Makine Öğrenmesi Algoritmaları ile Sahte Twitter Hesapların Tespiti

Detection of Fake Twitter Accounts with Machine Learning Algorithms

İlhan AYDIN
Bilgisayar Mühendisliği Bölümü
Fırat Üniversitesi
Elazığ, Türkiye
iaydın@firat.edu.tr

Mehmet SEVİ
Bilgi İşlem Daire Başkanlığı
Muş Alparslan Üniversitesi
Muş, Türkiye
m.sevi@alparslan.edu.tr

Mehmet Umut SALUR
Bilgisayar Mühendisliği Bölümü
Harran Üniversitesi
Şanlıurfa, Türkiye
umutsalur@harran.edu.tr

***Özet*— Sosyal ağlar günümüzde birçok alanda insan hayatının bir parçası olmuştur. Haberleşme, tanıtım, reklam, haber, gündem oluşturma gibi birçok aktivite sosyal ağlar üzerinden yapılmaya başlanmıştır. Twitter 'da bazı kötü niyetli hesaplar yanlış bilgi ve gündem oluşturma gibi amaçlar için kullanılmaktadır. Bu durum sosyal ağlardaki temel problemlerden biridir. Bu nedenle kötü niyetli hesapların tespit edilmesi önemli olmaktadır. Bu çalışmada insanları yanlış yönlendirebilecek sahte hesapların tespiti için makine öğrenmesi tabanlı yöntemler kullanılmıştır. Bu amaçla oluşturulan veri kümesi ön işlemden geçirilmiş ve makine öğrenmesi algoritmaları tarafından sahte hesaplar tespit edilmiştir. Sahte hesapların tespiti için karar ağacı, lojistik regresyon ve destek vektör makinaları algoritmaları kullanılmıştır. Bu yöntemlerin sınıflandırma başarımları karşılaştırılmıştır ve lojistik regresyonun daha başarılı sonuç verdiği ispatlanmıştır.**

***Anahtar Kelimeler*— Sosyal Ağlar, Twitter, Bot Tespiti, İleri Makine Öğrenmesi, Sınıflandırma.**

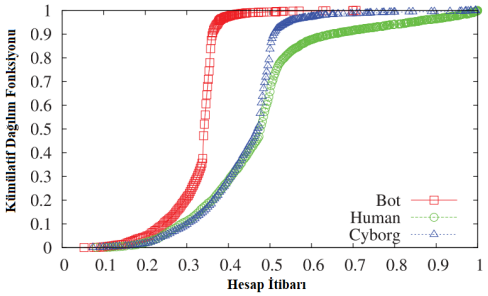
***Abstract* — Today, social networks have been part of many people's lives. Many activities such as communication, promotion, advertisement, news, agenda creation have started to be done through social networks. Some malicious accounts on Twitter are used for purposes such as misinformation and agenda creation. This is one of the basic problems in social networks. Therefore, detection of malicious account is significant. In this study, machine learning-based methods were used to detect fake accounts that could mislead people. For this purpose, the dataset generated was pre-processed and fake accounts were determined by machine learning algorithms. Decision trees, logistic regression and support vector machines algorithms are used for the detection of fake accounts. Classification performances of these methods are compared and the logistic regression proved to be more successful than the others.**

***Index Terms*— Social Networks, Twitter, Bot Detection, Advanced Machine Learning, Classification.**

I. GİRİŞ

Sosyal medyanın yoğun kullanılması, bu ortamların kötü niyetli insanlar tarafından ihmal edilmesini kolaylaştırmıştır. Kurumsal kimliği ile bilinen Twitter sosyal medya uygulaması toplumun her kesimi tarafından kullanılmaktadır. Alexa'ya göre Twitter dünyada en çok ziyaret edilen sitelerin arasında 13. sırada gelmesine rağmen, bu sosyal platformda da kötüye kullanılmalar artmıştır [1]. Twitter'ın uygulamasına farklı platformlardan kolay bir şekilde erişmek mümkündür [2]. Bu da kötü niyetli kullanıcıları daha çok cezbetmektedir. Genellikle kötü niyetli Twitter hesapları daha fazla takipçi edinmek, belli bir topluluğu etkileyerek insanları kendi organizasyonlarına üye yapma, insanları hisse senedi piyasası için manipüle etme, sahte haberleri yayma, özel bilgileri kullanarak insanlara şantaj yapma gibi hedefleri vardır.

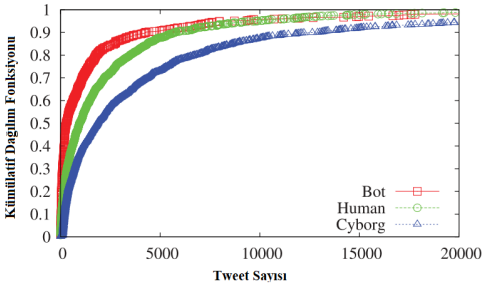
Sosyal ağlarda botlar, insan davranışlarını taklit eden otomatik reaksiyonlar gösteren bilgisayar yazılımlarıdır [3]. Günümüzde aktif Twitter kullanıcıları arasında %15 oranında bot olduğu düşünülmektedir [4]. Cyborg hesaplarda botlara benzemektedir. Bu hesaplar yarı bot yarı insan karakteristiğinde hesaplardır. İnsanlar tarafından açılır fakat bundan sonraki hareketleri bot hesaplara benzemektedir [5].



Şekil 1. Hesap itibarları [6].

978-1-5386-6878-8/18/$31.00 ©2018 IEEE

Şekil 1'de görüldüğü gibi bot ve cyborg hesapların hesap itibarlarının kümülatif dağılım fonksiyonları kıyaslanmaktadır. Hesap itibarı, takipçi sayısının takipçi ve arkadaş sayılarının toplamına oranıdır. Hesap itibarı en geniş olan grup sırasıyla insanlar, cyborglar ve botlardır. Şekil 1'e göre botların %60'ı arkadaş sayısından daha az takipçiye sahiplerdirler [6]. Bot ve cyborg hesapları, tespit edilmemeleri açısından yanlış bilgi veya spam yayması dışında bazen de hava durumu, deprem bilgileri gibi doğru bilgiler sağlamaktadırlar.



Şekil 2. Tweet sayıları [6].

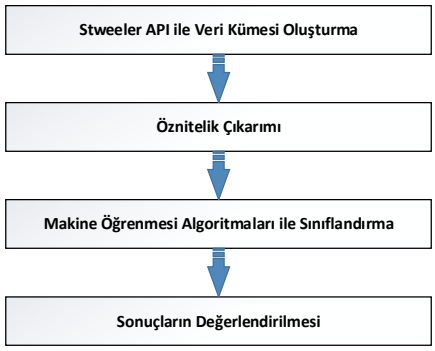
Şekil 2'de hesaplar tarafından atılan tweet sayılarının kümülatif dağılım fonksiyonları kıyaslanmaktadır. Botlar cyborglar ve insanlara göre daha fazla tweet atmışlardır. Cyborg hesaplarının büyük bir kısmı ticari şirketler ve web siteleri tarafından yeni bir medya kanalı ve müşteri hizmeti türü olarak kayıtlıdırlar [7]. Bu hesapların borsa manipülasyonları, şantaj ve yalan bilgi yayma gibi etkileri yüksek orandadır [8]. Sosyal medyanın yaygınlaşmasıyla bot hesapların tespit edilmesinin önemi her geçen gün daha da artmış bulunmaktadır.

Twitter bot-insan-Cyborg tespit edilmesinde Zi ve ark. entropi tabanlı katmanlı bir mimari kullanmışlardır [6]. Twitter botlarının karakteristiklerini dikkate alan ve otomatik öznitelik çıkarımı yapıp hesapları tespit eden yazılım çatıları da geliştirilmiştir [4]. Bir başka çalışmada kullanılan veri kümesinde farklı takipçi sayılarına göre Twitter profilleri ele alınmıştır [8]. Çalışmada çapraz doğrulma metodu sayesinde veri kümesindeki tüm örneklerin hem eğitim hem de test için kullanılmıştır. Ayrıca veri kümesindeki fazla öznitelik çeşitliliği de bot tespitinde önemli olmuştur. Twitter bot tespit edilmesi konusunda yapılan en kapsamlı yarışma 2016 yılı Amerikan Başkanlık seçimleri sırasında düzenlenen "The DARPA Twitter Bot Challenge" yarışmasıdır. Yarışmaya katılan 6 takım 7.038 hesap ve yaklaşık 4 milyon tweet üzerinde çalışmışlardır [9]. Bir başka çalışmada geliştirilen "warped correlation" tekniği ile Twitter'daki bot hesaplarının tespiti yapılmıştır. Bu yöntem ile yıllık 544.868 adet bot hesabı tespit edilmiş ve başarı oranı %94 şeklindedir [10]. Twitter ortamında bot tespiti toplumun güvenliği içinde önemli olmaktadır. Twitter bot tespitinde günümüzde yeni yeni derin öğrenme ağları da kullanılmaya başlanmıştır [11]. Bot hesapların tespit edilmesinde gerek makine öğrenmesi gerek derin öğrenme yöntemleri doğruluk, kesinlik ve F-skoru gibi metrikler kullanmaktadır Fred ve ark. bu metriklerin değerlendirilmesinde farklı bir bakış açısından katkı sağlamışlardır [12].

Bu çalışmada makine öğrenmesi tabanlı Twitter ortamındaki hesapların bot veya insan olarak sınıflandırması uygulaması gerçekleştirilmiştir. Çalışmanın ikinci kısmanda yapılan çalışmadan ve başarım değerlendirme metriklerinden, üçüncü kısımda veri kümesinden ve özniteliklerden ve dördüncü kısımda ise sınıflandırma sonuçlarından bahsedilecektir. Beşinci bölümde de genel çalışmanın sonucu verilecektir.

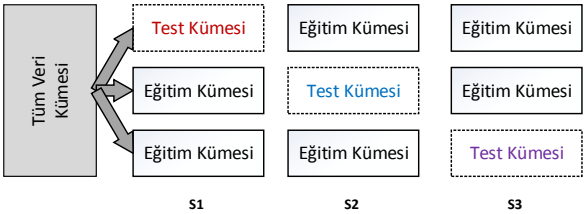
II. YAPILAN ÇALIŞMA VE ALGORİTMA BAŞARIM DEĞERLENDİRME ÖLÇÜTLERİ

Yapılan çalışmada Twitter hesapları ikili sınıflandırma kapsamında insan veya bot olarak sınıflandırılmıştır. Makine öğrenmesi algoritmalarının uygulaması için MATLAB programı kullanılmıştır. Çalışmada makine öğrenmesi sınıflandırma yöntemleri olarak; Karar Ağacı(KA), Lojistik Regresyon(LR) ve Destek Vektör Makine(DVM) algoritmaları kullanılmıştır. Yapılan çalışmanın genel adımları Şekil 3'te verilmiştir.



Şekil 3. Çalışma adımları.

Veri kümesinin sınıflandırılmasında k-kat çapraz doğrulama(K-Fold Cross Validation) yöntemi kullanılmıştır. Çapraz doğrulama sınıflandırma modelin değerlendirilmesi ve eğitilmesinde veri kümesinin parçalara ayırma yöntemlerinden biridir. Bu yöntem ile makine öğrenmesi algoritması veri kümesinin her bir örneğini eğitim ve test için kullanmış olmaktadır.



Şekil 4. Çapraz doğrulama

Şekil 4'te çapraz doğrulama yönteminin görseli bulunmaktadır. Bu yöntemde bütün veri kümesi alt bölümlere ayırılmakta ve her bölüm en az bir kez algoritmanın test edilmesinde kullanılmaktadır. Şekil 4'te görüldüğü üzere veri kümesi üç alt kümeye ayırılmış ve algoritmanın her iterasyonu S1, S2 ve S3 olarak isimlendirilmiştir.

Sınıflandırma algoritmalarının başarımlarının değerlendirilmesinde karışıklık matrisi(confisuon matrix) ve ROC(reciever operator characteristics curve) eğrisi gibi

kriterler kullanmaktadır. Karışıklık matrisi, makine öğrenme algoritmalarında kullanılan sınıflandırma performansını değerlendirmek için hedefe ait tahminlerin ve gerçek değerlerin karşılaştırıldığı bir matristir. Sınıflandırma tahminleri dört adet değerlendirmeden birine sahip olacaktır: gerçek pozitifler, gerçek negatifler, yanlış pozitifler ve yanlış negatifler.

ROC eğrisinde ise gerçek pozitif oranı(hassasiyet), farklı kesme noktaları için yanlış pozitif oranın(100-Özgüllük) fonksiyonunda çizilir. ROC eğrisindeki her nokta belirli bir karar eşiğine karşılık gelen bir duyarlılık ve özgüllük çiftini temsil eder. Mükemmel ayrımcılıkla(iki dağılımda çakışma olmaz) yapılan bir test, sol üst köşeden geçen bir ROC eğrisine sahiptir(% 100 hassasiyet,% 100 özgüllük). Bu nedenle, ROC eğrisinin sol üst köşeye yaklaştıkça, testin genel doğruluğu artar [13]. ROC eğrisi altındaki alan(AUC), bir parametrenin iki grubun ne kadar iyi ayırt edilebildiğinin bir ölçüsüdür. Bu ölçü 1'e ne kadar yakınsa o kadar mükemmeldir.

Sınıflandırma sonuçlarını değerlendirilmesinde karışıklık matrisi dışında kullanılan ölçütler aşağıdaki gibidir;

- Accuracy(isabet oranı): Çalışmadaki temel değerlendirme unsurudur. Toplam doğru tahmin sayının test edilen veriye oranıdır [5].
- Precision(kesinlik): Gerçek değeri pozitif olup pozitif değere sınıflandırılan sayısının, pozitif değere sınıflandırılanların toplamına oranıdır [5].
- Recall(hassasiyet): Gerçek değeri pozitif olup pozitif değere sınıflandırılan sayısının, gerçek değeri pozitif olanların tümüne oranıdır [5].
- F Score: Kesinlik ve hassasiyetin harmonik ortalamasıdır [5].
- ROC eğrisi altındaki alan(AUC), bir parametrenin iki grubun ne kadar iyi ayırt edilebildiğinin bir ölçüsüdür. Bu değer 1'e ne kadar yakınsa o kadar mükemmeldir.

III. VERİ KÜMESİ VE ÖZNİTELİKLER

Özellik çıkarımı işlemi bir boyut işlemidir. Buna göre gereksiz veya fazla olan bir özelliği veri kümesinden çıkararak daha basit bir problem haline indirgenir. Doğru yapılmış bir özellik çıkarımı işlemi sayesinde daha kesin sonuçlara daha hızlı ulaşılmaktadır.

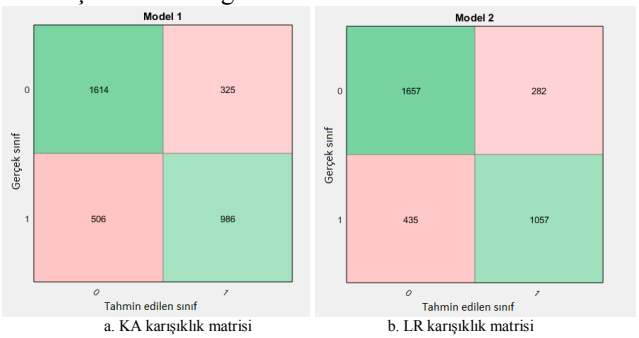
Bu çalışmada kullanmış olduğumuz veri kümesi Zafar Gilani tarafından Twitter'da bot sınıflandırması için Stweeler platformu ile oluşturulmuştur. Veri kümesinde 3431 Twitter hesabına dair çıkarılan öznitelikler bulunmaktadır. Stweeler platformu, Twitter Streaming API'yi kullanarak kullanıcıların bilgilerini toplayan Ruby yazılımıdır [8]. CSV(Comma-Separated Values) formatında bulunan bu veri kümesi MATLAB ortamında algoritmalara giriş olarak verilmiştir. Veri kümesinde bulunan öznitelikler Tablo 1'de verilmiştir [8]. Bu öznitelikler bir hesabın insana veya bota ait olduğunu belirlemede önemli birer ipuçlarıdır. Örneğin; 15 yıllık bir Twitter hesabının bot olması düşük bir ihtimaldir. Çünkü biz biliyoruz ki bot hesapları anlık veya günlük yanlış bilgi sızdırma veya bilgi kötü niyetli kullanma gibi amaçları vardır. Bu nedenle bu hesaplar genellikle gündeme bağlı olarak açıldığından 15 yıllık bir botun olması düşük bir ihtimal barındırmaktadır.

TABLO I. VERİ KÜMESİ ÖZNİTELİKLERİ

Öznitelik	Açıklama
Hesabın yaşı	Hesabın gün cinsinden yaşı. İnsan hesapların daha yaşlı olduğu görülmektedir.
Favorilerin tweetlere oranı	Her hesaptan beğeni ve favori sayıları. İnsan hesapları daha fazla favori almaktadır.
Kullanıcı başına listeler	Abone olunan listelerin sayısı. Botlar genellikle daha fazla listeleri takip ederler.
Takipçilerin arkadaşlara oranı	İnsan hesapların sahip oldukları oranın 1'e yakın olduğunu görülmektedir.
Kullanıcı favorileri	Kullanıcı tarafından tweetleri favori olarak etiketleme sayısı. İnsan hesaplarında daha fazladır.
Tweet başına favori ve beğeni sayısı	Kullanıcı tarafından alınan favorilerin sayısı. Orijinal içerik ve çeşitlilik nedeniyle insan hesapları daha fazla beğeni alırlar.
Tweet başına retweet sayısı	Retweet kullanıcılar tarafından yapılır. Orijinal içerik ve çeşitlilik nedeniyle insan hesaplar daha fazla retweet alır.
Kullanıcı cevapları	Tweetler kullanıcılar tarafından cevaplarının sayısı. İnsan hesapları daha çok cevap yazarlar.
Kullanıcı retweetleri	Kullanıcı tarafından yapılan retweet'lerin sayısı.
Tweet atma sıklığı	Kullanıcının bir günde ne kadar sıklıkla tweet attığını ifade eder. Botlar daha sık tweet atarlar.
Bağlantı sayısı	Twitter mesajları içerisindeki URL'lerin sayısı.
Kaynak sayısı	Kullanıcının kullandığı kaynak sayısı. İnsan hesapları daha fazla kaynak kullanırlar.
Yüklenmiş kaynak miktarı	Twitter yüklenen içerik miktarı. Ortalama bir bot hesabı daha fazla içerik yüklemeye meyillidir.
Kullanıcı tweetleri	Kullanıcı tarafından atılan tweet'lerin sayısı. Botlar daha çok tweet atmaktadırlar.

IV. UYGULAMA SONUÇLARI

Sonuçların değerlendirilmesinde veri kümesindeki gerçek sınıf ile tahmin edilen sınıf sayılarını içermektedir. Sınıflandırma tahminleri dört adet değerlendirmeden birine sahip olacaktır: gerçek pozitifler, gerçek negatifler, yanlış pozitifler ve yanlış negatifler. Veri kümesi KA algoritması tarafından çalıştırıldığında %75,8'lik bir isabet oranı elde edilmiştir. Şekil 5 a.'da KA sınıflandırması sonucunda elde edilen karışıklık matrisi görülmektedir.



a. KA karışıklık matrisi b. LR karışıklık matrisi

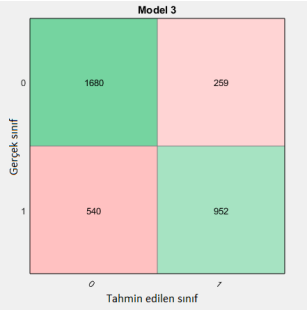
Şekil 5. KA ve LR karışıklık matrisleri.

KA algoritması insan olan 1939 hesaptan 1614 tanesini insan olarak sınıflandırmış ve 325 tanesini de yanlış bir şekilde bot olarak sınıflandırmıştır. Diğer taraftan 1492 tane bot hesaptan doğru bir şekilde 986 tanesini bot olarak etiketlemiş ve 506 tanesini de yanlış bir şekilde insan olarak etiketlemiştir.

Veri kümesine LR algoritması ile sınıflandırıldığında %79,1'lik bir isabet oranı elde edilmiştir. Şekil 5 b.'de LR algoritmasının karışıklık matrisi verilmiştir. Karışıklık matrisi

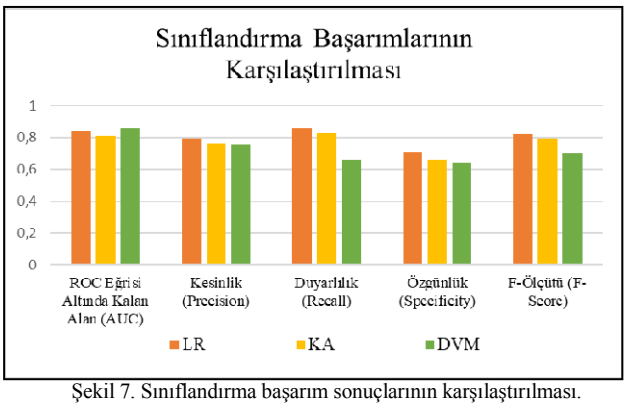
incelendiğinde %79,1 başarı oranı görülmektedir. Yüzdelik dilimler olarak sonuçlar ifade edilirse insan sınıfı için doğru pozitif oranı %85 ve bot sınıfı için %71 diğer taraftan yanlış negatif oranı insan sınıfı için %15, bot sınıfı için ise %29 olarak elde edilmiştir.

Veri kümesine DVM algoritması uygulandığında %76,7'lik bir isabet oranı elde edilmiştir. Şekil 6'da DVM karışıklık matrisi verilmiştir. Karışıklık matrisi incelendiğinde %76,7 başarı oranı elde edildiği görülmektedir.



Şekil 6. DVM karışıklık matrisi.

DVM insan olan 1939 kullanıcının 1680 tanesini insan olarak doğru sınıflandırmış, 259 tanesini ise yanlış bir şekilde bot olarak sınıflandırmıştır. Yüzdelik dilimler olarak sonuçlar ifade edilirse insan sınıfı için doğru pozitif oranı %87, bot sınıfı için %64 diğer taraftan yanlış negatif oranı insan sınıfı için %13, bot sınıfı için ise %36 olarak gözlemlenmiştir.



Şekil 7. Sınıflandırma başarım sonuçlarının karşılaştırılması.

Çalışmada kullanılan üç makine öğrenmesi algoritmasının sınıflandırma sonucundaki başarım ölçütleri Şekil 7'de verilmiştir. Yöntemlerin AUC'lerine bakıldığında iki sınıfı en iyi ayırt eden yöntemin DVM olduğu görülmektedir. Kesinlik ölçütü bakımından LR algoitması en iyi sonucu vermektedir. Onu KA ve DVM takip etmektedir. Kesinlik ölçütünü tek başına yorumlamak yanlış olabilir. Duyarlılık ölçütüne baktığımızda ise yine aynı sıralama görülmektedir. Kesinlik ve duyarlılık ölçütlerini beraber değerlendirmek için, her iki değerin harmonik ortalaması olan F1 ölçütüne baktığımızda yine LR en iyi sonuca sahip olduğu görülmektedir. Sadece isabet oranından farklı olarak KA ve DVM sonucu yer değiştirmektedir. Hassasiyet ölçütüne baktığımızda en yüksek oran LR algoritmasına aittir. Ayrıca en yüksek gerçek negatif rakamı da(FN) 325 ile LR algoritmasındadır. Bu da veri kümesindeki botların tespitinin daha zor olduğunu göstermektedir.

V. SONUÇLAR

Yaşadığımız bilgi çağından doğru ve güvenilir bilgi her geçen gün daha da değerli olmaktadır. Twitter ortamında botların yardımıyla paylaşılan bilgiler birçok kişi için kritik bir öneme sahiptir. Bu çalışmada da Twitter ortamındaki botların tespit edilmesinde makine öğrenmesi algoritmaları kullanılmış ve Twitter kullanıcıları insan veya bot olarak sınıflandırılmıştır. Çalışmada KA, LR ve DVM makine öğrenmesi algoritmaları kullanılmıştır. Algoritmaların başarımlar çeşitli metriklerle değerlendirilmiş ve en yüksek başarımlar LR algoritması ile elde edilmiştir. En yüksek F-1 ölçütü 0,821 olarak LR algoritmasından elde edilmiştir. F-1 ölçütü sırası ile KA için 0,794 ve DVM için 0,704 olarak elde edilmiştir. Çalışmanın devamında derin öğrenme yöntemleriyle bot belirlemesi yapılması planlanmaktadır.

KAYNAKÇA

[1] "Twitter.com Traffic." [URL]. Available: https://www.alexa.com/siteinfo/twitter.com. [Erişim:13-May-2018].

[2] M. U. Salur and İ. Aydin, "Üniversitelerin Paylaştığı Twitter Mesajlarının İnsanlara Erişiminin Bulanık Birliktelik Kuralları ile Değerlendirilmesi," *Harran Üniversitesi Mühendislik Derg.*, vol. 03, no. 02, pp. 25–39, 2017.

[3] M. Shafahi, L. Kempers, and H. Afsarmanesh, "Phishing through social bots on Twitter," in *Proceedings - 2016 IEEE International Conference on Big Data, Big Data 2016*, 2016, pp. 3703–3712.

[4] O. Varol, E. Ferrara, C. A. Davis, F. Menczer, and A. Flammini, "Online Human-Bot Interactions: Detection, Estimation, and Characterization," *arXiv Prepr. arXiv1703.03107*, 2017.

[5] E. Van Der Walt and J. Eloff, "Using Machine Learning to Detect Fake Identities: Bots vs Humans," *IEEE Access*, vol. 6, pp. 6540–6549, 2018.

[6] Z. Chu, S. Gianvecchio, and H. Wang, "Detecting Automation of Twitter Accounts :Are You a Human, Bot, or Cyborg?," *IEEE Trans. DEPENDABLE Secur. Comput.*, vol. 9, no. 6, pp. 811–824, 2012.

[7] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "DNA-Inspired Online Behavioral Modeling and Its Application to Spambot Detection," *IEEE Intell. Syst.*, vol. 31, no. 5, pp. 58–64, 2016.

[8] Z. Gilani, E. Kochmar, and J. Crowcroft, "Classification of Twitter Accounts into Automated Agents and Human Users," *Proc. 2017 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Min. 2017 - ASONAM '17*, pp. 489–496, 2017.

[9] V. S. Subrahmanian *et al.*, "The DARPA Twitter Bot Challenge," *Computer (Long. Beach. Calif).*, vol. 49, no. 6, pp. 38–46, 2016.

[10] N. Chavoshi, H. Hamooni, and A. Mueen, "DeBot: Twitter bot detection via warped correlation," *Proc. - IEEE Int. Conf. Data Mining, ICDM*, no. December, pp. 817–822, 2017.

[11] C. Cai, L. Li, and D. Zeng, "Behavior Enhanced Deep Bot Detection in Social Media," in *2017 IEEE International Conference Intelligence and Security Informatics (ISI)*, 2017, pp. 128–130.

[12] F. Morstatter, L. Wu, T. H. Nazer, K. M. Carley, and H. Liu, "A New Approach to Bot Detection : Striking the Balance Between Precision and Recall," in *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2016, pp. 533–540.

[13] M. H. Zweig and G. Campbell, "Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine," *Clin. Chem.*, vol. 39, no. 4, pp. 561–577, 1993.