YÖNETİM BİLİŞİM SİSTEMLERİ DERGİSİ
http://dergipark.gov.tr/ybs

Yayın Geliş Tarihi: 07.11.2019
Yayına Kabul Tarihi: 24.12.2019
Online Yayın Tarihi: 31.12.2019

Cilt:5, Sayı:2, Yıl:2019, Sayfa: 29-43
e-ISSN: 2630-550X

METİN MADENCİLİĞİ, MAKİNE VE DERİN ÖĞRENME ALGORİTMALARI İLE WEB SAYFALARININ SINIFLANDIRILMASI*

İlker ŞAHİN[1], Oumout CHOUSEINOGLOU[2]

[1] Comodo Inc., Ankara Türkiye

[2] Endüstri Mühendisliği Bölümü, Hacettepe Üniversitesi, Ankara Türkiye

ÖZET

Web sitelerine daha etkili ve hızlı ulaşmak ya da zararlı içeriği filtrelemek için Web siteleri içerikleri doğrultusunda farklı yöntemlerin kullanımı ile sınıflandırılabilmektedir. Fakat Web sitelerinin sayısının ve bu sitelerde bulunan içeriğin çok hızlı bir şekilde artması, bu sınıflandırma işleminin bilgisayarlar aracılığı ile otomatik yapılması ihtiyacını doğurmuştur. Bu çalışmada, literatürdeki makine öğrenmesi yöntemleri ve yapay sinir ağları kullanılarak Web sitesi sınıflandırma problemi İkili Sınıflandırma ve Çok Sınıflı Sınıflandırma olmak üzere iki farklı yaklaşım ile ele alınmış, performansları çalışma kapsamında toplanan Web siteleri üzerinde incelenmiştir. Tüm deneysel sonuçlar göz önüne alındığında İkili Sınıflandırma yaklaşımının, belirlenen bir Web site sınıfının filtrelenmesinde daha etkili olacağı tespit edilmiştir. Çoklu Sınıflandırma yaklaşımında, farklı kelime vektörleştirme yöntemleri uygulanıp performansları incelenmiştir. Bu çalışmada İkili ve Çoklu Sınıflandırma yaklaşımlarında uygulanan farklı algoritmaların farklı vektörleştirme yöntemleri ile beraber kullanılması, Web sayfalarının sınıflandırılması ve içerik filtrelenmesi problemlerinin ortak çözümlenmesi için bir yöntem olarak sunulmuştur.

Anahtar Kelimeler: Web Sayfa Sınıflandırması, Metin Madenciliği, Doğal Dil İşleme, Makine Öğrenmesi

WEB PAGE CATEGORIZATION WITH TEXT MINING, MACHINE AND DEEP LEARNING ALGORITHMS†

ABSTRACT

Web sites can be classified according to their content with the use of different methods in order to enable accessing them in a more efficient and fast way, or to filter malicious contents in them. However, the number and the content size of the Web sites increases rapidly creating the need to realize this classification automatically with the use of computers. In this study, the Web site classification problem is examined by using different machine learning methods and artificial neural networks with two different approaches, namely Binary Classification and Multiple Classification. Both approaches are tested and their performances are compared by using a number of Web sites collected for this study.

* Bu çalışma, birinci yazarın ikinci yazar danışmanlığında hazırladığı aynı başlıklı yüksek lisans tezinden üretilmiş ve 6. Uluslararası Yönetim Bilişim Sistemleri Konferansında (IMISC2019, 9-12 Ekim 2019, Kadir Has Üniversitesi, İstanbul) sunulmuştur.

† This manuscript is produced from the Master's Dissertation with the same title prepared by the first author under the supervision of the second author and was presented at the 6th International Management Information Systems Conference (IMISC2019, 9-12 October 2019, Kadir Has University, İstanbul).

Considering all experimental results, it is found that the Binary Classification approach is more effective when used to filter a specified Web site class. Different word vectorization methods have been employed and their performances have been compared within the Multiple Classification approach. In this study, the use of different algorithms together with different vectorization methods within the context of the Binary and Multi-class Classification approaches has been proposed as a method for the mutual solution of classification and content filtering problems on Web pages.

Keywords: Web Page Classification, Text Mining, Natural Language Processing, Machine Learning

GİRİŞ

Bireylerin ve kurumların İnternet erişim imkânları arttıkça, Web üstünden paylaşılan bilgiler ve Web sitelerinin sayısı orantılı olarak hızlı bir şekilde artmaktadır. 2019 yılı itibarı ile dünya çapında 1,7 milyar civarında Web sitesi bulunmakta, fakat bunlardan sadece 200 milyona yakını aktiftir (Internet Live Stats, 2019; Netcraft, 2019). Web sitelerinin ve bu sitelerde barındırılan sayfaların sayısının artması ile, kişilerin istedikleri bilgilere hızlı ve kolay erişmeleri ya da zararlı olabilecek içerikleri onlardan engellemek için, Web sayfalarını sınıflandırma ihtiyacı doğmuştur (Rekik vd., 2018). Qi ve Davison (2009) çalışmalarında, Web sayfalarına benzer şekilde Web sitelerinin de sınıflandırılabileceğini, bu sınıflandırma çalışmalarında da sadece Web sitesi içeriğine odaklanan, sadece Web sitesi yapısal özelliklerini kullanan, veya Web sitesi içeriği ve yapısal özellik bilgilerini beraber değerlendiren çalışmalar olduğunu belirtmektedirler. Web sitesi sınıflandırması ile, o sitenin başlıca konusunun üst başlığı öngörülerek ilgili ve yararlı olduğu düşünülen sayfalara daha etkin ve hızlı bir şekilde ulaşılabilirken, ilgisiz Web sitelerine ve zararlı içeriklere kişilerin erişimi engellenebilmektedir. Metin yoğun bir site sınıflandırma amacı ile incelendiğinde, site içeriğinin insan gözüyle işlenmesi ve sınıflandırılması için çok fazla efor ve zaman harcanabilir, sınıflandırma sürecinde hatalar yapılabilir. Dolayısıyla, Web sınıflandırma işleminin daha hızlı ve daha etkin olması için farklı yaklaşımlara ihtiyaç duyulmakta, bu kapsamda başta araştırmacılar, Web şirketleri, arama motorları ve siber güvenlik sağlayan özel şirketler olmak üzere farklı alanlardan kişi ve kurumlar, sınıflandırma algoritmaları ve bu algoritmaları kullanan sınıflandırma yöntemleri geliştirmektedir. Bu sayede içeriğe göre sınıflanmış olan Web sitelerini ziyaret eden İnternet kullanıcısının, gerekli görüldüğü durumlarda önceden belirlenen sınıflara sahip veya olası sahtekârlık unsurları içeren Web sitelerine erişimin engellenmesi hedeflenmektedir.

Qi ve Davison (2009) ve Ester, Kriegel ve Schubert'in de (2002) belirttikleri üzere, Web sitesi sınıflandırması Web sayfa içeriğinin değerlendirilmesi ve sınıflandırılması ile gerçekleştirilebilir ve Web sayfasının temel içerik bileşeninin metin olmasından dolayı, metin sınıflandırma yaklaşımları bu doğrultuda kullanılabilmektedir. Bu çalışmada Web sitesi sınıflandırma problemi bir metin sınıflandırma problemi olarak kabul edilmiş olup metin sınıflandırmasında kullanılan yöntemler ile dört alt başlık altında incelenmiştir; bunlar sırası ile Web sitesi sınıflandırması için gerekli verinin tanımı ve toplanması, toplanmış verinin ön işleme sürecinden geçirilmesi, kelime vektörleştirme ve sınıflandırmada kullanılan algoritmaların eğitim, test ve analiz sürecidir. Metin sınıflandırması; önceden oluşturulmuş ve belirlenmiş olan sınıflardan faydalanarak, incelenmekte olan bir metnin, belgenin ya da bir cümlenin bu belirlenmiş olan sınıflardan hangisine dâhil olacağının otomatik olarak hesaplanması işlemidir (Hartmann vd., 2019; Li ve Jain, 1998). Diğer bir deyişle metin sınıflandırması, ele alınmakta olan bir metne, bu metnin içeriğine göre ve önceden belirlenmiş yöntemler ile etiket veya kategori atama işlemi olarak tanımlanabilir.

Metin sınıflandırma işlemi, Denklem 1 (Sebastiani, 2002; Kowsari vd., 2019) şeklinde ifade edilebilir.

$$Y = f(X, \theta) + \varepsilon \quad (1)$$

Denklem 1'de,

***f*:** eğitim verilerini kullanarak tahmin eden sınıflandırıcı veya sınıflandırma modeli

***Y*:** incelenmekte olan metnin önceden tanımlanmış olan bir metin sınıfına üyeliğini belirten değer

***X*:** kelime veya kelime gruplarından oluşan bir metin vektörü

***θ*:** *f* fonksiyonuyla ilişkilendirilen bilinmeyen parametrelerin kümesi

***ε*:** sınıflandırma hatasıdır.

Mevcut çalışmada, Web sitelerinin sınıflandırılması sürecinde uygulanan metin sınıflandırma modelinin oluşturulmasında İkili Sınıflandırma ve Çok Sınıflı Sınıflandırma (Sebastiani, 2002) yaklaşımları kullanılmıştır. Hem Çok Sınıflı Sınıflandırma hem de İkili Sınıflandırma yaklaşımlarında model (*f*) olarak literatürde yaygın olarak kullanılmış ve başarıları çeşitli çalışmalarda test edilmiş olan başlıca makine öğrenmesi ve derin öğrenme yöntemleri uygulanmış, devamında da mevcut çalışma kapsamındaki başarım ve etkinlikleri uygun değerlendirme yöntemleri ile ölçülmüştür. Uygulama içeriği İngilizce tabanlı Web siteleri üzerinde gerçekleştirilmiş, örnekler ve bulgular bu şekilde sunulmuştur. Bu çalışmanın devamında; Kısım 2'de kısaca benzer çalışmalara değinilmekte ve Kısım 3'te kullanılmış olan metodoloji verilmektedir. Kısım 4'te yapılan örnek uygulama sonucunda elde edilen bulgular, bu bulgular ışığında da Kısım 5'te tartışma ve öneriler sunulmaktadır.

BENZER ÇALIŞMALAR

Gali, Mariescu-Istodor ve Fränti (2017), Web sayfalarını sınıflandırma problemini ele alırken şimdiye kadar yapılan çalışmaların büyük kısmının HTML yapılardan elde edilen yazı ve metin bilgilere dayalı yapıldığını ve bu yaklaşımın Web sayfasında karşılaşılan reklamlardan dolayı yanıltıcı olabileceğini ileri sürmüştür. Problemin çözümü için istatistiksel teknikler, dil bilgisi ve metin bölümleme metotlarını kullanmışlardır. Naive Bayes, k-En Yakın Komşu (k-nearest neighbors, KNN) ve Destek Vektör Makineleri (Support Vector Machine, SVM) sınıflandırıcılarla değerlendirdikleri bu yaklaşımlarında, metin analizi ve gruplandırılması bakımından önemi vurgulamışlardır. Diğer yandan Shen, Yang ve Chen (2007), Web sayfalarına gömülmüş çeşitli gürültü bilgiler nedeniyle, Web sayfası sınıflandırması probleminin saf metin sınıflandırmasından daha zor bir probleme dönüştüğünü iddia etmektedirler. Yazarlar, Web sayfası sınıflandırma performansını iyileştirmek için özetleme tekniklerinin kullanımı ile gürültünün yok edilmesini önermiş, bu doğrultuda Web sayfası düzenine dayanan özgün bir sayfa içeriği özetleme algoritması ortaya koyup, LookSmart Web dizinindeki diğer birçok güncel metin özetleme yöntemi ile birlikte değerlendirmişlerdir. Elde ettikleri deneysel bulgular, herhangi bir özetleme yaklaşımı uygulanmış olan Naive Bayes ve SVM sınıflandırma algoritmalarının, saf metin tabanlı sınıflandırma algoritmalarına kıyasla %5'den daha yüksek bir iyileştirme göstermiştir. Ayrıca bu çalışmada, farklı özetleme algoritmalarını birlikte kullanmak için bir topluluk özetleme yöntemi sunulmakta olup, saf metin tabanlı yöntemlere göre önerilmekte olan bu yenilikçi yaklaşımın %12'den daha fazla gelişme sağladığı belirtilmektedir.

Diğer yandan, Web sitesinde bulunan sayfaların içeriğini kullanmadan ve sadece Web sitesi URL yapısını inceleyerek siteleri sınıflandıran yaklaşımlar da önerilmiştir. Rajalakshmi ve Aravindan (2018) esnek hesaplamalı bir yaklaşım ve Naive Bayes sınıflandırma ile, sadece URL yapısından çıkarılmış olan kelime ve özellikleri kullanarak ve bu özellikleri ağırlıklandırarak, Web sitelerini 13 ön tanımlı sınıf üstünden sınıflandırmaktadır. Web sitesi metin içeriğine bakmayan bu tarz yaklaşımların, Web sitelerini sınıflandırma probleminde bir ön filtreleme yöntemi olarak kullanılabileceğini, fakat daha

etkin ve kapsamlı sonuçlar almak için Web sitesi içeriklerinin de devamında sınıflandırmada kullanılması gerektiğini belirtmektedirler. Whittaker, Ryner ve Nazif (2010) oltalama (phishing) saldırısı tehdidi bulunan ve bulunmayan Web sitelerini, site URL'i, sitenin barındırılma bilgisi ve Web sitesinde bulunan sayfaların HTML içeriğini kullanarak sınıflandırmaya çalışmaktadır. Geliştirmiş oldukları sınıflandırma yaklaşımı makine öğrenmesi yöntemlerini kullanarak %0,1'den daha az yanlış pozitif başarım oranı yakalamaktadır. Yazarların da belirttiği üzere, bu şekilde otomatik yöntemler kullanarak olatalama gibi kötü niyetli ve saldırı içerikleri barındıran Web sitelerini sınıflandırarak tanımlamaya çalışmak, saldırı girişiminde bulunan tarafın akıllı olmasından dolayı hep bir adım geriden gelecek, fakat buna rağmen bu sınıflandırma girişiminin de birçok kullanıcıyı kötü niyetli kişi ve Web sitelerinden koruyabilecektir.

Zhang vd. (2010), Web sitesi sınıflandırma probleminin sadece metin içeriğin incelenmesi ile yapılamayacağını, Web sitesinin aslında birbirleri ile HTML köprüleri vasıtasıyla bağlanmış sayfalar bütünü olduğunu belirterek sınıflandırma problemini Web sitesini oluşturan sayfaların topolojisini üstünden yapmaktadırlar. Önerdikleri yöntemde ilk önce Web sitesinin topolojisi bir yönlendirilmiş çizge olarak oluşturulmakta, devamında alt çizgeler elde edilerek bu alt çizgelerde PageRank algoritmaları ile konu başlığı vektörleri elde edilmektedir. Bu konu başlığı vektörleri bir SVM sınıflandırma algoritması vasıtası ile işleme alınarak Web sitesinin sınıflandırılması gerçekleştirilmektedir. Ren vd. (2019) yaptıkları çalışmada, SVM sınıflandırma algoritmasının özellik seti çok boyutlu olan veri setindeki uyumluluğunu incelemişler, yaptıkları hesaplamalar sonucunda iki farklı çekirdek (kernel) fonksiyonunu hedefleyen ve genetik algoritmalar ile optimize edilmiş bir SVM önerilmiştir. Bu sayede, metin sınıflandırma problemi için oluşturulan SVM modelinin genelleştirme ve öğrenme yeteneklerinin arttığı, simülasyon verileri kullanılarak metin sınıflandırma üzerinde gösterilmiştir. Chen, Cheng ve Cheng (2016) tarafından yapılan çalışmada, gıda ürünleri ile ilgili bir Web sitesinde farklı başlık ve konulara sahip kapsamlı gıda haber raporlarının bulunmasının sık karşılaşılan bir durum olduğu, bu raporlardan otomatik şekilde analizler alınması ihtiyacının bir Web sayfası metin sınıflandırma problemi olarak ele alınabileceği belirtilmiştir. Çalışmalarında, Uzun-ve-Kısa-Süreli-Hafıza (Long Short Term Memory) temelli toplu öğrenme kullanarak bir gıda güvenliği belge sınıflandırma yöntemi önerilmiştir. Ayrıca, gıda güvenliği ile ilgili olmayan belgelerden oluşan ve negatif örnek olarak tanımlanan çok sayıda etiketsiz haber metin genişletme yaklaşımında kullanılmıştır.

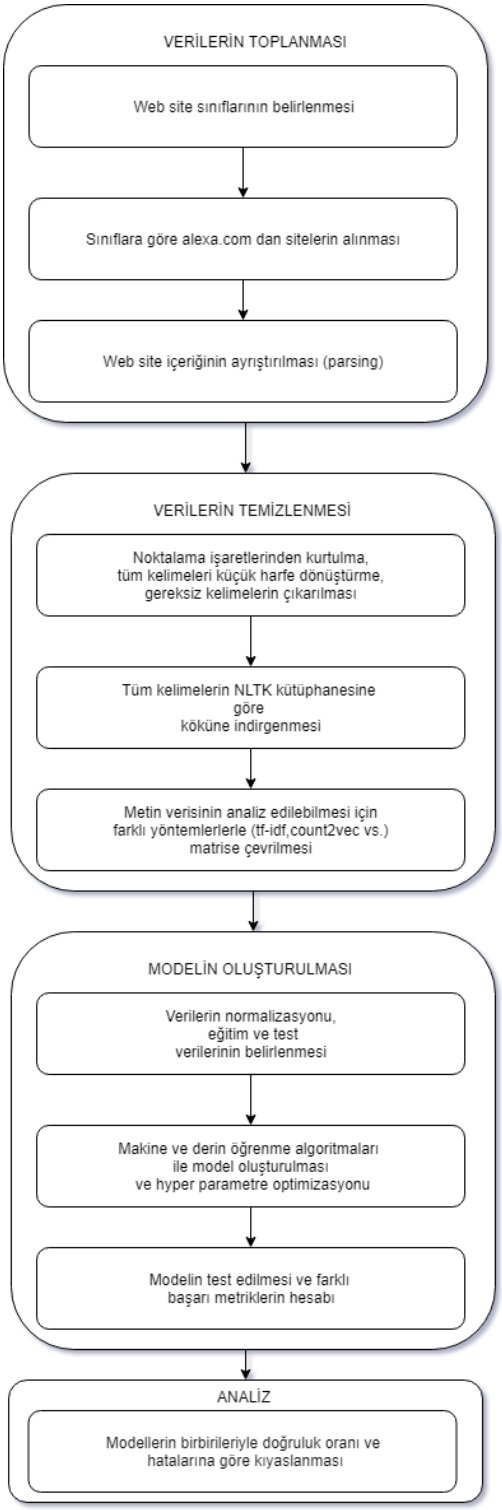
METODOLOJİ

Bu çalışmada, Web sitelerinin sınıflandırılması için gözetimli sınıflandırıcılar kullanılmıştır (Panigrahi ve Borah, 2019). Geliştirilmiş olan gözetimli metin sınıflandırma modeli, daha önce Web sayfa sınıfı ile etiketlenmiş Web sayfalarının içerikleri ile eğitilip (gözetimli öğrenme), yeni bir Web sayfası içerik metni girdi olarak verildiğinde eğitim kümesinde verilen sınıflardan biri ile etiketlenecek şekilde tasarlanmıştır. Bu kapsamda, gözetimli sınıflandırmada ilk olarak daha önce sınıflandırılmış Web sayfaları kullanılarak bir metin sınıflandırma modeli eğitilmiş, ardından test kümesindeki her bir Web sayfası için sınıf tahmini yapılmıştır.

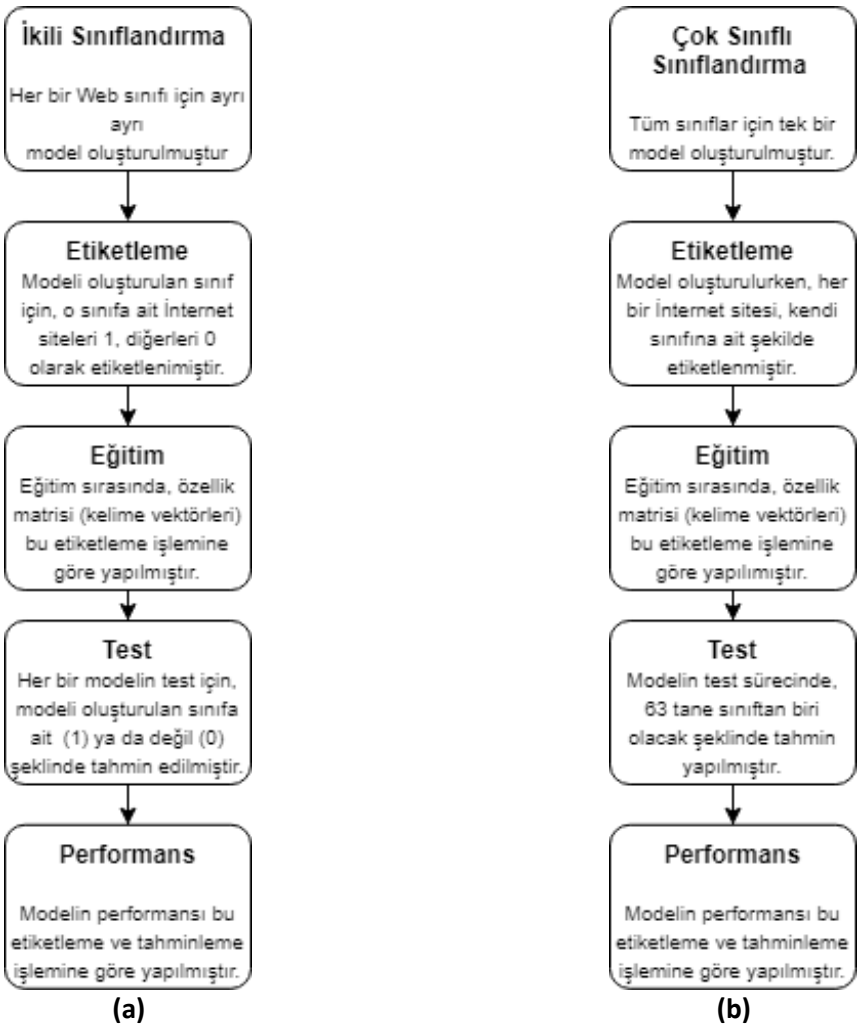
Çalışmada, Web sitesi sınıflandırması için İkili (Binom) Sınıflandırma ve Çok Sınıflı Sınıflandırma olmak üzere iki alternatif gözetimli öğrenme yaklaşımı denenmiş ve karşılaştırılmıştır (Qi ve Davison, 2009). İkili Sınıflandırıcı yönteminde, Şekil 2a'da gösterildiği üzere, her bir sınıf için ayrı ayrı ikili sınıflandırıcı modeller oluşturulmuş, devamında eğitim sırasında özellik seti ile birlikte etiketlenen değişkene, modeli oluşturulan sınıf için 1, geri kalan örnekler için 0 değeri verilmiştir (Takenouchi ve Ishii, 2018). Çok Sınıflı Sınıflandırma yaklaşımında ise tüm sınıfları içerecek şekilde çok sınıflı sınıflandırma modeli oluşturulmuştur. Bu yaklaşımda, Şekil 2b'de gösterildiği şekilde, model bir kez oluşturulmuş, tüm sınıflar, bu model içerisinde etiketlenen değişken olarak, özellik seti ile verilmiştir. Her iki yaklaşım farklı algoritmalar ile modellenip, bir örnek veri seti üstünde başarım

oranları incelenmiştir. Şekil 1'de bu çalışma kapsamında takip edilmiş olan adımlar gösterilmekte ve makalenin devamında her bir adımda yapılmış olan işlemler detaylı bir şekilde verilmektedir.

Şekil 1. Çalışmada Uygulanan Sınıflandırma Süreci



Şekil 2. Çalışmada Takip Edilen İkili Sınıflandırma Süreci (a) ve Çok Sınıflı Sınıflandırma Süreci (b)



Veri Tanımı ve Toplanması

Sınıf belirleme sürecinde, Web sitelerine bir sınıf atanabilmesi için olabilecek en geniş sınıf sayısı belirlenmiştir. Veri toplama işlemi kapsamında Amazon'un bir hizmeti olan Alexa Top Sites'tan[3] daha önceden yazarlar tarafından belirlenen 63 farklı sınıftan 45.543 adet Web sayfası alınmış, bu sayfalardaki kelime ve kelime grupları söz konusu bu 63 tane sınıfta kullanılacak olan veri olarak kabul edilmiştir. Çalışmaya özel olarak geliştirilmiş olan Python 3 betikleri ile Web sayfalarından veriler otomatik olarak toplanmış ve istenilen biçimde bir sonraki aşama için yapılandırılmıştır. Yazarlar tarafından belirlenmiş olan 63 Web sitesi sınıfı arasından toplamda en fazla kelime sayısına sahip olan ilk 10 sınıf, o sınıf içinde kullanılmış olan Web site sayıları ve bu Web sitelerine ait Web sayfalarından elde edilmiş olan toplam kelime sayıları Tablo 1'de örnek olarak verilmektedir.

[3]https://www.alexa.com/topsites

Tablo 1. Çalışmada Kullanılan ve En Fazla Toplam Kelime Sayısına Sahip İlk 10 Web Sayfa Sınıfı

Sınıf Adı	Bu Sınıfa Ait Web Site Sayısı	Web Sitelerinde Bulunan Toplam Kelime Sayısı
Political Issues	1.000	534.330
Community and Society	1.000	420.280
Education-Reference	1.000	411.976
Food-Drink	1.000	393.388
News	1.000	384.178
Religion-Spirituality	1.000	384.135
Paranormal	1.000	377.745
Adult Content	1.000	372.913
Travel	1.000	352.077
Software-Hardware	1.000	350.071

Çalışma kapsamında uygulanmış olan her iki sınıflandırma yaklaşımında da (İkili Sınıflandırma ve Çok Sınıflı Sınırlandırma) Alexa Top Sites'tan elde edilmiş olan toplam 45.543 Web site verisinden (içerik metni) 37.814 tanesi (%83,03) eğitim seti olarak kullanılmışken, test sırasında 7.729 tane Web site verisi (%16,97) kullanılmıştır. İkili Sınıflandırma sürecinde, bütünleşik veri seti karıştırılarak 10 defa farklı eğitim ve test veri setleri elde edilmiş, devamında da model tekrar oluşturulmuştur.

Veri Temizleme

Çalışmada kullanılmış olan 45.543 Web sitesinden elde edilen veri üstünde temizleme ve ön işleme süreçleri, seçilmiş makine ve derin öğrenme teknikleri uygulanmadan önce gerçekleştirilmiştir. Bunlar sırası ile; metindeki her türlü noktalama işaretlerinden kurtulma, önceden tanımlanmış olan etkisiz kelimelerin çıkarılması, metin içinde bulunan büyük/küçük harf uyumsuzluğundan kurtulmak için tüm kelimeleri küçük harfe dönüştürme ve her bir kelimenin kök öbeğine ulaşmadır. Bu çalışma kapsamında "etkisiz kelimeler" olarak bağlaç, imleç, sayı ve kalıplaşmış kısaltma gibi içerikten bağımsız kelimelere ve kelime yapıları kabul edilmiştir. Uygulamada kullanılmış olan Web sitelerine ait Web sayfaları sadece İngilizce içeriğe sahip şekilde seçilmiş olduğu için etkisiz (stop word) kabul edilen "the", "a", "I", "of" ve benzeri kelimeler metinden çıkartılmıştır. Web sitelerinden toplanmış olan metin verisinin temizlenmesi ve önişlenmesi sürecinde literatürdeki benzer çalışmalarda yaygın olarak atıfta bulunulan Python tabanlı bir doğal dil işleme kütüphanesi olan NLTK[4] ve NLTK'da tanımlanmış stop word listesi[5] kullanılmıştır (Loper ve Bird, 2002).

Kelime Vektörleştirme

Seçilmiş olan makine ve derin öğrenme yöntemlerinin uygulanmasından önce, Web sitelerinden toplanmış olan metinlerin kelime vektörleştirme işlemine alınması gerekmektedir. Kelime vektörleştirme bir dönüşüm işlemi olup, bu işlem ile birlikte her bir metin, kendi sınıfı içerisinde sayısal

[4] https://www.nltk.org/
[5] https://gist.github.com/sebleier/554280

veriler içeren vektörlere dönüştürülmektedir (Stein vd., 2019). Kelimelerin birbirine yakınlığı ya da kelimelerin sınıf içerisindeki sıklığı gibi farklı yöntemler kelime vektörleştirmede yaygın olarak kullanılmaktadır. Web sitelerinden elde edilmiş olan veriler metin tabanlı olduğundan, kelime vektörleştirme ile makine ve derin öğrenme algoritmaları tarafından eğitilebilecek ve analiz edilebilecek sayısal değerlere dönüştürülmüştür. Bu çalışmada kelime vektörleştirme işlemi için literatürde önerilen ve yaygın olarak kullanılan Kelime Torbası (Bag of Words, BOW) (Sinoara vd., 2019), Terim Sıklığı - Ters Metin Sıklığı (Term Frequency – Inverse Document Frequency, TF-IDF) ve Word2Vec (Stein vd., 2019) yöntemleri her bir algoritmada ayrı ayrı kullanılmış, devamında da kendi aralarındaki başarım performansları kıyaslanmıştır.

Sınıflandırma Sürecinde Kullanılan Algoritmalar

Ham verinin ön işleme adımlarından geçirilmesi ve analiz edilebilir hale getirilmesi sürecinden sonra; çalışmada kullanılmakta olan veri yapısına ve sınıflandırma problemine uygunluğu farklı alanlardaki birçok çalışma ile doğrulanmış makine ve derin öğrenme algoritmaları uygulanmak üzere seçilmiştir. Devamında bu algoritmalar, çalışmanın temel yaklaşımını teşkil eden İkili Sınıflandırma modellerini ve Çoklu Sınıflandırma modellerini (Qi ve Davison, 2009) oluşturmak için kullanılmıştır.

İkili (Binom) Sınıflandırma'da, tanımlanmış sınıf sayısı kadar (63 tane) sınıflandırıcı oluşturulmuştur. Devamında, İkili Sınıflandırma için ayrı ayrı Bernoulli Naive Bayes, Lojistik Regresyon (Logistic Regression) ve Tam Bağlantılı Yapay Sinir Ağları (Fully Connected Neural Network, FCNN) algoritmaları sırası ile kullanılmıştır. Her üç algoritmanın başarım oranları, hata matrisleri ve işlem süreleri incelenmiş ve birbirleri ile karşılaştırılmış (Manning vd., 2010; Hilbe, 2011), bulgular Tablo 3 ve Şekile 3'te verilmektedir.

Diğer yandan **Çok Sınıflı Sınıflandırma**'da 63 Web sitesi sınıfından her biri, kendi özellik kümesinde (kelime vektörü) temsil edilecek şekilde eğitilmiş ve bu şekilde Çok Sınıflı Sınıflandırma modeli oluşturulmuştur. Eğitim sürecinden sonra, algoritmaların testinde tahmin edilmesi beklenen çıktı yine her bir 63 sınıftan biridir. Bu sınıflandırma yaklaşımında İkili Sınıflandırmadan farklı olarak; 63 sınıftan her biri için ayrı ayrı sınıflandırıcılar oluşturmak yerine tek bir sınıflandırıcı oluşturulmuş ve tüm testler bu tek model üzerinden yapılmıştır. Multinominal Naive Bayes, Rastgele Orman (Random Forest) ve SVM sınıflandırıcılar ayrı ayrı kullanılarak Çok Sınıflı Sınıflandırma modelini oluşturulmuştur. Modelin oluşturulmasının devamında her bir sınıflandırıcının başarım oranları incelenmiş (Onan vd., 2016; Chen ve Hsieh, 2006; Xu vd., 2017), elde edilen sonuçlar Tablo 4 ve Şekil 4'te verilmektedir.

Sınıflandırıcı Performans Ölçme Yöntemleri

Mevcut çalışmada, İkili ve Çok Sınıflı Sınıflandırma için kullanılmış olan makine ve derin öğrenme yöntemlerinin performanslarının değerlendirilmesi için, yaygın olarak kullanılan bir ölçme yöntemi olan hata matrisinin bileşenleri ve bu bileşenlerden üretilmiş ve Denklem 2'de verilen Başarım Oranı (Accuracy), Denklem 3'te verilen F1 Skoru (F1 Score), Denklem 4'te verilen Kesinlik (Precision) ve Denklem 5'te verilen Duyarlılık (Recall) ölçüm değerleri kullanılmıştır. Tablo 2'de çalışmada kullanılmış olan hata matrisi gösterimi verilmekte olup bu gösterimde, negatif (0) Web sitesinin ilgili (sınıflandırıcısı oluşturulan) sınıfa ait olmaması durumunu; pozitif (1) gösterimi ise Web sitesinin ilgili sınıfa ait olması durumunu belirtmektedir.

Tablo 2. İkili Sınıflandırmada Kullanılan Hata Matrisi Gösterimi

		Gerçekte Olan	
		Negatif (0)	**Pozitif (1)**
Tahmin Edilen	**Negatif (0)**	Doğru Negatif (True negative, TN) 0 olarak tahmin edilen sınıfın gerçekte de 0 olması durumudur.	Yanlış Negatif (False negative, FN) 0 olarak tahmin edilen sınıfın gerçekte 1 olması durumudur.
	Pozitif (1)	Yanlış Pozitif (False positive, FP) 1 olarak tahmin edilen sınıfın gerçekte 0 olması durumudur.	Doğru Pozitif (True positive, TP) 1 olarak tahmin edilen sınıfın gerçekte de 1 olması durumudur.

- Başarım Oranı (Denklem 2); tüm test seti içerisinde doğru sınıflandırılmış Web sitesi oranıdır.

$$Başarım\ Oranı = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

- Kesinlik (Denklem 3); pozitif tahmin edilen Web sitelerinin kaçının doğru tahmin edildiğinin oranı olup, FP tahmin maliyeti yüksek olduğu durumlarda kullanılması önerilmektedir. Bu çalışma kapsamında filtrelenmek istenen bir sınıfa o sınıfa gerçekte ait olmayan bir Web sitesinin dahil edilerek filtrelenmesi maliyeti olarak düşünülebilir.

$$Kesinlik = \frac{TP}{TP + FP} \quad (3)$$

- Duyarlılık (Denklem 4); gerçekte pozitif sınıfa ait olan Web sitelerinin kaçının pozitif tahmin edildiği oranı olup, FN tahmin maliyeti yüksek olduğu durumlarda kullanılan bir ölçümdür. Bu çalışma kapsamında filtrelenmek istenen bir Web site sınıfına ait olan bir sitenin, o sınıfa dahil edilmeyerek filtrelenmemesinin maliyeti olarak düşünülebilir.

$$Duyarlılık = \frac{TP}{TP + FN} \quad (4)$$

- F1 Skoru (Denklem 5); kesinlik ve duyarlılık ölçümlerinin harmonik ortalaması olup, her ikisini eşzamanlı göz önünde bulunduran ve yaygın olarak benzer çalışmalarda kullanılan önemli bir göstergedir.

$$F1\ Skoru = \frac{Duyarlılık * Kesinlik}{Duyarlılık + Kesinlik} \quad (5)$$

İkili Sınıflandırma yöntemi kapsamında uygulanan algoritmaların performanslarının ölçümünde yukarıda detayları verilmiş olan dört ölçüm yöntemi de kullanılırken, diğer yandan Çoklu Sınıflandırmada kullanılmış algoritmaların performansını ölçmek için sadece Başarım değeri kullanılmıştır. Çoklu Sınıflandırma yaklaşımında Başarım ölçümü, doğru sınıflandırılmış (TP ve TN) toplam Web sitesi sayısının, test setinde kullanılan toplam Web sitesine oranı ile elde edilmiştir (Denklem 2).

BULGULAR

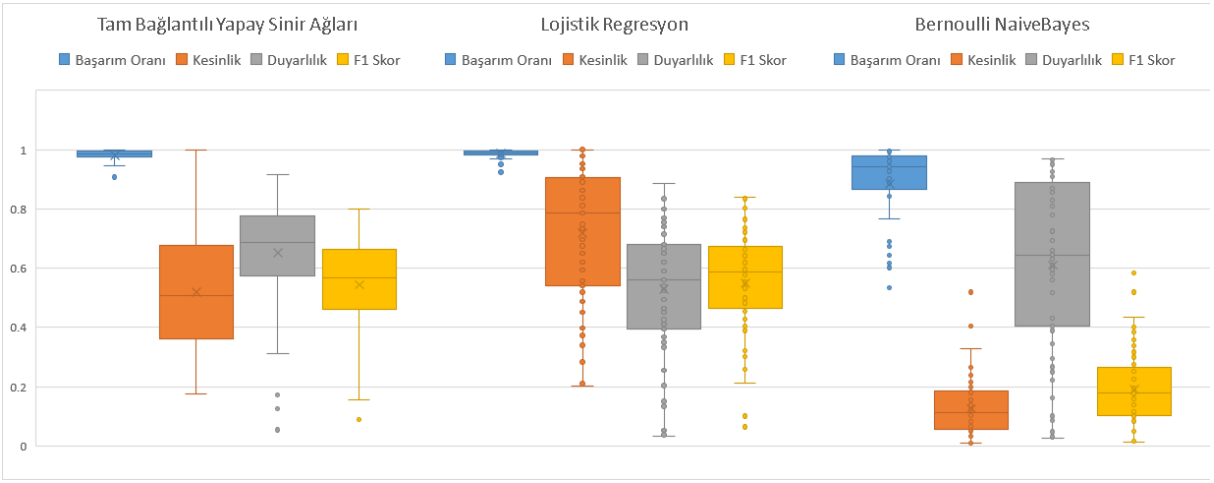
Bu çalışmada, ele alınmakta olan Web sitelerinin sınıflandırılması problemi kapsamında her bir sınıf için oluşturulan Çoklu Sınıflandırıcı modelleri ile TF-IDF, BOW ve Word2Vec kelime vektörleştirme yöntemleri ayrı ayrı kullanılmıştır. Diğer yandan, İkili Sınıflandırma yaklaşımı için ise sadece TF-IDF kullanılmıştır. Bu modeller belirlenmiş olan eğitim seti kullanılarak oluşturulduktan sonra, test setinde bulunan Web sitesi verileri kullanılarak sonuçlar elde edilmiş ve İkili Sınıflandırma ve Çoklu Sınıflandırma yaklaşımı olmak üzere iki farklı grupta incelenmektedir. Modeller Python 3 betikleri ile kodlanmış olup kodlama sürecinde Python'un scikit-learn[6], SciPy[7], NLTK, Keras[8], PyPI regex[9], Pandas[10] ve NumPy[11] kütüphanelerinde bulunan fonksiyonlar kullanılmıştır. Python betikleri Intel Core i7-7700HQ 2.80Ghz işlemcili 16 GB DDR4 RAM donanımına sahip ve Windows 10 İşletim Sistemli bir bilgisayarda çalıştırılmıştır.

İkili Sınıflandırma yaklaşımında kullanılan yöntemler Tam Bağlantılı Yapay Sinir Ağları, Lojistik Regresyon ve Bernoulli Naive Bayes olup her biri için kendi içinde Başarım, Kesinlik, Duyarlılık ve F1 Skor performans ölçüm değerlerinin temel istatistiki değerleri (ortalama, minimum, maksimum, mod ve medyan) ve ayrıca ortalama işlem süreleri hesaplanmıştır. Tablo 3 ve Şekil 3'te verilen bu değerlere bakıldığında, Kesinlik ve Başarım ölçümlerine göre en başarılı sınıflandırıcı Lojistik Regresyon olurken, Tam Bağlantılı Yapay Sinir Ağları'nın en yüksek duyarlılığa sahip olduğu görülmektedir. Diğer yandan F1 skorları kapsamında Bernoulli Naive Bayes sınıflandırıcısının düşük performans göstermektedir. Kullanılan yöntemlerin işlem süreleri ele alındığında Bernoulli Naive Bayes ve Lojistik Regresyon sınıflandırıcılarının, Tam Bağlantılı Yapay Sinir Ağlarına göre daha az sürede çalışmayı tamamladığı ve sonuç verdiği görülmüştür.

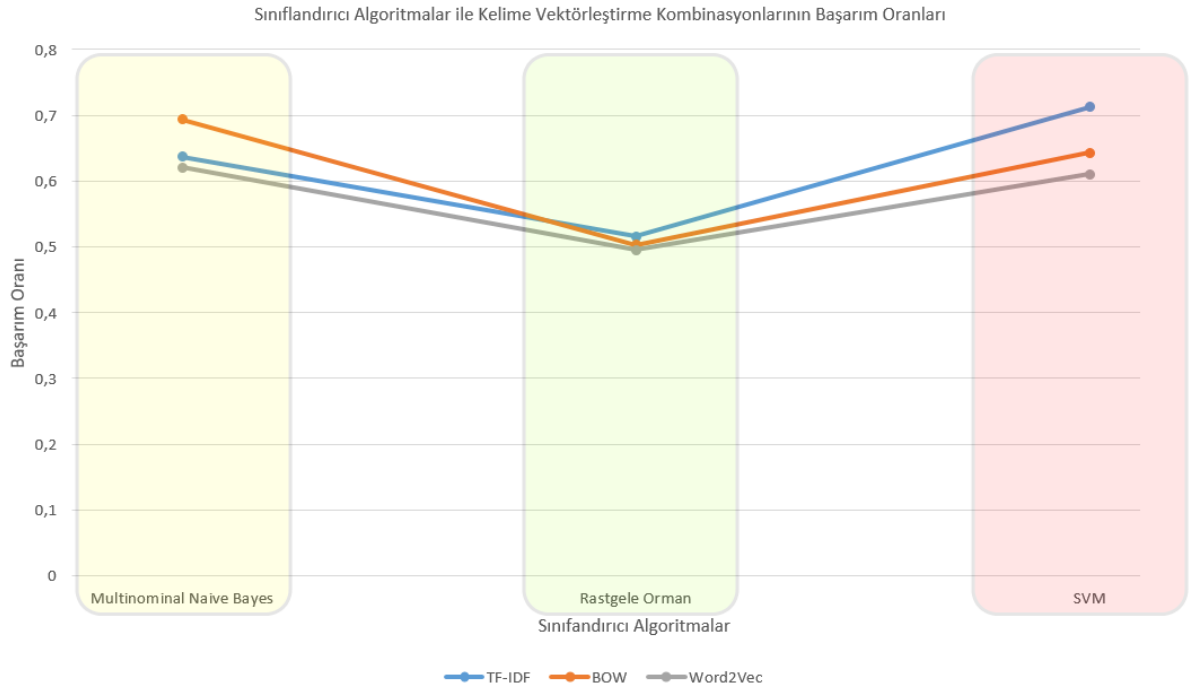
Çok Sınıflı Sınıflandırma yaklaşımında üç kelime vektörleştirme yöntemi ile üç sınıflandırma algoritması beraber kullanılarak dokuz farklı kombinasyonun başarım oranları 63 sınıfın tamamı için elde edilmiştir. Çok Sınıflı Sınıflandırma 10 defa rastsal şekilde farklı eğitim ve test kümeleri oluşturularak denenmiştir. Deney sonucunda elde edilen başarım oranlarını içeren Tablo 4 ve Şekil 4 incelendiğinde, farklı kelime vektörleştirme yöntemlerinin farklı sınıflandırıcılarda kullanımının sınıflandırmanın başarılı olmasında bir etki oluşturduğu düşünülmektedir. Sonuçlar SVM'nin çok sınıflı sınıflandırıcı olarak ve TF-IDF'in de kelime vektörleştirme yöntemi olarak beraber kullanılmasının, diğer alternatif kombinasyonlara göre en başarılı sonucu verdiği görülmüştür. Bu çalışma kapsamında yapılan tüm deneysel sonuçlar göz önüne alındığında TF-IDF yönteminin ortalamada en iyi sonucu verdiği, devamında da sırasıyla BOW ve Word2Vec'in bu kelime vektörleştirme yöntemini izlediği söylenebilir. İşlem süreleri incelendiğinde, Çok Sınıflı Sınıflandırma yaklaşımında denenmiş olan dokuz alternatif arasında önemli bir süre farkı gözlemlenmemiş, dolayısı ile diğer alternatiflere göre bir veya birkaçının bu kapsamda öne çıktığı söylenememektedir.

[6] https://scikit-learn.org
[7] https://www.scipy.org
[8] http://keras.io
[9] https://pypi.org /project /regex
[10] https://pandas.pydata.org
[11] https://numpy.org

Şekil 3. İkili Sınıflandırıcıların Ölçüm Değerleri



Şekil 4. Sınıflandırıcı Algoritmalar ile Kelime Vektörleştirme Kombinasyonlarının Başarım Oranları (%)



Tablo 3. İkili Sınıflandırıcı Algoritmaların Performansları

	İkili Sınıflandırma Algoritmaları														
	Tam Bağlantılı Yapay Sinir Ağları					**Lojistik Regresyon**					**Bernoulli Naive Bayes**				
	Ort. ($\bar{x}$)	**Min**	**Max**	**Mod**	**Medyan**	**Ort. ($\bar{x}$)**	**Min**	**Max**	**Mod**	**Medyan**	**Ort. ($\bar{x}$)**	**Min**	**Max**	**Mod**	**Medyan**
Başarım (%)	0,980	0,907	1	0,998	0,9865	0,987	0,923	1	0,986	0,9890	0,884	0,534	0,998	0,945	0,9430
Kesinlik (%)	0,519	0,175	1	0,175	0,5065	0,717	0,201	1	0,628	0,7680	0,125	0,008	0,520	0,065	0,1070
Duyarlılık (%)	0,651	0,054	0,915	0,835	0,6975	0,527	0,033	0,885	0,620	0,5650	0,610	0,025	0,970	0,950	0,6625
F1 Skor (%)	0,543	0,089	0,799	0,531	0,5635	0,548	0,065	0,840	0,626	0,5910	0,190	0,012	0,583	0,103	0,1680
İşlem Süresi (Sn.)	64,700	53,064	97,958	58,255	59,517	0,550	0,298	0,954	0,666	0,5500	0,045	0,039	0,078	0,049	0,0440

Tablo 4. Çok Sınıflı Sınıflandırıcı Algoritmaların Başarım Oranları

		Çok Sınıflı Sınıflandırma Algoritmaları *(Başarım Oranları)*		
		Multinominal Naive Bayes	**Rastgele Orman**	**SVM**
Kelime Vektörleştirme Yöntemi	**TF-IDF (%)**	0,637	0,516	0,713
	BOW (%)	0,694	0,503	0,643
	Word2Vec (%)	0,621	0,495	0,610

TARTIŞMA ve ÖNERİLER

Bu çalışmada literatürde önerilmiş olan İkili Sınıflandırma ve Çoklu Sınıflandırma temelli iki farklı Web sitesi sınıflandırma yöntemi geliştirilmiş, her yönteme uygun makine öğrenmesi ve derin öğrenme algoritmaları ilişkilendirilerek bir Web sitesi sınıflandırma metodolojisi oluşturulmuştur. Devamında, Alexa Top Sites'tan alınmış olan 45.543 Web sitesi ve bu Web sitelerine ait Web sayfaları kullanılarak bir metin sınıflandırma veri seti ve 63 tane Web sitesi sınıfı oluşturulmuş ve bu veri seti kullanılarak, tanımlanmış olan metodoloji veri setine uygulanmıştır. İkili ve Çok Sınıflı Sınıflandırma algoritmalarının ayrı ayrı ve farklı kelime vektörleştirme yöntemleri ile denenmesi; Web sayfalarının sınıflandırılması ve filtrelenmesi problemlerini birlikte ele alınmasını sağlamış olup, mevcut çalışmanın literatürdeki benzer çalışmalardan farkı ortaya konmuştur. Tablo 2 ve 3'te verilmekte olan Başarım oranları değerlendirildiğinde, İkili Sınıflandırmanın istenilen bir Web site sınıfının filtrelenmesi görevi için kullanıldığında, Çoklu Sınıflandırmaya göre daha başarılı sonuçlar ürettiği tespit edilmiştir. İkili Sınıflandırmada uygulanan yöntemlerin mevcut çalışmanın veri seti, çalışmada kullanılan betikler ve donanım özellikleri kapsamındaki işlemsel performansları süre olarak incelendiğinde, Lojistik Regresyon ve Bernoulli Naive Bayes sınıflandırıcılarının, Tam Bağlantılı Yapay Sinir Ağlarına göre 150 kat daha hızlı çalıştığı ve sonuç ürettiği gözlemlenmiştir.

Kullanılan makine öğrenme sınıflandırıcılarının performansları arasındaki fark Tablo 4'te gösterilmiştir. Bu performans farkları, sınıflandırıcıların veri setinde kullanılmış olan metin üzerindeki yetenekleri ve kapasiteleri ile doğrudan ilgili olabileceği düşünülmektedir. Modellerin oluşturulması ve eğitilmesi sırasında her bir Web sayfa sınıfındaki özellik matrisi boyutu için üst bir sınır belirlense de (max features: 50.000), bir çoklu karar ağacı algoritması olan Rastgele Ağaç algoritması için bu boyuttaki bir özellik matrisi, algoritmanın performansını olumsuz yönde etkilemiştir. Çok Sınıflı Sınıflandırma yöntemlerinden SVM'nin, Naive Bayes ve Rastgele Orman sınıflandırıcılarına göre daha iyi sonuç verdiği görülmüştür.

İkili Sınıflandırma algoritmaları kullanılırken karşılaşılmış olan bir problem, her bir sınıfa ait veri setlerinin dengesiz, yani veri setindeki Web site sayısı ve toplam kelime sayısının eşit olmaması, dolayısıyla eğitim sırasında belli sınıfların daha az temsil edilmesi durumudur. Bu problem çalışmadaki Kesinlik, F1 skoru ve Duyarlılık gibi ölçümler ile tespit edilmiştir. İleriki çalışmalarda veri dengesizliğinden kaynaklanan bu sorunun çözümü için kullanılan Web site sayısı ve bu sitelerden elde edilen kelimer ile oluşturulan veri setininin genişletilmesi veya veri üretme gibi tekniklere başvurulabilir. Aynı zamanda İkili Sınıflandırma yaklaşımlarından Tam Bağlantılı Yapay Sinir Ağları'nın diğer sınıflandırıcılara göre daha düşük başarı ile çalışmasının sebebi, verinin yeterli büyükte olmaması, her sınfta yeterli sayıda örnek olmaması ya da dengeli şekilde dağılmaması olabileceği

düşünülmektedir. Gelecek çalışmalarda, toplanan veri setindeki her sınıf için örnek sayısı artırılarak Yapay Sinir Ağlar'ının performansı arttırılabilir. Ayrıca mevcut çalışma İngilizce içerik bulunduran Web siteleri ile yapılmış olup, başta Türkçe olmak üzere, farklı dillerde içerikler bulunduran Web siteleri ile çalışmanın tekrarlanması, kullanılmış olan algoritmaların ve kelime vektörleştirme yöntemlerinin başarımı hakkında daha geniş ve detaylı fikir vereceği düşünülmektedir.

KAYNAKÇA

Chen, Y., Cheng, B. ve Cheng, X. (2016). Food safety document classification using LSTM-based ensemble learning. Revista Técnica de la Facultad de Ingeniería Universidad del Zulia, 39(10), 172-178.

Chen, R. C. ve Hsieh, C. H. (2006). Web page classification based on a support vector machine using a weighted vote schema. Expert Systems with Applications, 31(2), 427-435.

Ester, M., Kriegel, H.-P. ve Schubert, M. (2002). Web site mining: A new way to spot competitors, customers and suppliers in the World Wide Web. KDD '02: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, 249 - 258.

Gali, N., Mariescu-Istodor, R. ve Fränti, P. (2017). Using linguistic features to automatically extract web page title. Expert Systems with Applications, 79, 296-312.

Hartmann, J., Huppertz, J., Schamp, C. ve Heitmann, M. (2019). Comparing automated text classification methods. International Journal of Research in Marketing, 36(1), 20-38.

Hilbe, J. M. (2011). Logistic regression. International encyclopedia of statistical science, 755-758.

Internet Live Stats (2019). "Total Number of Websites", https://www.internetlivestats.com/total-number-of-websites/ (erişim tarihi: 16.05.2019)

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L. ve Brown, D. (2019). Text classification algorithms: A survey. Information, 10(4).

Li, Y. H. ve Jain, A. K. (1998). Classification of text documents. The Computer Journal, 41(8), 537-546.

Loper, E. ve Bird, S. (2002). NLTK: the natural language toolkit. arXiv preprint cs/0205028.

Manning, C., Raghavan, P. ve Schütze, H. (2010). Introduction to information retrieval. Natural Language Engineering, 16(1), 100-103.

Netcraft (2019). "July 2019 Web Server Survey", https://news.netcraft.com/archives/category/web-server-survey/ (erişim tarihi: 16.05.2019)

Onan, A., Korukoğlu, S. ve Bulut, H. (2016). Ensemble of keyword extraction methods and classifiers in text classification. Expert Systems with Applications, 57, 232-247.

Panigrahi, R. ve Borah, S. (2019). Classification and Analysis of Facebook Metrics Dataset Using Supervised Classifiers. In S. Borah, N. Dey, R. Babo & A. S. Ashour (Eds.), Social Network Analytics, Elsevier.

Qi, X., ve Davison, B. D. (2009). Web page classification: Features and algorithms. ACM computing surveys (CSUR), 41(2), 12.

Rajalakshmi, R. ve Aravindan, C. (2018). A Naive Bayes approach for URL classification with supervised feature selection and rejection framework. Computational Intelligence, 34(1), 363-396.

Rekik, R., Kallel, I., Casillas, J. ve Alimi, A. M. (2018). Assessing web sites quality: A systematic literature review by text and association rules mining. International Journal of Information Management, 38(1), 201-216.

Ren, X. Y., Shi, C., Zhang, D. ve Wang, W. S. (2019). An improved SVM web page classification algorithm. In Journal of Physics: Conference Series (Vol. 1187, No. 4, p. 042063). IOP Publishing.

Sebastiani, F. (2002). Machine learning in automated text categorization. ACM computing surveys (CSUR), 34(1), 1-47.

Shen, D., Yang, Q. ve Chen, Z. (2007). Noise reduction through summarization for Web-page classification. Information Processing & Management, 43(6), 1735-1747.

Sinoara, R. A., Camacho-Collados, J., Rossi, R. G., Navigli, R. ve Rezende, S. O. (2019). Knowledge-enhanced document embeddings for text classification. Knowledge-Based Systems, 163, 955-971.

Stein, R. A., Jaques, P. A. ve Valiati, J. F. (2019). An analysis of hierarchical text classification using word embeddings. Information Sciences, 471, 216-232.

Takenouchi, T. ve Ishii, S. (2018). Binary classifiers ensemble based on Bregman divergence for multi-class classification. Neurocomputing, 273, 424-434.

Whittaker, C., Ryner, B. ve Nazif, M. (2010). Large-scale automatic classification of phishing pages. ai.google

Xu, S., Li, Y. ve Wang, Z. (2017). Bayesian multinomial Naïve Bayes classifier to text classification. in Advanced multimedia and ubiquitous engineering. Springer, Singapore, 347-352.

Zhang, J. B., Xu, Z. M., Xiu, K. L. ve Pan, Q. S. (2010). A Web Site Classification Approach Based On Its Topological Structure. Int. J. of Asian Lang. Proc., 20(2), 75-86.