Fırat Üniversitesi Müh. Bil. Dergisi
35(1), 283-290, 2023

Araştırma Makalesi
https://doi.org/10.35234/fumbd.1217219

Açık Kaynaklardan Test Otomasyon Araçlarıyla Siber Tehdit İstihbaratı Çıkarılması

Anıl SEZGİN[1,2*], Aytuğ BOYACI[3]
[1] ATASAREN, Milli Savunma Üniversitesi, İstanbul, Türkiye
[2] Siemens, İstanbul, Türkiye
[3] Bilgisayar Mühendisliği Bölümü, Milli Savunma Üniversitesi, Hava Harp Okulu, İstanbul, Türkiye
[*1,2] anilszgn@gmail.com, [3] aboyaci@hho.msu.edu.tr

(Geliş/Received: 11/12/2022; Kabul/Accepted: 08/02/2023)

Öz: Siber tehdit istihbaratı siber tehdit odaklı, yapılandırılmış ve analiz edilmiş bilgiler için kullanılan bir kavramdır. Bu çalışmada sosyal medya ve siber güvenlik siteleri gibi çeşitli açık kaynaklardan, test otomasyon araçları aracılığıyla elde edilen ham verilerin yapısal verilere dönüştürülmesini ve veriye dayalı tehdit analizine hazır, siber güvenlik istihbaratı elde edilmesini sağlayacak bir model önerilmiştir. Toplanan ve yapılandırılmamış olan ham veriler, modelimiz tarafından yapılandırılmış verilere dönüştürülmekte ve siber güvenlik yazılımlarını besleyebilecek standart formlara getirilmektedir.

Anahtar kelimeler: Açık kaynak istihbaratı, siber tehdit istihbaratı, test otomasyon, veri toplama.

Extracting Cyber Threat Intelligence with Test Automation Tools from Open Sources

Abstract: Cyber threat intelligence is a concept used for cyber threat-oriented, structured and analyzed information. In this study, we propose a model to transform raw data obtained from various open sources such as social media and cybersecurity sites through test automation tools into structured data and to obtain cybersecurity intelligence ready for data-driven threat analysis. The collected and unstructured raw data is converted into structured data by our model and standardized into forms that can feed cyber security softwares.

Key words: Open-source intelligence, cyber threat intelligence, test automation, data gathering.

1. Giriş

Açık kaynak istihbaratı (Open Source Intelligence, OSINT) analiz ve değerlendirme gibi bilgi işleme süreçleri aracılığıyla anlamlı istihbarat toplamayı amaçlayan bir süreçtir. Bu süreç, gazete, dergi, sosyal medya siteleri, video paylaşım siteleri, blog ve içerik paylaşım siteleri gibi halka açık bilgileri esas alarak toplanan istihbaratlardan oluşmaktadır. OSINT süreci tarafından oluşturulan istihbaratın iki koşulu sağlaması gerekir. Birincisi, analiz edilen bilgilerin içeriğinin ve kaynağının doğrulanmasıdır. Diğeri ise analiz edilen ve değerlendirilen istihbaratın amacına uygun olarak yararlı ve anlamlı hale getirilmesi gerekliliğidir.

Siber tehdit istihbaratı (Cyber Threat Intelligence, CTI) analiz edilmiş, yapılandırılmış siber tehditler hakkında bilgi için kullanılan bir kavramdır. Kuruluşlar son yıllarda artan siber saldırı tehdidine karşı siber tehdit istihbaratını geliştirebilmek için büyük yatırımlar yapmaya başlamıştır. Bu bilgiler kuruluşlara zarar verebilecek farklı düzeylerdeki mevcut riskleri anlamalarına yardımcı olmak için kullanılmaktadır. Ayrıca bu istihbarat, kuruluşların savunma amaçlı karşı önlemler planlamasına ve kendilerine zarar verebilecek saldırılara karşı korunabilmelerine de yardımcı olabilmektedir. Temelde veri odaklı olan bu süreçte, birçok kuruluş geleneksel olarak güvenlik bilgi ve olay yönetim sistemleri (Security Information and Event Management System, SIEM), günlük dosyaları, ağ saldırı tepit ve önleme sistemleri gibi dahili sistemlerden veri toplayığ analiz ederek ortaya çıkan tehditler ve kilit tehdit aktörleri hakkında tahminler yapmaya çalışmaktadır.

Bir güvenlik sisteminin, altyapıyı doğru ve düzgün bir şekilde izlemek, sürdürmek ve güvenceye almak için zamanında ve ilgili tehdit istihbaratlarını elde etmesi gerekmektedir. Bu durum, güvenlik analistlerini çeşitli bilgi akışlarını toplayarak ve okuyarak tehdit farkındalığı için çalışmaya yöndirmektedir. Ancak bu süreçlerin manuel olarak yapılması, büyük miktarda veri göz önüne alındığından çok fazla kaynak gerektirmektedir.

Siber güvenlik istihbaratı, siber güvenlikle ilişkili verilerin analiz edilmesi ve değerlendirilmesiyle oluşturulur. Siber güvenlik istihbaratı kullanarak siber tehdide karşı etkin bir güvenlik ve savunma stratejisi oluşturmak mümkündür.

* Sorumlu yazar: anilszgn@gmail.com. Yazarların ORCID Numarası: [1,2] 0000-0002-5754-1380, [3] 0000-0003-1016-3439

Siber güvenlik bilgi kaynakları iki grupta incelenebilir. NVD ve US-CERT gibi resmi kaynaklar ve blog, forum, Twitter, Reddit gibi sosyal medya platformları, güvenlik açıkları, tehditler ve saldırılarla ilgili bilgi sağlamaktadır. Günlük olarak çok fazla veri akışı olan bu kaynakların manuel olarak taranması, bilgi çıkarılması ve saldırının gerçekleşebileceği çeşitli senaryoların tasarlanması neredeyse imkansızdır. Açık kaynak istihbaratından elde edilen ilgili bilgilerin otomatik olarak çıkarılması araştırmacıların ilgisini çekmektedir [1] [2].

Kurumlar, şirketler, endüstriyel tesisler siber suç grupları tarafından sıklıkla hedef alınmaktadır. Saldırılar çoğu zaman ana bilgisayara kötü amaçlı yazılım yükleyerek kontrolü ele geçirme üzerine kuruludur. Son yıllarda ise Botnet saldırıları artmıştır. Bu saldırıların büyümesinin arkasındaki nedenin, nesnelerin interneti ve endüstriyel nesnelerin interneti kullanımındaki artış olarak açıklanabilir.

Ağ trafiğini izleyen, tehditler ve ilgili kuralların eşleşmelerini tespit etmek için paket yüklerini analiz eden, anomali tespiti yapan ve uyarılar üreten saldırı tespit sistemleri (intrusion detection system, IDS) ve önleme sistemleri (intrusion prevention system, IPS) kullanılarak saldırılardan korunulması amaçlanmaktadır.

Siber güvenlik alanında, OSINT aracılığıyla sağlanan bilgiler, geleneksel güvenlik sistemleri, IDS ve IPS gibi izleme araçları aracılığıyla elde edilen verileri birleştirerek sistemlerin güncel kalması sağlanabilir. Siber tehdit istihbaratı genellikle tehdit bilgilerini yapılandırılmış veri biçiminde resmi olarak yayınlayan açık kaynaklardan elde edilebilir. Yapılandırılmış tehdit istihbaratı, XML ya da Json şeması gibi ortak biçim ve yapıya sahip bir modele bağlıdır. Bu nedenle, yapılandırılmış siber tehdit istihbaratı, güvenlik tehditlerini uygun şekilde analiz etmek ve bunlara yanıt verebilmek için güvenlik araçları tarafından kolay ayrıştırılabilir. Siber tehdit istihbaratının resmi kaynaklarına örnek olarak Common Vulnerabilities and Exposures (CVE) veritabanı ve National Vulnerability Database (NVD) verilebilir.

Bu çalışma kapsamında siber tehdit istihbaratlarını Twitter sosyal medya ağından ve çeşitli siber güvenlik sitelerinden otomatik olarak toplayan bir sistem önerilmiştir. Sistem, Selenium gibi çeşitli test otomasyon araçlarını kullanarak yapısal olmayan verilerden yapısal veri oluşturmayı amaçlamaktadır. Bu sayede yapısal hale getirilen veriler IDS, IPS gibi siber güvenlik sistemlerinin beslenebilmesi amacıyla kullanılabilecektir.

2. Geçmiş Çalışmalar

Siber tehdit istihbaratı siber güvenlik sorunlarını çözmek için makine öğrenmesi ve büyük veri analitiği kullanmaya odaklanmakta ve araştırma topluluklarında yoğun ilgi görmektedir. [3] çalışmasında siber güvenliğin geleceğinin sorun çözmede veriye dayalı bir yaklaşımı benimsemekte yattığı öne sürülmüştür. Bu çalışmada şu konulara yönelik sorunlar ve yaklaşımlar açıklanmıştır: i) saldırı tespiti için veriye dayalı bilim, veri güvenilirliği ve politika esas paylaşımın temelleri ve iii) güvenlik metrikerine risk esaslı bir yaklaşım.

Sosyal medya siteleri açık kaynak istihbaratı için önemli bir kaynak haline gelmiştir. Sosyal medya sitelerine ait veriler, araştırmacılar tarafından doğal afetler [4] [5], terör saldırıları [6], hükümet seçimleri [7], borsa tahmini [8] gibi alanlarda istihbarat toplamak için kullanılmıştır. Bu alanların yanı sıra siber tehditlerle ilgili bilgilerin güvenlik amacıyla kullanılmasına yönelik çalışmalar büyük ilgi görmüştür. Siber tehdit istihbaratını Twitter'dan otomatik olarak toplamak için çeşitli yöntemler kullanılmıştır [9] [10].

Siber güvenlik tehditleriyle ilgili tweet'leri toplamak için geleneksel yöntemlerden biri CVE tanımlayıcısını içeren tweet'leri aramaktır. [11] çalışmasında bu yöntem kullanılarak veriler toplanmış ve gerçek dünyadaki siber exploit için tahminler üretilmiştir. Exploit dedektörleri, modelin kesinliğini arttırmak ve erken istismar uyarıları oluşturmak için toplanan siber tehdit tweet'lerini kullanmıştır. Ancak CVE tanımlayıcısıını içermeyen tweet'ler göz ardı edildiğinden, içinde potansiyel exploitlerin bulunduğu tehdit tweet'leri sürece dahil edilememiştir.

Sosyal medya üzerindeki veriler yüksek hacimli verilerdir ve bu nedenle anlamlı bilgi çıkarabilmek için yüksek hacimli, akan verilerin işlenebilmesi gerekmektedir. [12] çalışmasında SONAR adı verilen bir model önerilmiştir. Bu model sayesinde, Twitter akışı üzerinden veri çekebilen, coğrafi olarak konumlandırabilen ve kategorize edebilen bir framework geliştirilmiştir.

Sayıları giderek artan saldırı olaylarıyla başa çıkma çabaları ve siber tehdit istihbaratı bilgilerini ücretsiz olarak dağıtan çeşitli kanalların ortaya çıkması, standart format ve protokolün oluşturulmasıyla sonuçlanmıştır. [13] çalışmasında çeşitli kanallardan toplanan OSINT bilgilerini graph veritabanına yüklerken paylaşılabilen siber tehdit istihbaratının standartlaştırılmış biçime dayalı bir yönetim yapısı ve yöntem önerilmektedir.

Günümüzde siber tehditlerin sayısı sürekli artmakta ve saldırı teknikleri giderek daha gelişmiş ve akıllı hale gelmektedir. Bu durumla ilgili dikkat edilmesi gereken önemli bir husus, bir siber saldırı için aynı IP, domain ve zararlı kodu kullanan benzer siber olaylardaki belirgin artıştır. Bu nedenle benzer siber saldırıları tespit etmek ve anında yanıt vermek için aynı saldırı altyapısının farklı siber saldırılar için yeniden kullanılması nedeniyle meydana gelen siber saldırılar arasındaki ilişkiyi anlamak önemlidir. Siber saldırılar arasındaki ilişkiyi anlamak için siber saldırıların yöntem ve teknikleri ile ilgili verilerin toplanabilmesi gerekmektedir. [14] çalışmasında bu

ihtiyaçlara yönelik siber tehdit istihbaratı toplama sistemi önerilmektedir. Önerilen sistem, saldırı altyapısı verilerini çeşitli açık kaynaklardan toplamaktadır. Önerilen sistem bir sanallaştırma yapısı ve dağıtık işleme teknolojisi kullanmaktadır.

Periyodik olarak güncellenen tehdit bilgisi sağlayan ve çeşitli analitik çözümleri beslemek için kullanılan çok sayıda OSING kaynağı mevcuttur. Bu noktada hem yapılandırılmış hem de yapılandırılmamış kaynaklardan büyük hacimli veri üretilmektedir. [15] çalışmasında yapılandırılmamış siber güvenlik bilgi kaynaklarından yaklaşık %70 hassasiyetle tehdit akışlarını çıkarabilmek için doğal dil işleme kullanılmıştır. Önerilen model, yaygın olarak kabul gören bir endüstri standardı olan STIX standardında kapsamlı tehdit raporları oluşturabilmektedir.

Açık kaynaklı siber tehdit istihbaratı madenciliği, siber güvenlik profesyonlleri için siber tehditleri hızlı bir şekilde anlama ve zamanında öneylici tedbirler alma konusunda önemli bir rol oynamaktadır. [16] çalışmasında açık kaynaklı siber tehdit istihbaratını araştırmak için bir yaklaşım önerilmiştir. Yaklaşım arama motorlarını kullanarak açık kaynaklı veri beslemesi sağlamaktadır. Arama motorlarından elde edilen sonuçlar analiz edilirken, bu durum ikili sınıflandırma problemine indirgenmektedir.

Veri madenciliği teknikleriyle siber tehdit istihbaratı oluşturulmasını sağlayacak bir modelin önerildiği [17] çalışmasında ağ trafiği kayıtları öğrenilen saldırı tiplerine göre sınıflandırılarak siber tehdit istihbaratları standart formatta otomatik olarak üretilebilmektedir. Sistem, bilinmeyen saldırıların uzman görüşü ile tespit edilmesini sağlayarak eğitim setini yeni saldırı tipleri ile güncelleyebilmektedir. Sistemin başarımlarını doğrulamak için literatürdeki çalışmaların sonuçlar ve Weka aracı ile elde edilen doğruluğun önerilen sistemin sonuçları ile benzer olduğu gösterilmiştir.

Bilgi sistemlerini ve kişisel bilgileri siber tehditlere karşı koruyabilmek için [18] çalışmasında dağıtık ve otonom bir sistem inşa ederek Web üzerindeki kaynaklardan veriler toplanmıştır. Önerilen model 3 alt sistemden oluşmaktadır. Web alanı, otonom işbirlikçi Web tarayıcısı ve Crawler. Geliştirilen Crawler için bir sanallaştırma mimarisi kullanılarak kötü niyetli Web sitelerinin algılanabilmesi için dinamik yeniden yapılandırma gerçekleştirilebilmektedir.

Çeşitli açık kaynaklardan alınan verilere dayalı olarak tehdit analizi yapılan [19] çalışmasında geliştirilen analitik motoru için yapılandırılmamış verilerin Regex tabanlı sınıflandırmasına yönelik bir token ayrıştırma tekniği önerilmiştir. Geliştirilen motor açık kaynaklardan zaman serileri halinde verileri tarayıp getirmekte, verileri analiz etmekte ve getirilen parametreye zaman bilgisi de eklenerek kullanıcıya anlamlı bir bilgi sunulabilmektedir. Toplanan ve yapılandırılmamış olarak görünen veriler, motor tarafından yapılandırılmış bir veri olarak görünecek şekilde dönüştürülmekte ve ardından veritabanına eklenmektedir. Analiz motoru tehdit verilerini modelleyerek analiz etmektedir. Ancak buradaki zorluk analiz için kullanışlı yapılandırılmış bir veriye sahip olmaktır.

Açık, sosyal ve karanlık web son yıllarda uygun araçlar ve yöntemler kullanıldığında taranabilecek ve istihbarata dönüştürülebilecek değerli siber güvenlik bilgilerinin zengin kaynakları olarak görülmektedir. [20] çalışmasında bilgi toplama görevine odaklanılmış olup kaynak olarak açık web'deki güvenlik web siteleri, sosyal web'deki güvenlik forumları ve karanlık web'deki hacker forumları/pazarları kullanılmıştır. Önerilen mimari, veri toplama için iki aşamalı bir yaklaşımdan oluşmaktadır. Birinci aşamada veri toplama için makine öğrenmesi tabanlı bir crawler kullanılırken ikinci aşamada toplanan veriler üzerinde istatistiksel dil modelleme teknikleri kullanılmıştır. Önerilen modelde sadece açık kaynaklı araçlar tercih edilmiştir.

Temelde veri odaklı bir süreç olan siber tehdit istihbaratında birçok kuruluş geleneksel olarak günlük kayıtlarından veri toplayıp analiz ederek reaktif istihbarat elde etmektedir. Web'de bulunan siber güvenlik topluluklarının yaptıkları paylaşımlar, kuruluşları daha önce farkında olmadıkları tehditlere karşı uyararak önemli bir proaktif istihbarat değeri sunabilir. Çeşitli platformlar arasında forumlar zengin meta veriler sağlayabilir. [21] çalışmasının amacı saldırganların kullandıkları yöntem ve teknikleri sürekli olarak toplamak için tarama karşıtı önlemleri geçebilecek ve bu yöntemleri önceden tanımlanmış kategorilere göre otomatik olarak sınıflandırmak için derin öğrenme yöntemleri kullanılarak gerçek zamanlı siber tehdit istihbaratını oluşturabilecek bir altyapı sunmaktır.

Siber tehdit ortamı gittikçe daha karmaşık ve polimorfik hale geldikçe, saldırganı ve çalışma şeklini anlamak daha kritik hale gelmektedir. Siber güvenlik topluluğu, savunma bileşenlerini bilgilendirmek için teknik göstergeleri tanımlama ve paylaşma konusunda belirli bir olgunluk geliştirmiş olsa da, tehdit aktörü bağlamı gibi tek tip olmayan, yapılandırılmamış ve belirsiz üst düzey bilgilerle hala mücadele ediyoruz, bu da daha bağlamsal, doğru ve ilgili istihbarat elde etmek için farklı kaynaklarla korelasyon kurma yeteneğimizi sınırlıyor. Siber tehdit istihbaratı üretme ve daha iyi operasyonel hale getirme kabiliyetini arttırmak için bu sınırlamanın üstesinden gelinmesi gerekmektedir. [22] çalışmasında tehidt aktörlerini ve faaliyetlerini karakterize etmek için üzerinde mutabık kalınan kontrollü kelime dağarcıklarının siber tehdit istihbaratını zenginleştmek ve açıklanabilir, sorgulanabilir daha yüksek bir bağlamsal düzeyde yeni bilgiler çıkarmak için nasıl kullanılabileceğini göstermektir. Özellikle, tehdit aktörlerinin türlerini kişiliklerine dayalı olarak otomatik olarak çıkarmak, doğalarını

anlamak ve zaman içinde davranış ve özelliklerindeki çok biçimliliği ve değişiklikleri yakalamak için ontolojik bir yaklaşım sunulmuştur. Böyle bir yaklaşım, yüksek düzeyde bağlamsal siber tehdit istihbaratını paylaşmak için yapılandırılmış bir yol ve araç sağlayarak birlikte çalışabilirliği mümkün kılmakla kalmaz, aynı zamanda makine hızında yeni bilgiler türetir ve manuel sınıflandırma yaklaşımlarının gerektirdiği bilişsel önyargıları en aza indirir.

3. Sistem Tasarımı

Bu çalışma, açık kaynak verileri test otomasyon araçlarıyla toplayarak gerçek zamanlı siber güvenlik sistemlerinin tasarlanması ve uygulanmasına yönelik mevcut yöntemleri genişleterek ve geliştirerek mevcut bilgi tabanına katkıda bulunmaktadır. Bu yaklaşım, tehdit aktörleri için tercih edilebilecek bir dizi kaynağa görünürlük sağlamamıza ve sıfır gün güvenlik açıkları dahil olmak üzere siber güvenlik istihbaratını zamanında belirlememize olanak tanımaktadır.

Çalışmamız OSINT kaynakları üzerinden test otomasyon araçları ile siber güvenlik tehdit istihbaratının yapılabilmesi için veri toplamaya odaklanmaktadır. Çalışma kapsamında geliştirdiğimiz framework, yapısal ve yapısal olmayan açık kaynaklardan toplanan verileri siber güvenlik araçlarının beslenmesini sağlayabilecek standart formatlarda çıktı sunmaktadır.

Bu çalışma ile çözülmeye çalışılan problemlerden biri yapısal olmayan veri paylaşımı yapan açık kaynaklardan düzenli bir şekilde verilerin çekilebilmesi ve yapısal hale getirilebilmesidir. Siber güvenlik alanında çok fazla paylaşım yapan kaynak bulunmaktadır bu yüzden yayınlanan verilerin türleri birbirinden farklıdır ve yeni siber saldırı terimleri ortaya çıktıkça bu durum tespitin doğruluğunu etkilemektedir. Çalışmamızda çeşitli siber güvenlik istihbarat kaynakları için şeffaf bir şekilde tarama altyapısı sağlayabilen yeni bir mimari sunulmaktadır. Bu mimari açık kaynak taraması ve düzenli ifade (regular expresion, regex) tabanlı filtrelemenin bir kombinasyonuna başvurularak gerçekleştirilmektedir. Önerilen mimari: toplama (gathering), işleme (processing) ve yayınlama (publishing) olmak üzere 3 temel modülden oluşmaktadır.

Toplama modülünde veriler ham olarak toplanmakta ve bir veritabanında saklanmaktadır. Her bir açık kaynak için geliştirilen servisler sayesinde sosyal medya, blog, siber güvenlik haber siteleri üzerinden veriler toplanmaktadır. Her bir veri kaynağı için tahsis edilmiş servis bulunmaktadır. Bu sayede aynı anda birçok kaynak üzerinden kesintisiz olarak veri akışı gerçekleştirilebilmektedir. Veri toplama modülü ile sadece ham veri toplamaya odaklanılmasının en önemli sebebi büyük hacimli verilerin hızlı ve kayıpsız bir şekilde kaydedilmek istenmesidir. Özellikle sosyal medyadan gelen verilerin hacmi oldukça büyüktür. Bu modül sayesinde açık kaynak verileri hızlı, kayıpsız ve kesintisiz bir şekilde veritabanına aktarılmaktadır. Eş zamanlı olarak veri işleme modülü bu ham verileri çekerek analiz etmekte ve elde edilen çıktıları yapısal verilerin bulunduğu veritabanına aktarmaktadır. Bu sayede farklı kaynaklardan toplanan ve bir standart forma sahip olmayan veriler hem karşılaştırılabilmekte hem de tekrar eden ya da hatalı olabilecek veriler elenmektedir.

Siber güvenlik alanında veri toplanabilecek kaynaklar kimlik doğrulama, kısıtlama ve gizleme gibi tarama karşıtı önlemler kullanabilmektedirler. Bunun yanı sıra başta sosyal medya siteleri olmak üzere birçok kaynak, platformlardaki veriyi uygulama programlama arayüzleri (Application Programming Interface, API) aracılığıyla sunmaktadır. Ancak çoğunlukla ücretli olarak sunulan API aracılığıyla elde edilen verilerde çeşitli filtreler ve kısıtlamalar mevcut olabilmektedir.

Web kazıma (Web Scraping, Web Data Extraction ya da Web Harvesting), web sitelerinden verileri çıkarmak için kullanılan bir tekniktir. Bu yöntem genellikle web üzerindeki yapılandırılmamış veya büyük verilerin (HTML/XML belgeleri) kullanıcı sorgusuna göre organize bilgilere dönüştürülmesine odaklanır. Web kazıma süreci 2 aşamada incelenebilir:

i) Web kaynaklarının elde edilmesi

ii) Elde edilen kaynaklardan ilgili verilerin çıkarılması

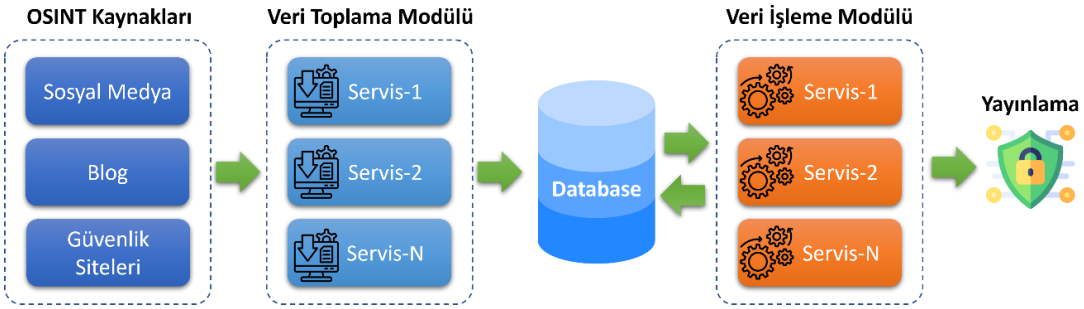
Geleneksel olarak Web kazıma yazılımları HTTP protokolünü yürüterek veya gelişmiş internet tarayıcısını kullanarak bir kullanıcının Web sayfası üzerindeki hareketlerini simüle eder. Web kazıma, bir bot veya Web tarayıcısı kullanarak Web üzerindeki verileri kaydeden ve çoğu Web arama aracında uygulanan kapsayıcı bir yöntem olan web indeksleme süreçlerinde tercih edilmektedir.

Çalışma kapsamında önerdiğimiz modelin ilk modülü hem Web Scraping hem de Web Crawling stratejilerini kullanarak HTML DOM tabanlı mimari aracılığıyla sosyal medya ve siber güvenlik sitelerinden veri çekilmesine dayanmaktadır.

Modelimizin geliştirilmesinde programlama dili olarak Python tercih edilmiş olup, veri toplama için Beautiful Soup 3 ve Selenium kütüphaneleri kullanılarak bir test otomasyonu ortamı hazırlanmıştır. Beautiful Soup, Web sayfasının verilerini HTML veya XML biçiminde getiren bir kütüphanedir. İçinden bulunan bir ayrıştırıcı ile,

DOM yaklaşımına dayalı bir ayrıştırma ağacı oluşturur ve ardından belirli bir etiketi, dizgiyi, öznitelikleri bulmak için farklı filtre işlevleri uygulanabilir. Selenium, test otomasyonları için tercih edilen bir modüldür. Google, Firefox vb. birçok tarayıcıyı desteklemektedir. Web uygulamalarını test etmek için birçok test işlemi sağlayabilmektedir. Selenium doğası gereği dinamik olan web sayfalarını destekleyen bir web sürücüdür, yani sayfanın kendisi yeniden yüklenmeden öğeleri değişebilen bir sayfanın o anki öğelerine erişim sağlayabilmektedir. Bu özelliği sayesinde Web tabanlı uygulamalarda test otomasyonu geliştirme amaçlı kullanılan bir üründür.

Test otomasyon ve Web kazıma araçları kullanılarak açık kaynak istihbaratından siber güvenlik istihbaratının çıkarılmasından sorumlu modelimizin çalışma akışı Şekil 1'de gösterilmektedir.



Şekil 1. Çalışma akışı

4. Sonuç

Siber tehdit istihbaratı, kuruluşların siber tehditleri anlamasına, öngörmesine ve bunlara karşı savunma yapmasına yardımcı olan bilgileri ifade etmektedir. Siber tehdit istihbaratının önemli bir kaynağı internet siteleri ve sosyal medyadır. Test otomasyon araçları, görevleri bir bilgisayarda veya başka bir cihazda otomatik olarak gerçekletirebilen araçlardır. Bu araçlar, veri madenciliği, web kazıma ve web sayfalarıyla etkileşim dahil olmaz üzere çok çeşitli aktiviteleri otomatikleştirmek için kullanılabilir.

Genellikle, Web sitelerinin büyük bir yüzdesi, sunucularına basit bir http isteği yapıldığında başarılı bir yanıt verir, ancak birçoğunun gerçek bir istek ile bir bot arasında ayrım yapan mekanizmaları vardır. Web kazıma, çeşitli web sitelerinden büyük miktarda veri çıkarmak için kullanılan etkili bir tekniktir. Bu veriler daha sonra herhangi bir dosyada ya da bir veritabanında saklanabilir. Web kazıma, Web sayfasının kodunun çıkarılmasını içermektedir. Modelimiz ile veri çıkarılması aşamasında öntanımlı olarak girilen OSINT kaynakları bulunmaktadır. Bu kaynaklar sosyal medya hesaplarından ve siber güvenlik sitelerinden oluşmaktadır. Bu bilgiler veri toplama modülüne parametre olarak verilmektedir. Modüle gelen adrese, Selenium web sürücüsü aracılığıyla gidilir ve her web sayfasındaki veriler için kazıma işlemi gerçekleştirilir. Kazılan tüm veriler ham olarak, zaman ve kaynak bilgisi ile veritabanına kaydedilir. Tüm bu işlemler her bir OSINT kaynağı için tanımlanan farklı servisler ile eşzamanlı olarak yapılmaktadır.

Veritabanına kaydedilen verilerin yapısal hale getirilebilmesi için çeşitli veri ön işleme teknikleri uygulanmış ve düzenli ifadeler (Regular Expression, regex) ve örüntüler kullanılarak ham verilerin içinde istenilen verilerin olup olmadığı kontrol edilerek bu verinin ayrıştırılması sağlanmaktadır.

Örüntülerin tanımlanmasında uygulama isimleri ve uzantıları, paylaşılan URL ve IP verileri, CVE kodları (Common Vulnerabilities and Exposures), öntanımlı saldırı tekniklerinin ve illegal aktivitelerin isimleri kullanılmaktadır. Bu örüntülere uyan verilerin, ham veriler içinden ayıklanıp yapısal veriler haline getirilerek veritabanına kaydedilmesi sağlanmaktadır. Her bir OSINT kaynağından elde edilen ham verilerin o kaynağa tanımlı olarak bir servis ayağa kalkmaktadır. Bu sayede her bir kaynak için paralel olarak servisler koşabilmekte ve eşzamanlı olarak örüntüler yakalanabilmektedir. Veri işleme servisleri ile yapısal hale getirilen verilerin örnek JSON ve XML çıktıları Şekil 2, Şekil 3 ve Şekil 4'te gösterilmektedir.

```
{
  "Source": "SocialMedia",
  "SourceDetail": "Twitter",
  "Channel": "MalwareHunterTeam",
  "Type": "Malware",
  "Platform": "Android",
  "FileName": "ChatService_master.apk",
  "Date": "2022-11-22 22:46:06.000"
}
```


Şekil 2. Twitter üzerinden çıkarılan örnek Malware verisinin JSON çıktısı


```
{
  "Source": "SocialMedia",
  "SourceDetail": "Twitter",
  "Channel": "CXSecurity",
  "Type": "Exploit",
  "Title": "ZTE ZXHN-H108NS Authentication Bypass",
  "DetailPage": "https: //cxsecurity.com/issue/WLB-2022110035,
  "Date": "2022-11-22 22:46:06.000"
}
```


Şekil 3. Twitter üzerinden çıkarılan örnek Exploit verisinin JSON çıktısı


```
<Data>
    <Source>Website</Source>
    <SourcePage>https://www.exploit-db.com</SourcePage>
    <Type>Exploit</Type>
    <CVE>2022-37661</CVE>
    <ExploitInfo>SmartRG Router SR510n 2.6.13 - Remote Code
        Execution</ExploitInfo>
    <ExploitType>Remote</ExploitType>
    <Platform>Hardware</Platform>
    <Date>2022-11-11 00:00:00.000</Date>
</Data>
```


Şekil 4. Exploit-db üzerinden çıkarılan örnek Exploit verisinin XML çıktısı

Bu çalışmada önerilen model, yapısı ve işlevleri sayesinde açık kaynaklardan yararlanarak bir siber tehdit bilgi toplama sistemi geliştirilmiştir. Önerilen model bünyesinde 24 sosyal medya hesabından ve 3 siber güvenlik sitesinden veriler toplanmıştır. Veriler, mevcut hesapların ve sitelerin geçmiş kayıtları da dahil olmak üzere yaklaşık 100.000 siber saldırıdan toplanmıştır. Tablo 1'de toplama kanalı türleri, etkilenen platform bilgisi ve saldırı tipleri yer almaktadır.

Tablo 1. Siber saldırı öznitelikleri

Siber Saldırı Öznitelikleri	İçerik
Saldırı Tipi	Malware, Ransomware, Exploit, Phishing
Kaynak	Sosyal Medya, Siber Güvenlik Siteleri
Etkilenen Platform	Windows, MacOS, Linux, Unix, Hardware, OpenBSD, Android, iOS, Solaris, FreeBSD

Açık kaynaklardan yapılan taramalar ile siber tehdit odaklı istihbarat toplayabilen ve yapısal veri olarak standart formatlarda yayın yapabilen modelimiz sayesinde hem siber güvenlik araçlarının güncel veriler ile beslenebilmesi hem de siber tehdit modellemesi yapmakla ilgilenen araştırmacıların yararlanabileceği bir platform geliştirilmesi amaçlanmıştır.

5. Gelecek Çalışmalar

Siber tehdit istihbaratı, kuruluşların kendilerini siber saldırılara karşı korumaları ve bu saldırıların potansiyel etkilerini azaltmaları için kritik öneme sahiptir. Bu istihbaratın temel kaynaklarından biri, Twitter ve bloglar gibi sosyal medya platformlarının yanı sıra halka açık diğer çevrimiçi kaynakları içeren açık Web'dir.

Siber tehdit istihbaratı toplamak için test otomasyon araçlarını kullanmanın çeşitli faydaları vardır. Birincisi, normalde manuel olarak yapılması gereken tekrarlayan görevleri otomatikleştirerek zamandan ve kaynaktan tasarruf edebilirler. İkinci olarak, çok çeşitli kaynakları hızlı ve sürekli olarak tarayarak tehdit ortamının daha kapsamlı ve güncel bir görünümünü sağlayabilirler.

Test otomasyon araçları, bu kaynaklardan istihbaratı verimli ve etkili bir şekilde toplamak için kullanılabilmektedir. Kuruluşlar, ilgili kaynakları belirleyerek, bu kaynakları ayıklamak için komut/script dosyaları yazarak ve toplanan verileri analiz ederek, siber güvenlik stratejilerini ve uygulamalarını bilgilendirmek için siber tehdit istihbaratı toplayabilir. Ancak test otomasyon araçlarını kullanarak açık kaynaklardan istihbarat toplamanın da bazı kısıtlamaları vardır. Verileri toplamak için kullanılan kaynaklar ve anahtar kelimeler yeterli genişlikte bir istihbarat ağını temsil etmeyebileceğinden toplanan verilerin doğruluğunun teyit edilmesi gerekebilir. Açık kaynaklardan toplanan verilerin güvenilirliği ve doğruluğu değişebileceğinden kuruluşlar topladıkları istihbaratın doğruluğunu arttırmak için adımlar atmalıdır.

Gelecekte test otomasyon araçlarını kullanarak açık kaynaklardan tehdit istihbaratı toplama etkinliğini arttırmak için yeni çalışmalar yapılabilir. Açık kaynakları ayıklamak ve istihbaratı çıkarmak için daha karmaşık algoritmalar ve teknikler geliştirilerek araştırma yapılabilir. Ayrıca toplanan verilerdeki yanlılığı azaltmanın ve istihbaratın güvenilirliğini ve doğruluğunu arttırmanın yolları üzerine araştırma yapılabilir. Siber tehditlere ilişkin daha kapsamlı bir görüş sağlamak üzere test otomasyon araçlarını HUMINT ve SIGINT gibi farklı kaynaklarla entegre ederek yeni çalışmalar yapılabilir.

Teşekkür

Bu çalışma Milli Savunma Üniversitesi, ATASAREN bünyesindeki doktora tezinin bir parçasıdır. Çalışmanın ortaya çıkmasında verdiği destek için Siemens Türkiye'ye teşekkür ederiz.

Kaynaklar

[1] F. Neri and P. Geraci, "Mining Textual Data to Boost Information Access in OSINT," in International Conference Information Visualisation, 2009.

[2] P. Maciolek and G. Dobrowolski, "Cluo: Web-Scale Text Mining System For Open Source Intelligence Purposes," Computer, vol. 14, no. 1, pp. 45-62, 2013.

[3] B. M. Thuraisingham, M. Kantarcıoğlu, K. W. Hamlen and L. Khan, "A Data Driven Approach for the Science of Cyber Security: Challenges and Directions," in International Conference on Information Reuse and Integration (IRI), 2016.

[4] S. Yamada, K. Utsu and O. Uchida, "An Analysis of Tweets During the 2018 Osaka North Earthquake in Japan -A Brief Report," in International Conference on Information and Communication Technologies for Disaster Management (ICT-DM), 2018.

[5] B. Shah, V. Agarwal, U. Dubey and S. Correia, "Twitter Analysis for Disaster Management," in International Conference on Computing Communication Control and Automation (ICCUBEA), 2018.

[6] P. Garg, H. Garg and V. Ranga, "Sentiment analysis of the Uri terror attack using Twitter," in International Conference on Computing, Communication and Automation (ICCCA), 2017.

[7] J. Wang and J. Q. Gan, "Prediction of the 2017 French election based on Twitter data analysis," in Computer Science and Electronic Engineering (CEEC), 2017.

[8] D. S. A. Fernandes, M. G. C. Fernandes, G. A. Borges and F. A. A. M. N. Soares, "Decision-Making Simulator for Buying and Selling Stock Market Shares Based on Twitter Indicators and Technical Analysis," in International Conference on Systems, Man and Cybernetics (SMC), 2019.

[9] N. Dionísio, F. Alves, P. M. Ferreira and A. N. Bessani, "Cyberthreat Detection from Twitter using Deep Neural Networks," in International Joint Conference on Neural Networks (IJCNN), 2019.

[10] F. Alves, A. Bettini, P. Ferreira and A. N. Bessani, "Processing tweets for cybersecurity threat awareness," Information Systems, vol. 95, 2021.
[11] C. Sabottke, O. Suciu and T. Dumitraş, "Vulnerability disclosure in the age of social media: exploiting twitter for predicting real-world exploits," in USENIX Conference on Security Symposium, 2015.
[12] Q. L. Sceller, E. B. Karbab, M. Debbabi and F. Iqbal, "SONAR: Automatic Detection of Cyber Security Events over the Twitter Stream," in International Conference on Availability, Reliability and Security, 2017.
[13] S. Lee, H. Cho, N. Kim, B. Kim and J. Park, "Managing Cyber Threat Intelligence in a Graph Database: Methods of Analyzing Intrusion Sets, Threat Actors, and Campaigns," in International Conference on Platform Technology and Service (PlatCon), 2018.
[14] M. Kim, S. Lee, B. Cho, -I. Kim and M. Jun, "Design of a Cyber Threat Information Collection System for Cyber Attack Correlation," in International Conference on Platform Technology and Service (PlatCon), 2018.
[15] Y. Ghazi, Z. Anwar, R. Mumtaz, S. Saleem and A. Tahir, "A Supervised Machine Learning Based Approach for Automatically Extracting High-Level Threat Intelligence from Unstructured Sources," in International Conference on Frontiers of Information Technology (FIT), 2018.
[16] P. Zhang, J. Ya, T. Liu and J. Shi, "Mining Open-Source Cyber Threat Intelligence with Distant Supervision from the Web," in International Conference on Data Science in Cyberspace (DSC), 2021.
[17] S. M. Arıkan and S. Acar, "A Data Mining Based System for Automating Creation of Cyber Threat Intelligence," in International Symposium on Digital Forensics and Security (ISDFS), 2021.
[18] Y. Kawano and E. Nunohiro, "A Proposal of Distributed Autonomous Cooperative System about Exclusive Web Crawling for Cyber Security," in International Conference on Network-Based Information Systems (NBiS), 2016.
[19] M. H. Mohd Pakhari, N. Jamil, M. E. Rusli and A. A. Abdul Rahim, "Implementation of Token Parsing Technique for Regex Based Classification of Unstructured Data for Cyber Threat Analysis," in International Conference on Information Technology and Multimedia (ICIMU), 2020.
[20] P. Koloveas, T. Chantzios, C. Tryfonopoulos and S. Skiadopoulos, "A Crawler Architecture for Harvesting the Clear, Social, and Dark Web for IoT-Related Cyber-Threat Intelligence," in IEEE World Congress on Services (SERVICES), 2019.
[21] R. Williams, S. Samtani, M. Patton and H. Chen, "Incremental Hacker Forum Exploit Collection and Classification for Proactive Cyber Threat Intelligence: An Exploratory Study," in IEEE International Conference on Intelligence and Security Informatics (ISI), 2018.
[22] V. Mavroeidis, R. Hohimer, T. Casey and A. Jesang, "Threat Actor Type Inference and Characterization within Cyber Threat Intelligence," 2021 13th International Conference on Cyber Conflict (CyCon), 2021, pp. 327-352.