Kimlik Avı Websitesi Sınıflandırma Veri Kümesi

Kişiler:

Rami M. Mohammad, Fadi Thabtah, Lee McCluskey

Amaç:

Bu çalışmanın amacı kimlik avcılığı web sitelerini makine öğrenmesi ile tespit etmektir. Kimlik avcılığı sitelerini öngörmede sağlıklı ve etkili olduğu kanıtlanan önemli özellikleri kullandık ve üzerine kendi özelliklerimizi de ekledik. Bunun sonucunda 30 farklı özellik eklenip bunlara göre eğitim yapılmıştır. Özellik kategorileri ve kısa açıklamaları tablo 1'de verilmiştir.

Tablo 1. Özellik kümelerinden bazıları

Özellik Kümesi	Açıklama
IP Adresini Kullanma	Bir IP adresi, URL'deki "http://125.98.3.123/fake.html" gibi bir alan adının alternatifi olarak kullanılıyorsa, kullanıcılar birinin kişisel bilgilerini çalmaya çalıştığından emin olabilirler.
Şüpheli Kısmı Gizlemek için Uzun URL	URL'nin uzunluğu 54 karakterden büyük veya ona eşitse, URL'nin kimlik avı olarak sınıflandırıldığını göstermiştir. Veri setimizi inceleyerek, 1220 URL uzunluğunu, toplam veri setinin% 48.8'ini oluşturan 54 veya daha fazlasına eşit olduğunu tespit ettik.
URL Kısaltma Hizmetlerinin Kullanılması "TinyURL"	URL kısaltması, bir URL'nin çok daha küçük hale getirilebildiği ve hala gerekli web sayfasına yönlendirilebildiği "World Wide Web" de bir yöntemdir.
URL'de "@" Sembolü Var	URL'deki "@" sembolünün kullanılması, tarayıcının "@" sembolünden önceki her şeyi görmezden gelmesine neden olur ve gerçek adres genellikle "@" sembolünü izler.
"//" kullanarak yönlendiriliyor	URL yolunda "//" varlığı, kullanıcının başka bir web sitesine yönlendirileceği anlamına gelir.
Etki Alanına (-) ile Ayrılmış Önek veya Sonek Ekleme	Kısa çizgi sembolü, yasal URL'lerde nadiren kullanılır. Kimlik avcıları, (-) ile alan adlarına ayrılan önek veya son ekleri eklemeye meyillidirler, böylece kullanıcılar meşru bir web sayfasıyla ilgilendiklerini hissederler.
Alt Etki Alanı ve Çok Alt Etki Alanları	Bir etki alanı adı, örneğimizde "tr" olan ülke kodu en üst düzey etki alanlarını (ccTLD) içerebilir. "edu" kısmı "akademik" için kısaltılmış, birleştirilmiş "edu.tr" ikinci seviye alan adı (SLD) ve "hud" alanın asıl adıdır.

Açılır Pencereyi Kullanma	Kullanıcılardan kişisel bilgilerini bir açılır pencereden göndermelerini isteyen yasal bir web sitesi bulmak olağandışıdır.
Domain Yaşı	Bu özellik WHOIS veritabanından çıkarılabilir (Whois 2005). Çoğu phishing web sitesi kısa bir süre yaşar. Veri setimizi inceleyerek meşru alanın asgari yaşının 6 ay olduğunu tespit ediyoruz.
Google Index	Bu özellik, bir web sitesinin Google'ın dizininde olup olmadığını inceler. Bir site Google tarafından indekslendiğinde, arama sonuçlarında görüntülenir (Webmaster kaynakları, 2014). Genellikle, phishing web sayfalarına kısa bir süre için erişilebilir durumdadır ve bunun sonucunda Google dizininde birçok phishing web sayfası bulunmayabilir.

Diğer özellikler: *HTTPS, Etki Alanı Kayıt Uzunluğu, favicon, Standart Olmayan Bağlantı Noktasını Kullanma, URL'nin Etki Alanı Bölümünde "HTTPS" Sertifikasının Varlığı, URL Yönlendirme, URL of Anchor, <Meta>, <Script> ve <Link> etiketlerindeki bağlantılar, Sunucu Formu İşleyicisi (SFH), E-postaya Bilgi Gönderme, Anormal URL, Web Sitesi Yönlendirme, Durum Çubuğu Özelleştirme, Sağ Tıklamayı Devre Dışı Bırakma, IFrame Yönlendirme, DNS Kaydı, Website Trafiği, Sayfa Puanı, Sayfaya İşaret Eden Bağlantıların Sayısı, İstatistiksel Raporlara Dayalı Özellik*

Tüm özellikler önişleme adımında standartlaştırılarak özellik seçimi yöntemine verilmektedir.

Elde edilen başarı oranları Tablo 2'de verilmiştir. Başarı değerleri KNeighborsClassifier LogisticRegression, GaussianNB, DecisionTreeClassifier yöntemleri kullanılarak elde edilmiştir. Tablo 2'ye göre en yüksek başarı tüm özellik kümelerinin birlikte DecisionTreeClassifier yöntemi kullanıldığı deneyde elde edilmiştir.

Tablo 2. Makine öğrenmesi yöntemlerinin sonuçları

Sınıflandırıcı	**Başarı**	**F1 Skor**
KNeighborsClassifier	0.9439421338155516	0.9494839760999457
LogisticRegression	0.9132007233273056	0.9223300970873787
GaussianNB	0.6009644364074744	0.42931034482758623
DecisionTreeClassifier	0.9656419529837251	0.9690385659967409

Halil DİLAVER 2014010205027
Mustafa Oğuzhan ÖZDEMİR 2014210205006