1-11. https://doi.org/10.1007/s13278-018-0493-2

[115] **Freeman, L. C.** (1977). A set of measures of centrality based on betweenness. *Sociometry,* 40(1), 35-41, https://doi.org/10.2307/3033543

[116] **Peng, W., BaoWen, X., YuRong, W., & XiaoYu, Z.** (2015). Link prediction in social networks: the state-of-the-art. *Science China-Information Sciences, 58*(1), 1-38. https://doi.org/10.1007/s11432-014-5237-y

[117] **Bonacich, P.** (1987). Power and centrality: A family of measures. *American Journal of Sociology, 92*(5), 1170-1182. https://doi.org/10.1086/228631

[118] **Akoglu, L., Tong, H., & Koutra, D.** (2015). Graph based anomaly detection and description: A survey. *Data Mining and Knowledge Discovery, 29*, 626-688. https://doi.org/10.1007/s10618-014-0365-y

[119] **Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E.** (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, 2008*(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

[120] **Reddy, V., Kolli, N., & Balakrishnan, N.** (2021). Malware detection and classification using community detection and social network analysis. *Journal of Computer Virology and Hacking Techniques, 17*(4), 333-346. https://doi.org/10.1007/s11416-021-00387-x

[121] **Newman, M. E., & Girvan, M.** (2004). Finding and evaluating community structure in networks. Physical Review E, *69*(2), 026113. https://doi.org/10.1103/PhysRevE.69.026113

[122] **Cetin, P., & Amrahov, Ş. E.** (2022). A new network-based community detection algorithm for disjoint communities. *Turkish Journal of Electrical Engineering and Computer Sciences, 30*(6), 2190-2205. https://doi.org/10.55730/1300-0632.3933

[123] **Wang, W., Tang, B., Zhu, C., Liu, B., Li, A., & Ding, Z.** (2020, July). Clustering using a similarity measure approach based on semantic analysis of adversary behaviors. In 2020 *IEEE Fifth International Conference on Data Science in Cyberspace (DSC)* (pp. 1-7). IEEE. https://doi.org/10.1109/DSC50466.2020.9194468

[124] **Raghavan, U. N., Albert, R., & Kumara, S.** (2007). Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E, 76*(3), 036106. https://doi.org/10.1103/PhysRevE.76.036106

[125] **Wenhao, S., Dongtao, C., Qian, W., & Li, S.** (2024). Neighborhood relation-based incremental label propagation algorithm for partially labeled hybrid data. *Machine Learning (113),*9. 6293-6339. https://doi.org/10.1007/s10994-024-06560-9

[126] **Cheng, W., Zhu, T., Chen, T., Yuan, Q., Ying, J., Li, H., Xiong, C., Li, M., Lv, M., & Chen, Y.** (2024). CRUcialG: Reconstruct Integrated Attack Scenario Graphs by Cyber Threat Intelligence Reports. arXiv preprint arXiv:2410.11209

[127] **Liu, F., Wen, Y., Zhang, D., Jiang, X., Xing, X., & Meng, D.** (2019, November). Log2vec: A heterogeneous graph embedding based approach for detecting cyber threats within enterprise. In *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security* (pp. 1777-1794). https://doi.org/10.1145/3319535.3363224

[128] **Li, L., Qiang, F., & Ma, L.** (2024, April). Advancing cybersecurity: Graph neural networks in threat intelligence knowledge graphs. In *Proceedings of the International Conference on Algorithms, Software Engineering, and Network Security* (pp. 737-741). https://doi.org/10.1145/3677182.3677314

[129] **Wang, C., & Zhu, H.** (2022). Wrongdoing monitor: A graph-based behavioral anomaly detection in cyber security. *IEEE Transactions on Information Forensics and Security, 17,* 2703-2718. https://doi.org/10.1109/TIFS.2022.3191493

[130] **Gao, X., Xiao, B., Tao, D., & Li, X.** (2010). A survey of graph edit distance. *Pattern Analysis and Applications, 13,* 113-129**.** https://doi.org/10.1007/s10044-008-0141-y

[131] **Kiss, R., & Szűcs, G.** (2024). Unsupervised graph representation learning with inductive shallow node embedding. *Complex & Intelligent Systems, 10*(5), 7333-7348. https://doi.org/10.1007/s40747-024-01545-6

[132] **Yılmaz, A., & Das, R.** (2025). A novel hybrid approach combining GCN and GAT for effective anomaly detection from firewall logs in campus networks. *Computer Networks,* 111082. https://doi.org/10.1016/j.comnet.2025.111082

[133] **Kim, H., Lee, B. S., Shin, W. Y., & Lim, S.** (2022). Graph anomaly detection with graph neural networks: Current status and challenges. *IEEE Access*, *10*, 111820-111829. https://doi.org/10.1109/ACCESS.2022.3211306

[134] **Gao,C., Zheng, Y., Li, N., Li, Y., Qin, Y., Piao, J., Quan,Y., Chang,J., Jin,D., He,X., Li, Y.** (2023). A survey of graph neural networks for recommender systems: Challenges, methods, and directions. *ACM Transactions on Recommender Systems*, *1*(1), 1-51. https://doi.org/10.1145/3568022

[135] **Lin, H., Yan, M., Ye, X., Fan, D., Pan, S., Chen, W., & Xie, Y.** (2022). A comprehensive survey on distributed training of graph neural networks. *arXiv preprint arXiv:2211.05368*

[136] **Adarsh, S., & Zacharias, J.** (2024, June). Handling Missing Data in Graph Networks with Adaptive Graph Convolutional Network. In *2024 International Conference on Advancements in Power, Communication and Intelligent Systems (APCI)* (pp. 1-6). IEEE. https://doi.org/10.1109/APCI61480.2024.10616601

[137] **D. Demirol, R. Das, ve D. Hanbay.** (2019, Eylül). Büyük veri üzerine perspektif bir bakış, içinde *2019 International Artificial Intelligence and Data Processing Symposium (IDAP)*, IEEE, ss. 1-9. https://doi.org/10.1109/IDAP.2019.8875902
[138] **Yuan, H., Yu, H., Gui, S., & Ji, S.** (2022). Explainability in graph neural networks: A taxonomic survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(5), 5782-5799.
[139] **Syed, Z., Padia, A., Mathews, M. L., Finin, T., & Joshi, A.** (2016, February). UCO: A unified cybersecurity ontology. In *Proceedings of the AAAI Workshop on Artificial Intelligence for Cyber Security* (pp. 195-202). http://dblp.uni-trier.de/db/conf/aaai/security2016.html#SyedPFMJ16
[140] **ESET** (2023). WeLiveSecurity. [Çevrimiçi]. Erişim adresi: https://www.welivesecurity.com/
[141] **CyberMonitor.** (2023). APT Cyber Criminal Campaign Collections. [Çevrimiçi]. Erişim adresi: https://github.com/CyberMonitor/APT_CyberCriminal_Campagin_Collections
[142] **M. Fenniak *vd.*** (2022).The PyPDF2 library. [Çevrimiçi]. Erişim adresi: https://pypi.org/project/PyPDF2/
[143] **M. Corporation.** (2024). Mitreattack-python: Python library for interacting with the MITRE ATT&CK framework.
[144] Inc. Fortinet. FortiGuard Labs Threat Research.
[145] **Hooi, E. K. J., Zainal, A., Maarof, M. A., & Kassim, M. N.** (2019, September). TAGraph: knowledge graph of threat actor. In *2019 International Conference on Cybersecurity (ICoCSec)* (pp. 76-80). IEEE.
[146] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 2980-2988).
[147] **Brin, S., & Page, L.** (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems, 30*(1-7), 107-117. https://doi.org/10.1016/S0169-7552(98)00110-X
[148] **Demirol, D., Das, R., & Hanbay, D.** (2025). A Novel Approach for Cyber Threat Analysis Systems Using BERT Model from Cyber Threat Intelligence Data. *Symmetry*, 17(4), 587. https://doi.org/10.3390/sym17040587

ÖZGEÇMİŞ

Ad-Soyad : Doygun DEMİROL

ÖĞRENİM DURUMU:

- **Lisans :** 2010, Fırat Üniversitesi, Teknik Eğitim Fakültesi, Elektronik-Bilgisayar Eğitimi Bölümü, Bilgisayar Öğretmenliği
- **Yüksek Lisans :** 2012, Fırat Üniversitesi, Fen Bilimleri Enstitüsü, Elektronik-Bilgisayar Eğitimi Bölümü
- **Doktora :** (2016 - Devam ediyor.), İnönü Üniversitesi, Bilgisayar Mühendisliği Anabilim Dalı, Donanım Bilim Dalı

MESLEKİ DENEYİM:

- 2014 yılından beri Bingöl Üniversitesi Genç Meslek Yüksekokulu'nda öğretim görevlisi olarak çalışmaktadır.

TEZ SÜRESİNCE TÜRETİLEN BİLİMSEL ÇALIŞMALAR

- **Demirol, D., Das, R., & Hanbay, D. (2019, Eylül).** Büyük veri üzerine perspektif bir bakış. *International Conference on Artificial Intelligence and Data Processing (IDAP)* (pp. 1-9). IEEE. https://doi.org/10.1109/IDAP.2019.8875902
- **Demirol, D., Das, R., & Hanbay, D. (2023).** A key review on security and privacy of big data: issues, challenges, and future research directions. *Signal, Image and Video Processing*, *17* (4), 1335-1343. https://doi.org/10.1007/s11760-022-02341-w
- **Demirol, D., Das, R., & Hanbay, D. (2025).** A Novel Approach for Cyber Threat Analysis Systems Using BERT Model from Cyber Threat Intelligence Data. *Symmetry*, 17(4), 587. https://doi.org/10.3390/sym17040587