Düzce Üniversitesi Bilim ve Teknoloji Dergisi

Araştırma Makalesi

Türkçe YouTube Yorumları Üzerinde Spam Filtreleme

Sevinj SHİRZADOVA [a], Alper Kürşat UYSAL [b,*]

[a] *Bilgisayar Mühendisliği Anabilim Dalı, Lisansüstü Eğitim Enstitüsü, Eskişehir Teknik Üniversitesi, Eskişehir, TÜRKİYE*

[b] *Bilgisayar Mühendisliği Bölümü, Rafet Kayış Mühendislik Fakültesi, Alanya Alaaddin Keykubat Üniversitesi, Antalya, TÜRKİYE*

** Sorumlu yazarın e-posta adresi: alper.uysal@alanya.edu.tr*

DOI:10.29130/dubited.974309

Öz

Sosyal medya kullanıcıları tarafından en çok tercih edilen platformlardan birisi YouTube'tur. YouTube kullanımının artması beraberinde bazı problemleri de getirmiştir. Genellikle paylaşılan video içerikleriyle alakası olmayan, reklam amaçlı ve sürekli tekrarlayan istenmeyen (spam) yorumlar boşuna kaynak kullanımına sebep olmaktadır. Bu çalışmada, YouTube yorumları üzerinde istenmeyen yorumların otomatik tespit edilmesi amaçlanmaktadır. Metin sınıflandırma problemlerinin çözümü için diğer dillerde gerekli sistemler geliştirilse de Türkçe için yapılan çalışmalar oldukça sınırlıdır. Bu çalışmada Türkçe YouTube yorumlarından oluşan veri setleri oluşturulmuş ve veri setleri üzerinde otomatik metin sınıflandırma algoritmalarının performansları değerlendirilmiştir. Bu çalışmanın önemli bir katkısı da gelecek akademik çalışmalarda kullanılmak üzere erişime açık olacak 5 adet Türkçe veri seti oluşturulmuş olmasıdır. Çalışmada, Weka veri madenciliği aracı kullanılarak doğruluk ve hız açısından iyi sonuçlar veren sınıflandırma algoritmalarının performansları karşılaştırılmıştır. Doğruluk değerleri açısından bakıldığında SMO makine öğrenimi algoritması Türkçe YouTube yorumları sınıflandırma problemi üzerinde diğerlerine göre daha başarılı olarak görünmektedir. Bunun yanısıra öznitelik seçiminin sınıflandırma performansına etkisi araştırılmış ve genellikle az miktarda da olsa sınıflandırma doğruluk değerlerinde iyileşmelere sebep olduğu görülmüştür.

Anahtar Kelimeler: *Spam filtreleme, Metin sınıflandırma, YouTube*

Spam Filtering on Turkish YouTube Comments

Abstract

One of the most preferred platforms by social media users is YouTube. The increase in the use of YouTube has brought some problems with it. Unwanted (spam) comments, which are generally unrelated to the shared video content, for advertising purposes and constantly repetitive, cause useless resource use. In this study, it is aimed to automatically detect unwanted comments on YouTube comments. Although some systems have been developed in other languages to solve text classification problems, studies for Turkish are very limited. In this study, datasets consisting of Turkish YouTube comments were created and the performances of automatic text classification algorithms on the datasets were evaluated. An important contribution of this study is the creation of 5 Turkish datasets that will be public for use in future academic studies. In the study, the performances of classification algorithms that give good results in terms of accuracy and speed were compared using the Weka data mining tool. In terms of accuracy values, the SMO machine learning algorithm seems to be more successful than the others on the classification problem of Turkish YouTube comments. In addition, the effect of feature selection on classification performance has been investigated and it has been observed that it generally leads to slight improvements in classification accuracy.

Keywords: *Spam filtering, Text classification, YouTube*

Geliş: 25/07/2021, Düzeltme: 03/02/2022, Kabul: 07/02/2022

I. GİRİŞ

Günümüzde sosyal medya insanların görüşlerini kolayca ifade edebildikleri küresel bir platforma dönüşmüştür. Sosyal medya, internet kullanıcılarının birbirleriyle bilgi ve görüşlerini görsel ya da işitsel bir şekilde paylaşarak iletişim kurmaları için olanak sağlayan araçlar ve web sitelerini içermektedir [1]. Sahip olduğu avantajlar nedeniyle sosyal medya diğer geleneksel medya platformlarından daha fazla tercih edilerek gelişimini sürdürmekte ve kendine yeni özellikler katarak çok hızlı yayılmaktadır. "We Are Social 2020" raporuna göre Türkiye'de nüfusun %74'ü internet kullanıcısı, nüfusun %64'ü sosyal medya kullanıcısı ve nüfusun %92'si mobil telefon kullanıcısıdır. Bu da 2019 yılı raporuyla kıyaslandığında internet kullanımında %4, sosyal medya kullanımında %4.2 ve mobil telefon kullanımda %3.4 oranlarında artış olduğunu göstermektedir [2]. Sosyal medyanın herkesin söz hakkı olduğu, insanların düşüncelerini özgürce ifade edip topluma etki edebildiği bir platform olması kötü amaçlı kullanıcıların sosyal medyayı olumsuz yönde kullanmalarına neden olmuştur. Bunlara sahte ürün reklamları, kötücül yazılımlar, kötü içerikli linkler vb. örnek olarak gösterilebilir. Genellikle rahatsız edici ve insanlar için tehlike oluşturan bu tür iletilerin spam olarak adlandırıldığı bilinmektedir.

Sosyal medyada popüler ortamlardan biri haline gelen YouTube, 2005 yılında PayPal çalışanları tarafından bir video barındırma web sitesi olarak kurulmuş ve bir sonraki yıl ise Google tarafından satın alınmıştır [3]. Google'ın YouTube için seçtiği "broadcast yourself" yani Türkçe anlamıyla "kendini yayınla" sloganı çok sayıda kullanıcının siteye abone olmasını sağlamıştır. Böylece YouTube, sadece ünlülerin değil, aynı zamanda sıradan insanların da kendine ait videolarını yayınlama imkânı olduğu bir platforma dönüşmüştür. YouTube'un para kazandırma özelliği, kısa sürede kullanıcı sayısının daha çok artmasına neden olmuştur. Bu avantajlarıyla YouTube kötü niyetli kullanıcıların da hedefi olmuştur. Bu kullanıcılar, videolar altına video ile ilgisi olmayan yorumlar yapmakla, sahte vaatlerle kullanıcıları yanlış yönlendirme ve bilgi çalma gibi amaçlarla kullanıcılar için tehlike oluşturacak biçimde karşımıza çıkmaktadır. YouTube, böyle uygunsuz içeriklerin işaretlenmesi ve önlemlerin uygulanması için hem gerçek kişilerden hem de teknolojiden yararlanmaktadır. İşaretlemeler, otomatik işaretleme sistemlerinden, Güvenilir İşaretleyici Programı'nın üyelerinden (STK'lar, devlet kurumları ve şahıslar) veya genel Youtube topluluğundaki kullanıcılardan gelebilir. Yeterli sayıda kullanıcı bir yorumu spam olarak işaretlerse o yorum gizlenir. Ancak bu yöntemler yorum denetleme için yeterli olmadığından spam hacmi gün geçtikçe artmaya devam etmektedir.

Bu çalışmada, Türkçe YouTube yorumları üzerinde otomatik spam tespiti yapılması amaçlanmıştır. Literatürdeki çalışmalar incelendiğinde konu ile ilgili genelde İngilizce iletiler için makine öğrenme yöntemleri kullanılarak spam tespiti yapıldığı görülmektedir. Buna karşın, Türkçe için yapılan çalışmaların çok daha kısıtlı olduğu söylenebilir. Bu çalışmanın genel olarak akademik literatür için iki ana katkısı bulunmaktadır. Bunlardan ilki, oluşturulan veri setlerinin bu çalışma ile birlikte internet ortamında gelecek akademik çalışmalar için araştırmacıların erişimine açılacak olmasıdır. İkinci katkı ise Türkçe YouTube yorumları üzerinde 10 farklı sınıflandırma algoritması içeren şemaların başarılarının kapsamlı şekilde analiz edilmesi ve en başarılı sınıflandırma şemalarının tespit edilmesidir. Çalışma kapsamında, Türkiye'de geçmiş yıllarda çok izlenen 5 YouTube video klibi seçilerek bu kliplere yapılan yorumlar kaydedilmiş ve 5 tane farklı veri seti oluşturulmuştur. Sonrasında, oluşturulan veri setleri araştırmacılar tarafından etiketlenmiştir. En sonunda da her biri üzerinde makine öğrenme algoritmalarının performansları ölçülerek sonuçlar karşılaştırılmış ve böylece daha başarılı sınıflandırma performansları veren sınıflandırma şemaları tespit edilmiştir.

Çalışmanın bundan sonraki akışı şu şekildedir: İkinci bölümde konu ile ilgili literatürdeki çalışmalar ele alınmaktadır. Üçüncü bölümde, deneysel çalışmalarda uygulanan sınıflandırma şemalarının bileşenlerinden bahsedilmektedir. Yapılan deneysel çalışmalar dördüncü bölümde detaylı şekilde ele alınmaktadır. Son bölümde ise çalışmanın sonuçları yorumlanacak ve ileriki muhtemel çalışmalardan bahsedilecektir.

II. LİTERATÜR TARAMASI

Literatürde spam tespiti ile ilgili çalışmaları incelediğimizde web spam tespiti [4], blog spam tespiti [5, 6], spam e-posta tespiti [7, 8, 9, 10], spam SMS mesaj tespiti [11, 12, 13] ile ilgili çok sayıda çalışma olduğunu görmekteyiz. Literatürdeki çalışmaların büyük çoğunluğunda çeşitli standart makine öğrenme yaklaşımları kullanılmakta olup bazı çalışmalarda derin öğrenme teknikleri de uygulanmıştır. Örneğin, Roy vd. spam SMS mesaj tespiti için Evrişimli Sinir Ağları (CNN) ve LSTM algoritması kullanmışlardır [13]. Literatürdeki çalışmalar genelde İngilizce iletilerden oluşan veri setleri üzerinde yapılmış olup Türkçe için bu anlamda çalışma sayısı oldukça azdır. Diğer sosyal ağlarda ve e-postalarda yayılan spam iletilerden farklı olarak YouTube'da yayınlanan spam yorumlar gerçek kullanıcılar tarafından oluşturularak genelde popüler videolarda kendi kendini tanıtmayı hedeflemektedir. Bu nedenle, bu tür mesajlar meşru mesajlara benzeyebilmektedir ve tespit edilmesi daha zordur.

2013 yılında yapılan bir çalışmada, YouTube videolarına yapılan video tipindeki yorumlar içerisinde spam videolar tespit edilmeye çalışılmıştır [14]. Çalışmada, video spam'leri otomatik olarak tespit edebilecek bir yaklaşım önerilmiştir. Bu yaklaşım tek sınıflı sınıflandırma algoritmasına dayanmaktadır. Video tipinde yorumlardan oluşan veri seti oluşturularak önerilen yaklaşım bu veri seti üzerinde denenmiştir. Yapılan deneysel çalışma, önerilen yaklaşımın %80 oranında doğruluk gösterdiğini ortaya çıkarmıştır. Abd, Altabrawe ve Ajmi, 2018 yılında yaptıkları çalışmada YouTube spam yorumlarını tespit etmek için Yapay Sinir Ağı modelleri kullanmıştır [15]. Bu çalışmada, aynı amaçla daha önce Alberto [16] tarafından yapılan başka bir çalışmayla karşılaştırma yapılmıştır. Alberto tarafından yapılan çalışmada TubeSpam adı verilen otomatik spam filtreleme yöntemi önerilmiş ve 5 tane en çok dinlenen şarkının yorumlarından oluşturulmuş 5 adet veri seti kullanılmıştır. Burada da yine aynı veri setleri kullanılarak Yapay Sinir Ağı modeli kullanılarak sınıflandırma yapılmış ve sonuçların Alberto'nun yaptığı çalışmadan daha iyi olduğu görülmüştür. Samsudin vd. 2019 yılında yaptıkları çalışmada YouTube spam yorumların tespiti için Basit Bayes ve Lojistik Regresyon sınıflandırma algoritmalarının performanslarını Weka ve Rapid Miner veri madenciliği programlarında karşılaştırmıştır [17]. Weka'da Basit Bayes ve Lojistik Regresyon %87.21 ve %85.29 oranlarında doğru sonuç verirken, Rapid Miner'da Basit Bayes ve Lojistik Regresyon %80.41 ve %80.88 oranlarında doğru sonuç vermiştir. Sonuç olarak Basit Bayes algoritmasının daha yüksek sınıflandırma doğruluğu verdiği görülmüştür. Uysal tarafından 2018 yılında yapılan bir başka çalışmada YouTube spam filtreleme için 2 farklı sınıflandırma algoritması kullanılarak 5 adet öznitelik seçim metodunun performansları analiz edilmiştir [18]. Sınıflandırıcı olarak Basit Bayes ve Karar Ağaçları algoritmaları kullanılmıştır. Veri seti olarak Alberto'nun sunduğu 5 şarkıcının klibine yapılan İngilizce yorumları içeren 5 farklı veri seti kullanılmıştır. Değerlendirme için Makro-F1 başarı ölçütü ve performans değerlendirilmesi için 3-katmanlı çapraz doğrulama yöntemi kullanılmıştır. Sonuçlar, Distinguishing Feature Selector (DFS) ve Gini Index (GI) metotlarının diğer üç öznitelik seçim yöntemine göre daha yüksek sınıflandırma doğruluğu verdiğini göstermiştir. Bununla beraber YouTube spam filtrelemede genellikle Karar Ağaçları algoritmasının Basit Bayes algoritmasından daha yüksek sınıflandırma doğruluğu verdiği görülmüştür.

Literatürde, Türkçe YouTube verileri üzerinde spam tespiti ile alakalı hemen hemen hiç çalışma bulunmamaktadır. Bunun nedenlerinden birisinin araştırmacıların erişimine açık Türkçe YouTube yorumları içerikli veri setinin bulunmaması olduğu değerlendirilmektedir.

III. SINIFLANDIRMA ŞEMALARI

Bu çalışmada öncelikle sınıflandırıcı performansları bölümün devamında anlatılan 10 farklı sınıflandırıcı ile analiz edilmiştir. Bir diğer yandan ise öznitelik seçiminin sınıflandırıcı başarımlarına etkisi araştırılmıştır. Çalışmada kullanılan sınıflandırma algoritmaları ve öznitelik seçim yöntemleri ile ilgili açıklamalar sonraki alt bölümlerde yer almaktadır.

A. SINIFLANDIRMA ALGORİTMALARI

Verilerin ön işlemlerden geçmesinden sonra Weka yazılım kütüphanesi [19] içerisindeki sınıflandırma algoritmalarından faydalanılarak deneysel çalışma sürdürülmüştür. Çalışmada, Weka içerisinde yer alan J48, Rastgele Orman (Random Forest), REPTree, Karar Tablosu (Decision Table), JRip, IBk, SMO, Bayes Ağı (BayesNet), Basit Bayes (Naive Bayes), Çok Terimli Basit Bayes (Multinominal Naive Bayes) isimli sınıflandırma algoritmaları kullanılmıştır. Aşağıda, bu çalışmada kullanılan sınıflandırma algoritmalarının çalışma mantıkları kısaca açıklanmıştır.

J48: Bir karar ağacı algoritması olup makine öğreniminde yaygın kullanılan C4.5 algoritmasının Weka uyarlamasıdır. Ağaç yapısı düğüm ve dallardan oluşup, düğümler sınıf etiketlerini ifade eder [20].

Rastgele Orman (Random Forest): Çok sayıda karar ağacı oluşturarak sınıflandırma başarısını artırmayı amaçlayan bir algoritmadır [21]. Karar ağaçlarının her biri farklı eğitim verileriyle eğitilerek elde edilen sonuçlar birleştirilir ve yeni gelen örneğin sınıflandırılması bu bireysel sınıflandırma ağaçlarının tahminlerinden elde edilen oylara dayanarak yapılır.

REPTree: Regresyon ağacı mantığına dayanan, farklı yinelemelerle çok sayıda karar ağacı oluşturup aralarından en iyisini seçmeyi hedefleyen bir karar ağacı algoritmasıdır [22]. REPTree algoritması varyanstan kaynaklanan hatayı en aza indirme ve entropi ile bilgi kazanımı ilkesini benimsemektedir.

Karar Tablosu (Decision Table): Oluşturduğu karar tablosuna göre sınıflandırma yapan bir algoritmadır [23]. Eğitim setindeki verilerin özelliklerine dayanarak karar tablosu oluşturur. Yeni gelen örnek, oluşturduğu karar tablosundaki kurallara göre sınıflandırılır.

JRip: Tekrarlayan artımlı budama yöntemiyle hatayı azaltma mantığına dayanan algoritma, kural tabanlı sınıflandırma tekniği kullanır [23]. JRip algoritması önce eğitim setini budama ve gelişen kurallar listesi olmak üzere iki alt kümeye ayırır, daha sonra gelişen kurallar listesinde yer alan örneklere göre bir kural oluşturur. Bu işlemden sonra kurallar listesi artık yeni gelen bir örneğin sınıflandırılması için kullanılabilir.

IBk: K-en yakın komşu algoritmasıdır [24]. Burada k farklı değerler alabilir. En yakın mesafede olan komşuların sınıfına bakılarak sınıflandırma işlemi yapılmaktadır.

SMO: Sıralı minimal optimizasyon algoritması olan SMO algoritması Destek Vektör Makinesi tabanlıdır [24]. Standart Destek Vektör Makinesi algoritmasından farklı olarak analitik kuadratik programlama tekniği kullanan SMO algoritması ekstra matrise ihtiyaç duymadan kuadratik programlama problemlerini hızlı bir şekilde çözebilmektedir.

Bayes Ağı (BayesNet): Bayes ağı, istatistiksel modelin bir çeşidi olan olasılıksal yönlü dönüşsüz çizge modelidir [24].

Basit Bayes (Naive Bayes): Bayes teoremine dayanan olasılıksal sınıflandırma yöntemidir [25].

Çok Terimli Basit Bayes (Multinominal Naive Bayes): Basit Bayes sınıflandırıcısının bir çeşidi olup genelde metin sınıflandırma problemleri için kullanılır [25]. Sınıflandırıcının kullandığı öngörücüler metin dokümanlarında bulunan kelimelerin geçme sıklığıdır.

B. ÖZNİTELİK SEÇİM YÖNTEMLERİ

Bu çalışmada Weka'da bulunan 2 adet öznitelik seçim yöntemi deneysel bölümde kullanılmıştır. Aşağıda, bu yöntemlerin çalışma mantıkları kısaca açıklanmıştır.

Korelasyon tabanlı Öznitelik Altküme Seçim Değerlendirici (Correlation based Feature Subset Selection-CfsSubsetEval): Filtreleme mantığına dayanan bu öznitelik seçim algoritması, alt özellik kümelerini korelasyon bazlı değerlendirip aralarından en iyi olanı bulmaya çalışır [26].

Kazanım Oranı Öznitelik Değerlendirici (GainRatioAttributeEval): Bir özniteliğin değerini, özniteliğin sınıfa göre kazanım oranını ölçerek hesaplar [27].

Çalışmada, Weka ortamında öznitelik seçim yöntemleri uygulanırken bu yöntemlerle birlikte 2 farklı araştırma algoritması uygulanmıştır. Aşağıda, bu yöntemlerin çalışma mantıkları kısaca açıklanmıştır.

Açgözlü Adımsal Araştırma (GreedyStepwise): Öznitelik alt kümesi üzerinde ileriye veya geriye doğru açgözlü araştırma yapar [19]. Öznitelik uzayındaki herhangi bir noktadan başlayabilir. Kalan özniteliklerin eklenmesi veya silinmesi değerlendirmede bir düşüşe neden oluncaya kadar devam eder.

Sıralayıcı (Ranker): Her bir özniteliği ayrı ayrı değerlendirerek sıralama yapar [27]. Weka ortamında GainRatioAttributeEval, ClassifierAttributeEval, OneRAttributeEval gibi öznitelik değerlendiricileri ile beraber kullanılır.

IV. DENEYSEL ÇALIŞMA

Bu bölümde gerekli verilerin toplanması, işlenmesi ve gerçekleştirilen deneylere yönelik detaylar anlatılmaktadır. Öncelikle, ön işlemden geçirilen veri setleri Weka'da kullanılabilmeleri için ARFF uzantılı dosya formatına çevrilmiştir. Ayrıca, deneylerde TF-IDF [28] ile ağırlıklandırma yapılmıştır. TF-IDF yöntemi, dokümanlardaki terimleri terim frekansına (tf) ve ters doküman frekansına (idf) göre ağırlıklandıran bir ağırlıklandırma şemasıdır. Çalışmada 10 farklı sınıflandırma algoritması kullanılmıştır. Weka'da hem ön işlemden geçirilmemiş veri, hem de ön işlemden geçirilmiş verinin sınıflandırma sonuçları karşılaştırılmıştır. Daha sonra ise öznitelik seçim metotlarının sınıflandırmaya etkisini araştırmak amacıyla 2 farklı öznitelik seçim yöntemi uygulanmıştır. Öznitelik seçiminde amaç, uygun olmayan özniteliklerin elenip sınıflandırma performansının iyileştirilmesi ve işlem yükünün azaltılmasıdır.

A. VERİ SETİNİN OLUŞTURULMASI

Literatürdeki ilgili çalışmalar araştırıldığında Türkçe YouTube yorumları üzerinde herhangi çalışma bulunmamakla beraber Türkçe YouTube yorumlarını içeren veri setinin de olmadığı ortaya çıkmıştır. Bu yüzden çalışmaya öncelikle veri setini oluşturulmasıyla başlanmıştır. Bunun için 2017 yılının en çok izlenen video klipleri içerisinden 5 tanesi seçilmiş ve bu kliplere yapılan yorumlar Youtube Comment Scraper (YouTube Yorum Ayrıştırıcı) aracı yardımıyla indirilerek kaydedilmiştir. Daha sonra her bir dosyada yorum metinleri spam ve normal olarak el yordamıyla etiketlenerek .txt dosya formatında ilgili dokümanlara kaydedilmiştir. Böylelikle spam ve normal yorumlar içeren 5 tane veri seti oluşturulmuştur. Veri setleri hakkında detaylar aşağıdaki tabloda gösterilmiştir. Oluşturulan 5 adet veri setini Weka formatında içeren ve TurkishYouTubeSpamDoc v1.0 adı verilen koleksiyon bu çalışma ile birlikte https://drive.google.com/file/d/1i-12McY9LfHlH6_kyJl43yMQQ7pehLix/view?usp=sharing adresinde araştırmacıların erişimine açılmıştır. Veri setlerindeki sınıflar bazındaki dağılım Tablo 1'de gösterilmektedir.

Tablo 1. *Veri setleri*

Veri seti	Spam	Normal	Toplam
Ece Seçkin	135	269	404
Gülşen	173	194	367
Hadise	145	189	334
Hande Yener	94	179	273
Tarkan	147	184	331

Ece Seçkin isimli veri setinin ön işlem uygulanmamış ve uygulanmış versiyonlarında sırasıyla 1849 ve 1187 adet öznitelik bulunmaktadır. Gülşen isimli veri setinin ön işlem uygulanmamış ve uygulanmış versiyonlarında sırasıyla 1558 ve 1036 adet öznitelik bulunmaktadır. Hadise isimli veri setinin ön işlem uygulanmamış ve uygulanmış versiyonlarında sırasıyla 1608 ve 1053 adet öznitelik bulunmaktadır. Hande Yener isimli veri setinin ön işlem uygulanmamış ve uygulanmış versiyonlarında sırasıyla 1378 ve 860 adet öznitelik bulunmaktadır. Tarkan isimli veri setinin ön işlem uygulanmamış ve uygulanmış versiyonlarında sırasıyla 1476 ve 934 sadet öznitelik bulunmaktadır.

B. SINIFLANDIRMA ŞEMALARININ BAŞARIM ANALİZİ

Veri madenciliğinde, kullanılan verinin kalitesinin sonuçları doğrudan etkilediği bilinmektedir. Çalışmamızda, sınıflandırma algoritmalarının başarımını artırmak adına bir takım ön işlemler gerçekleştirilmiştir. Ön işlem adımları olarak küçük harf dönüşümü, gereksiz işaretlerin ve durak kelimelerin metinden çıkarılması, dizgeciklere ayırma (tokenization), gövdeleme (stemming) uygulanmıştır. Gövdeleme için Zemberek Türkçe dil işleme kütüphanesinden faydalanılmıştır [29]. Yapılan işlemler Java programlama dili kullanılarak gerçekleştirilmiştir.

Verinin ön işlemden geçirilmesinin sonuçlara etkisini daha açık şekilde gözlemlemek için ön işlem uygulanmayan ve ön işlem uygulanan veri seti üzerinde sınıflandırma algoritmaları uygulanmıştır. Sınıflandırma sonuçlarının sunulduğu tablolarda veri setlerinin isminin sonuna "-S" eklenerek gösterilen sonuçlar ön işlem uygulanan durumları temsil edilmektedir. Sınıflandırma başarımını göstermek için doğruluk oranlarından faydalanılmıştır. 5 veri setinin her iki durumu için de sonuçlar Tablo 2-6'da gösterilmektedir. Tablolar içerisinde en iyi sınıflandırma doğruluğunu temsil eden sonuçlar kalın font ile gösterilmiştir. Ayrıca, model oluşum zamanları sistemin eğitim süresini temsil etmektedir.

Tablo 2'deki Ece veri setinde SMO, JRip, Basit Bayes (Naive Bayes), REPTree algoritmalarının %90 üzerinde doğruluk gösterdiği görülmektedir. Ece-S veri setinde, SMO, Çok Terimli Basit Bayes (NaiveBayesMultinominal- NBM), Rastgele Orman (RandomForest), Basit Bayes (Naive Bayes), JRip, Bayes Ağı (BayesNet) algoritmalarının %90 üzerinde doğruluk gösterdiği görülmektedir.

Tablo 3'te yer alan sınıflandırma doğruluğu değerleri göz önüne alındığında; Gülşen veri seti için, REPTree, Karar Tablosu (DecisionTable), JRip ve SMO sınıflandırma algoritmalarının diğer algoritmalara nazaran nispeten daha yüksek doğruluk sağladığı görülmektedir. NBM ve IBk algoritmaları hızlı olsa da doğruluk oranları az daha düşük olmuştur. Gülşen veri seti üzerinde, ön işlem gerçekleştirilen deneylerde, ön işlem gerçekleştirilmeyenlere nazaran tüm sınıflandırıcıların sınıflandırma performanslarının genel olarak yüzde 5 oranında daha yüksek olduğu gözlenmiştir. SMO, J48, Rasgele Orman (RandomForest), Karar Tablosu (DecisionTable), JRip, Bayes Ağı (BayesNet) daha yüksek başarı oranı gösteren algoritmalardandır.

Tablo 2. *Ece ve Ece-S veri setlerinin sınıflandırma sonuçları*

	Ece				Ece-S			
Sınıflandırma algoritması	Doğru olarak sınıflandırılan doküman sayısı	Yanlış olarak sınıflandırılan doküman sayısı	Doğruluk yüzdesi (%)	Model oluşum zamanı	Doğru olarak sınıflandırılan doküman sayısı	Yanlış olarak sınıflandırılan doküman sayısı	Doğruluk yüzdesi (%)	Model oluşum zamanı
J48	335	69	82.92	3.68 sn	354	50	87.62	1.4 sn
RandomForest	359	45	88.86	11.2 sn	378	26	93.56	6.46 sn
REPTree	364	40	90.10	3.12 sn	373	31	92.33	1.73 sn
DecisionTable	360	44	89.11	41.3 sn	370	34	91.58	18.7 sn
JRip	371	33	91.83	9.14 sn	376	28	93.07	5.13 sn
İbk	350	54	86.63	0.01 sn	356	48	88.12	0.01 sn
SMO	373	31	**92.33**	0.37 sn	380	24	94.06	0.34 sn
BayesNet	362	42	89.60	0.9 sn	376	28	93.07	0.54 sn
NaiveBayes	370	34	91.58	0.52 sn	378	26	93.56	0.37 sn
NBM	357	47	88.37	0.01 sn	381	23	**94.31**	0.01 sn

Tablo 3. *Gülşen ve Gülşen-S veri setlerinin sınıflandırma sonuçları*

	Gülşen				Gülşen-S			
Sınıflandırma algoritması	Doğru olarak sınıflandırılan doküman sayısı	Yanlış olarak sınıflandırılan doküman sayısı	Doğruluk yüzdesi (%)	Model oluşum zamanı	Doğru olarak sınıflandırılan doküman sayısı	Yanlış olarak sınıflandırılan doküman sayısı	Doğruluk yüzdesi (%)	Model oluşum zamanı
J48	331	36	90.19	2.33 sn	353	14	**96.19**	1.25 sn
RandomForest	333	34	90.74	6.43 sn	351	16	95.64	4.94 sn
REPTree	335	32	91.28	1.96 sn	350	17	95.37	0.91 sn
DecisionTable	337	30	**91.83**	16.8 sn	352	15	95.91	9.91 sn
JRip	336	31	91.55	3.58 sn	352	15	95.91	2.29 sn
İbk	333	34	90.74	0.01 sn	350	17	95.37	~ 0 sn
SMO	336	31	91.55	0.49 sn	353	14	**96.19**	0.2 sn
BayesNet	329	38	89.65	0.61 sn	352	15	95.91	0.53 sn
NaiveBayes	323	44	88.01	0.4 sn	348	19	94.82	0.29 sn
NBM	327	40	89.10	0.01 sn	347	20	94.55	0.01 sn

Tablo 4. *Hadise ve Hadise-S veri setlerinin sınıflandırma sonuçları*

	Hadise				Hadise-S			
Sınıflandırma algoritması	Doğru olarak sınıflandırılan doküman sayısı	Yanlış olarak sınıflandırılan doküman sayısı	Doğruluk yüzdesi (%)	Model oluşum zamanı	Doğru olarak sınıflandırılan doküman sayısı	Yanlış olarak sınıflandırılan doküman sayısı	Doğruluk yüzdesi (%)	Model oluşum zamanı
J48	302	32	90.42	1.89 sn	303	31	90.72	1.16 sn
RandomForest	303	31	90.72	6.61 sn	314	20	94.01	4.75 sn
REPTree	302	32	90.42	1.67 sn	302	32	90.42	1.24 sn
DecisionTable	302	32	90.42	20.24 sn	306	28	91.62	13.32sn
JRip	303	31	90.72	4.13 sn	310	24	92.81	2.65 sn
İbk	299	35	89.52	0.04 sn	309	25	92.52	~ 0 sn
SMO	306	28	**91.62**	0.36 sn	315	19	**94.31**	0.21 sn
BayesNet	300	34	89.82	0.72 sn	309	25	92.52	0.35 sn
NaiveBayes	303	31	90.72	0.38 sn	311	23	93.11	0.28 sn
NBM	300	34	89.82	0.01 sn	314	20	94.01	~ 0 sn

Tablo 5. *Hande ve Hande-S veri setlerinin sınıflandırma sonuçları*

	Hande				Hande-S			
Sınıflandırma algoritması	Doğru olarak sınıflandırılan doküman sayısı	Yanlış olarak sınıflandırılan doküman sayısı	Doğruluk yüzdesi (%)	Model oluşum zamanı	Doğru olarak sınıflandırılan doküman sayısı	Yanlış olarak sınıflandırılan doküman sayısı	Doğruluk yüzdesi (%)	Model oluşum zamanı
J48	240	33	87.91	2 sn	257	16	94.14	1.12 sn
RandomForest	251	22	91.94	6.45 sn	261	12	95.60	3.56 sn
REPTree	248	25	90.84	1.54 sn	257	16	94.14	0.74 sn
DecisionTable	246	27	90.11	19.83sn	256	17	93.77	8.88 sn
JRip	247	26	90.48	4.19 sn	262	11	**95.97**	1.5 sn
İbk	160	113	58.61	0.02 sn	225	48	83.52	~ 0 sn
SMO	252	21	**92.31**	0.27 sn	262	11	**95.97**	0.21 sn
BayesNet	243	30	89.01	0.56 sn	258	15	94.51	0.34 sn
NaiveBayes	250	23	91.58	0.36 sn	257	16	94.14	0.21 sn
NBM	231	42	84.62	0.01 sn	238	35	87.18	0.01 sn

Tablo 6. *Tarkan ve Tarkan-S veri setlerinin sınıflandırma sonuçları*

	Tarkan				Tarkan-S			
Sınıflandırma algoritması	Doğru olarak sınıflandırılan doküman sayısı	Yanlış olarak sınıflandırılan doküman sayısı	Doğruluk yüzdesi (%)	Model oluşum zamanı	Doğru olarak sınıflandırılan doküman sayısı	Yanlış olarak sınıflandırılan doküman sayısı	Doğruluk yüzdesi (%)	Model oluşum zamanı
J48	286	45	86.40	1.77 sn	313	18	94.56	1.62 sn
RandomForest	303	28	91.54	6.92 sn	312	19	94.26	4 sn
REPTree	303	28	91.54	1.74 sn	313	18	94.56	0.84 sn
DecisionTable	303	28	91.54	22.16sn	308	23	93.05	12.89 sn
JRip	308	23	**93.05**	4.67 sn	312	19	94.26	2.21 sn
İbk	293	38	88.52	0 sn	308	23	93.05	0.02 sn
SMO	307	24	92.75	0.23 sn	314	17	94.86	0.34 sn
BayesNet	291	40	87.92	0.64 sn	309	22	93.35	0.46 sn
NaiveBayes	301	30	90.94	0.38 sn	317	14	**95.77**	0.24 sn
NBM	301	30	90.94	0.04 sn	306	25	92.45	0.01 sn

Tablo 4'te, Hadise veri setinde algoritmaların birbirine yakın sonuçlar verdiği görülmektedir. SMO, Basit Bayes (NaiveBayes), JRip, Rastgele Orman (RandomForest) sınıflandırıcıları diğerlerine göre nispeten daha doğru sonuç göstermiştir. Hadise-S veri setinde SMO, NBM ve Rastgele Orman (RandomForest) algoritmaları yüksek doğruluk göstermiştir.

Tablo 5'teki Hande veri setinde SMO, Rastgele Orman (RandomForest), Basit Bayes'in doğru sınıflandırma yüzdeleri diğerlerine göre daha iyi görünmektedir. SMO ve Basit Bayes (NaiveBayes) aynı zamanda hızlı sonuç veren algoritmalardır. Hande-S veri setinde en yüksek doğru sınıflandırma yüzdesi JRip ve SMO algoritmaları ile elde edilmiştir.

Tablo 6'daki Tarkan veri setine baktığımız zaman JRip ve SMO algoritmalarının daha yüksek doğruluk oranı gösterdiğini görebiliriz. Tarkan-S veri setinde algoritmalar genellikle yakın sınıflandırma doğruluk yüzdelerine sahiptir fakat Basit Bayes (NaiveBayes) algoritmasının doğru sınıflandırma oranı diğerlerinden daha yüksektir.

Tablo 2-6'daki sonuçlara baktığımızda genel olarak tüm algoritmaların ön işlem uygulanmış veri seti üzerinde sınıflandırma performansının ön işlem uygulanmayan veri seti üzerindekine göre çok daha iyi olduğu görünmektedir. Ayrıca, veri setlerinin hemen hemen hepsinde SMO algoritmasının diğerlerine göre daha başarılı sınıflandırma performansları verdiği görülmektedir. SMO algoritması hem hız hem de doğruluk yüzdesi açısından başarılı olmuştur.

Deneylerin devamında öznitelik seçimi metotlarının sınıflandırma üzerindeki etkisi incelenmiştir. Weka'da bulunan metotlar uygulanarak öznitelik seçiminibn sınıflandırma performansını nasıl etkilediği araştırılmıştır. Bunun için 2 tane öznitelik değerlendirici ve araştırma yöntemi kombinasyonu kullanılmıştır. Öznitelik seçimi yöntemleri kullanılarak yapılan sınıflandırma sonuçları Tablo 7-16'da ifade edilmiştir. Tablolarda varsayılan olarak kullanılan gösterim, öznitelik sayısı için

bir sabit sayı belirtilmediği ve Weka programının belli bir eşik değerine göre seçilen öznitelik sayısını ayarladığı durumu temsil etmektedir.

Tablo 7. *Ece veri setinin özellik seçim metotları ile sınıflandırma sonuçları*

Sınıflandırma algoritması	CfsSubsetEval + GreedyStepwise (varsayılan)	Model oluşum zamanı	GainRAE + Ranker (varsayılan)	Model oluşum zamanı	GainRAE + Ranker (seçilen özellik sayısı: 100)	Model oluşum zamanı
J48	81.19	2.15 sn	82.92	3.3 sn	81.19	0.85 sn
RandomForest	91.58	2.74 sn	88.86	12.12 sn	93.07	2.16 sn
REPTree	89.60	2.32 sn	90.10	3.88 sn	89.85	0.71 sn
DecısıonTable	89.60	3.19 sn	89.11	43.9 sn	90.35	2.44 sn
JRip	91.34	1.84 sn	91.83	11.71 sn	92.08	1.11 sn
İBk	91.34	2.06 sn	86.63	0.81 sn	91.83	0.73 sn
SMO	91.34	1.91 sn	92.33	1.01 sn	**92.57**	1.09 sn
BayesNet	88.61	1.95 sn	89.60	1.47 sn	89.85	0.78 sn
NaıveBayes	89.36	1.79 sn	91.58	1.25 sn	91.34	0.74 sn
NBM	90.59	2.01 sn	88.37	0.98 sn	91.09	0.75 sn

Tablo 8. *Ece-S veri setinin özellik seçim metotları ile sınıflandırma sonuçları*

Sınıflandırma algoritması	CfsSubsetEval + GreedyStepwise (varsayılan)	Model oluşum zamanı	GainRAE + Ranker (varsayılan)	Model oluşum zamanı	GainRAE + Ranker (seçilen özellik sayısı: 100)	Model oluşum zamanı
J48	87.38	1.18 sn	87.62	2.23 sn	87.38	0.61 sn
RandomForest	92.08	1.86 sn	93.56	8.73 sn	**94.55**	2.19 sn
REPTree	90.84	1.33 sn	92.33	2.1 sn	92.08	0.71 sn
DecısıonTable	89.11	1.16 sn	91.58	22.13 sn	92.33	1.95
JRip	91.09	0.95 sn	93.07	5.19 sn	92.82	1.03 sn
İBk	91.58	1.14 sn	88.12	0.48 sn	93.07	0.43 sn
SMO	91.09	1.31 sn	94.06	0.88 sn	**94.55**	0.64 sn
BayesNet	89.36	1.12 sn	93.07	1.22 sn	93.32	0.53 sn
NaıveBayes	89.60	1.19 sn	93.56	0.9sn	92.57	0.61 sn
NBM	91.34	1.32 sn	94.31	0.6 sn	94.31	0.44 sn

Tablo 9. *Gülşen veri setinin özellik seçim metotları ile sınıflandırma sonuçları*

Sınıflandırma algoritması	CfsSubsetEval + GreedyStepwise (varsayılan)	Model oluşum zamanı	GainRAE + Ranker (varsayılan)	Model oluşum zamanı	GainRAE + Ranker (seçilen özellik sayısı: 100)	Model oluşum zamanı
J48	90.19	2.03 sn	90.19	2.83 sn	90.19	0.64 sn
RandomForest	91.01	2.45 sn	90.74	8.9 sn	91.83	2.25 sn
REPTree	91.28	2.39 sn	91.28	2.87 sn	91.28	0.45 sn
DecısıonTable	91.83	2.91 sn	91.83	24.53 sn	91.83	2.08 sn
JRip	91.83	2.36 sn	91.55	5.45 sn	91.83	0.82 sn
İBk	90.19	2.32 sn	90.74	0.58 sn	90.19	0.57 sn
SMO	91.01	2.29 sn	91.55	1 sn	**92.10**	0.63 sn
BayesNet	90.19	2.22 sn	89.65	0.92 sn	89.65	0.71 sn
NaıveBayes	89.37	2.02 sn	88.01	0.89 sn	87.47	0.41 sn
NBM	88.83	2.5 sn	89.10	0.71 sn	88.01	0.44 sn

Tablo 10. *Gülşen-S veri setinin özellik seçim metotları ile sınıflandırma sonuçları*

Sınıflandırma algoritması	CfsSubsetEval + GreedyStepwise (varsayılan)	Model oluşum zamanı	GainRAE + Ranker (varsayılan)	Model oluşum zamanı	GainRAE + Ranker (seçilen özellik sayısı: 100)	Model oluşum zamanı
J48	**96.19**	1.11 sn	**96.19**	1.93 sn	**96.19**	0.5 sn
RandomForest	**96.19**	1.29 sn	95.64	6.57 sn	95.91	2.71 sn
REPTree	95.64	1.27 sn	95.37	1.49 sn	95.64	0.52 sn
DecısıonTable	95.91	1.29 sn	95.91	13.74 sn	95.91	1.62 sn
JRip	**96.19**	1.22 sn	95.91	2.87 sn	95.91	0.55 sn
İBk	95.64	1.06 sn	95.37	0.35 sn	94.55	0.51 sn
SMO	**96.19**	1.1 sn	**96.19**	0.57 sn	95.91	0.6 sn
BayesNet	95.91	1.2 sn	95.91	0.76 sn	95.91	0.69 sn
NaıveBayes	95.64	1.44 sn	94.82	0.62 sn	94.28	0.41 sn
NBM	94.82	1.16 sn	94.55	0.45 sn	90.74	0.4 sn

Tablo 11. *Hadise veri setinin özellik seçim metotları ile sınıflandırma sonuçları*

Sınıflandırma algoritması	CfsSubsetEval + GreedyStepwise (varsayılan)	Model oluşum zamanı	GainRAE + Ranker (varsayılan)	Model oluşum zamanı	GainRAE + Ranker (seçilen özellik sayısı: 100)	Model oluşum zamanı
J48	89.22	1.62 sn	90.42	3.15 sn	90.42	0.7 sn
RandomForest	89.82	2.36 sn	90.72	7.87 sn	91.02	5.2 sn
REPTree	88.92	2.02 sn	90.42	1.83 sn	90.42	1.31 sn
DecısıonTable	88.92	2.13 sn	90.42	27.47 sn	90.42	12.73 sn
JRip	89.22	2.2 sn	90.72	4.77 sn	91.32	0.86 sn
İBk	88.92	1.62 sn	89.52	0.59 sn	91.62	0.5 sn
SMO	89.82	2.42 sn	91.62	0.75 sn	**91.92**	0.62 sn
BayesNet	88.32	2.31 sn	89.82	1.18 sn	89.82	0.63 sn
NaıveBayes	88.92	1.79 sn	90.72	0.67 sn	90.72	0.65 sn
NBM	88.62	2.25 sn	89.82	0.59 sn	91.02	0.53 sn

Tablo 12. *Hadise-S veri setinin özellik seçim metotları ile sınıflandırma sonuçları*

Sınıflandırma algoritması	CfsSubsetEval + GreedyStepwise (varsayılan)	Model oluşum zamanı	GainRAE + Ranker (varsayılan)	Model oluşum zamanı	GainRAE + Ranker (seçilen özellik sayısı: 100)	Model oluşum zamanı
J48	88.62	1.09 sn	90.72	1.49 sn	90.12	0.41 sn
RandomForest	89.82	1.88 sn	94.01	5.28 sn	94.01	1.39 sn
REPTree	89.52	1.34 sn	90.42	1.59 sn	91.02	0.55 sn
DecısıonTable	89.82	1.47 sn	91.62	13.64 sn	91.62	1.69 sn
JRip	89.82	1.49 sn	92.81	3.93 sn	93.41	0.89 sn
İBk	89.22	1.24 sn	92.52	0.45 sn	93.41	0.45 sn
SMO	89.82	1.31 sn	**94.31**	0.58 sn	93.11	0.39 sn
BayesNet	89.52	1.3 sn	92.52	1.03 sn	92.52	0.39 sn
NaıveBayes	89.52	1.21 sn	93.11	0.63sn	93.11	0.46 sn
NBM	88.62	1.3 sn	94.01	0.29 sn	93.71	0.41 sn

Tablo 13. *Hande veri setinin özellik seçim metotları ile sınıflandırma sonuçları*

Sınıflandırma algoritması	CfsSubsetEval + GreedyStepwise (varsayılan)	Model oluşum zamanı	GainRAE + Ranker (varsayılan)	Model oluşum zamanı	GainRAE + Ranker (seçilen özellik sayısı: 100)	Model oluşum zamanı
J48	84.62	3.3 sn	87.91	2.4 sn	87.91	0.48 sn
RandomForest	90.84	3.28 sn	91.94	7.17 sn	93.04	1.76 sn
REPTree	90.48	2.83 sn	90.84	1.81 sn	91.58	0.47 sn
DecısıonTable	90.48	3.11 sn	90.11	17.08 sn	90.84	1.61 sn
JRip	90.48	2.74 sn	90.48	4.79 sn	91.94	0.84 sn
İBk	89.01	2.76 sn	58.61	0.44 sn	90.48	0.44 sn
SMO	90.84	3.14 sn	92.31	0.68 sn	**92.68**	0.49 sn
BayesNet	87.91	2.57 sn	89.01	1.08 sn	89.01	0.5 sn
NaıveBayes	89.01	2.88 sn	91.58	0.65 sn	91.94	0.43 sn
NBM	89.38	3.22 sn	84.62	0.52 sn	90.84	0.46 sn

Tablo 14. *Hande-S veri setinin özellik seçim metotları ile sınıflandırma sonuçları*

Sınıflandırma algoritması	CfsSubsetEval + GreedyStepwise (varsayılan)	Model oluşum zamanı	GainRAE + Ranker (varsayılan)	Model oluşum zamanı	GainRAE + Ranker (seçilen özellik sayısı: 100)	Model oluşum zamanı
J48	93.77	0.49 sn	94.14	1.12 sn	94.14	0.41 sn
RandomForest	95.60	0.93 sn	95.60	4.04 sn	**96.70**	1.38 sn
REPTree	94.14	0.59 sn	94.14	0.89 sn	94.51	0.43 sn
DecısıonTable	93.77	0.66 sn	93.77	9.35 sn	93.77	1.76 sn
JRip	95.60	0.64 sn	95.97	1.62 sn	95.97	0.47 sn
İBk	95.24	0.72 sn	83.52	0.3 sn	94.14	0.22 sn
SMO	95.60	0.87 sn	95.97	0.32 sn	96.34	0.54 sn
BayesNet	92.67	0.56 sn	94.51	0.58 sn	94.51	0.29 sn
NaıveBayes	95.24	0.64 sn	94.14	0.47sn	94.51	0.28 sn
NBM	94.14	0.68 sn	87.18	0.2 sn	91.58	0.21 sn

Tablo 15. *Tarkan veri setinin özellik seçim metotları ile sınıflandırma sonuçları*

Sınıflandırma algoritması	CfsSubsetEval + GreedyStepwise (varsayılan)	Model oluşum zamanı	GainRAE + Ranker (varsayılan)	Model oluşum zamanı	GainRAE + Ranker (seçilen özellik sayısı: 100)	Model oluşum zamanı
J48	85.50	1.63 sn	86.40	1.93 sn	85.80	0.81 sn
RandomForest	91.24	1.44 sn	91.54	6.34 sn	**93.36**	1.84 sn
REPTree	90.63	1.56 sn	91.54	1.81 sn	91.54	0.7 sn
DecısıonTable	89.43	1.46 sn	91.54	21.57 sn	90.63	1.81 sn
JRip	91.24	1.57 sn	93.05	6.34 sn	92.45	0.84 sn
İBk	90.63	1.52 sn	88.52	0.51 sn	90.94	0.53 sn
SMO	91.24	1.34 sn	92.75	0.7 sn	93.35	0.6 sn
BayesNet	85.50	1.37 sn	87.92	1.14 sn	87.92	0.56 sn
NaıveBayes	90.03	1.41 sn	90.94	0.8 sn	90.63	0.5 sn
NBM	90.33	1.22 sn	90.94	0.64 sn	93.35	0.61 sn

Tablo 16. *Tarkan-S veri setinin özellik seçim metotları ile sınıflandırma sonuçları*

Sınıflandırma algoritması	CfsSubsetEval + GreedyStepwise (varsayılan)	Model oluşum zamanı	GainRAE + Ranker (varsayılan)	Model oluşum zamanı	GainRAE + Ranker (seçilen özellik sayısı: 100)	Model oluşum zamanı
J48	93.05	0.88 sn	94.56	1.52 sn	94.86	0.44 sn
RandomForest	93.96	0.94 sn	94.26	4.14 sn	95.17	1.73 sn
REPTree	93.35	0.7 sn	94.56	1.23 sn	94.56	0.48 sn
DecısıonTable	93.05	1.04 sn	93.05	10.14 sn	93.05	1.37 sn
JRip	92.75	0.77 sn	94.26	2.37 sn	94.26	0.54 sn
İBk	93.96	0.66 sn	93.05	0.37 sn	93.05	0.4 sn
SMO	94.26	0.83 sn	94.86	0.5 sn	94.26	0.41 sn
BayesNet	90.33	0.8 sn	93.35	0.81 sn	93.35	0.38 sn
NaıveBayes	93.66	0.69 sn	**95.77**	0.59 sn	95.47	0.39 sn
NBM	93.96	0.82 sn	92.45	0.41 sn	94.86	0.35 sn

Tablo 7-8'deki Ece ve Ece-S veri setleri üzerinde SMO + (GainRAE + Ranker) kombinasyonu ve RandomForest+ (GainRAE + Ranker) kombinasyonları ile en yüksek sonuçlar elde edilmiştir. Öznitelik seçimi ile elde edilen sonuçların doğruluk oranının öznitelik seçimi uygulanmadan elde edilen sonuçlara göre nispeten yüksek olduğu gözlenmektedir.

Tablo 9'daki Gülşen veri seti üzerinde SMO + (GainRAE + Ranker) kombinasyonu ile en yüksek sonuçlar elde edilmiştir. Diğer yandan, Tablo 10'daki Gülşen-S veri seti üzerinde SMO + (GainRAE + Ranker) kombinasyonunun yanısıra birden fazla kombinasyon için en yüksek başarı oranı olan %96.19 doğruluğun elde edildiği görülmektedir. Gülşen veri seti için öznitelik seçimi başarımı nispeten artırmış ancak Gülşen-S veri setinde iki durum için de aynı en yüksek doğru sınıflandırma oranı elde edilmiştir.

Tablo 11-12'deki Hadise ve Hadise-S veri setleri üzerinde SMO + (GainRAE + Ranker) kombinasyonu ile en yüksek doğruluk oranları elde edilmiştir. Hadise veri seti için öznitelik seçimi başarımı nispeten artırmış ancak Hadise-S veri setinde iki durumda da aynı en yüksek başarım oranları elde edilmiştir.

Tablo 13-14'teki Hande ve Hande-S veri setleri üzerinde SMO + (GainRAE + Ranker) kombinasyonu ve RandomForest + (GainRAE + Ranker) kombinasyonları ile en yüksek doğruluk oranları elde edilmiştir. Öznitelik seçimi ile elde edilen sonuçların doğruluk oranının öznitelik seçimi uygulanmadan elde edilen sonuçlara göre nispeten yüksek olduğu gözlenmektedir.

Tablo 15'deki Tarkan veri seti üzerinde RandomForest + (GainRAE + Ranker) kombinasyonu ile en yüksek doğruluk oranı elde edilmiştir. Diğer yandan, Tablo 16'daki Tarkan-S veri seti üzerinde NaiveBayes + (GainRAE + Ranker) kombinasyonu ile en yüksek doğruluk oranının elde edildiği görülmektedir.

Tüm veri setleri için genel olarak (GainRAE + Ranker) yöntemi ile 100 öznitelik seçildiğinde sınıflandırma algoritmalarının daha iyi doğruluk oranları verdiği değerlendirilmektedir. (CfsSubsetEval + GreedyStepwise) yöntemi ile ise bu oranda başarılı sonuçlar elde edilememiştir.

V. SONUÇ

Bu çalışmada, Türkçe YouTube yorumları üzerinde otomatik spam tespiti üzerine kapsamlı deneyler yapılmış ve başarılı olan sınıflandırma şemaları tespit edilmiştir. Amaç doğrultusunda 2017 senesinde en çok izlenen 5 Türkçe video klipe yapılan yorumlar online bir araçla çekilerek veri seti oluşturulmuş, verinin ön işlemden geçirilmesi için Java programlama dilinden ve Türkçe verilerin işlenmesine olanak sağlayan Zemberek kütüphanesinden yararlanılmıştır. Sınıflandırma aşaması için ise açık kaynak kodlu Weka veri madenciliği aracı kullanılmıştır. Deneysel sonuçlar değerlendirildiğinde genel olarak tüm algoritmaların ön işlem uygulanmış veri seti üzerinde sınıflandırma performansının ön işlem uygulanmayan veri setine göre çok daha iyi olduğu görünmektedir. Tüm veri setleri üzerinde Rastgele Orman (RandomForest) ve SMO algoritmalarının daha yüksek sınıflandırma doğruluğu verdiği görülmektedir. Rastgele Orman (RandomForest) algoritması hız açısından biraz düşük performans gösterse de, SMO algoritması hem hız hem de doğruluk yüzdesi açısından başarılı olmuştur. Bu algoritmalar ön işlem uygulanmayan veri setleri üzerinde %88.86 - %91.94 ve %91.55 - %92.75 arasında doğruluk gösterirken, ön işlemden geçirilmiş veri setleri üzerinde %93.56 - %95.64 ve %94.06 - %96.19 arasında doğruluk göstermişlerdir. Hemen ardından JRip ve Basit Bayes (NaiveBayes) algoritmaları da genellikle yüksek başarım gösteren algoritmalardandır.

Deneylerin devamında aynı veri setlerinin öznitelik seçimi metotları uygulanarak sınıflandırılan sonuçları sunulmuştur. Öznitelik seçiminin başarım oranlarını nispeten artırdığı görülmektedir. Öznitelik seçim metotları uygulandıktan sonra yapılan sınıflandırmada da yine Rastgele Orman

(RandomForest) ve SMO algoritmaları diğerlerine göre daha başarılı olmuştur. Algoritmaların başarı oranları ön işlem uygulanmayan veri setleri üzerinde %91.02 -%93.36 ve %91.92-%93.35 arasında iken, ön işlemden geçirilmiş veri setleri üzerinde %94.01-%96.70 ve %93.11-%96.34 arasında olmuştur. Diğer yandan, (CfsSubsetEval + GreedyStepwise) yöntemi ile başarılı sonuçlar elde edilememiştir.

Bu çalışma ile ilişkili ileriki çalışmalarda aynı yöntemler kullanılarak YouTube'dan farklı sosyal medya ortamlarında spam tespitinin performansının analiz edilmesi düşünülebilir.

VI. KAYNAKÇA

[1] R. Dolan, J. Conduit R., J. Fahy, S. Goodman, "Social media: communication strategies, engagement and future research directions", *International Journal of Wine Business Research*, vol. 29, no. 1, pp. 1-19, 2017.

[2] H. Bayrak. (2020, 23 Şubat). Türkiye İnternet Kullanımı ve Sosyal Medya İstatistikleri. [Online]. Erişim: https://dijilopedi.com/2020-turkiye-internet-kullanimi-ve-sosyal-medya-istatistikleri/.

[3] P. de Bérail, M. Guillon, C. Bungener, "The relations between YouTube addiction, social anxiety and parasocial relationships with YouTubers: A moderated-mediation model based on a cognitive-behavioral framework", *Computers in Human Behavior*, vol. 99, pp. 190-204, 2019.

[4] T. Singh, M. Kumari, S. Mahajan, "Feature oriented fuzzy logic based web spam detection", *Journal of Information and Optimization Sciences*, vol. 38, no. 6, pp. 999-1015, 2017.

[5] C. Romero, M.G. Valdez, A. Alanis, "A comparative study of machine learning techniques in blog comments spam filtering", *The 2010 International Joint Conference on Neural Networks (IJCNN)*, 2010, pp. 1-7.

[6] W. Yang, L. Kwok, "Improving blog spam filters via machine learning", *International Journal of Data Analysis Techniques and Strategies*, vol. 9, no. 2, pp. 99-121, 2017.

[7] Z. Li, H. Shen, "Soap: A social network aided personalized and effective spam filter to clean your e-mail box", *IEEE INFOCOM*, 2011, pp. 1835-1843.

[8] G. Sanghani, K. Kotecha, "Incremental personalized E-mail spam filter using novel TFDCR feature selection with dynamic feature update", *Expert Systems with Applications*, vol. 115, pp. 287-299, 2019.

[9] B. K. Dedeturk, B. Akay, "Spam filtering using a logistic regression model trained by an artificial bee colony algorithm", *Applied Soft Computing*, vol. 91, pp. 1-18, 2020.

[10] C. Wang, Q. Li, T. Y. Ren, X. H. Wang, G. X. Guo, "High Efficiency Spam Filtering: A Manifold Learning-Based Approach", *Mathematical Problems in Engineering*, vol. 2021, pp. 1-7, 2021.

[11] J.M.G. Hidalgo, T.A. Almeida, A. Yamakami, "On the validity of a new SMS spam collection", *11th International Conference on Machine Learning and Applications*, 2012, pp. 240-245.

[12] Ö. Örnek, "Orange 3 ile Türkçe ve İngilizce SMS Mesajlarında Spam Tespiti", *Eskişehir Türk Dünyası Uygulama ve Araştırma Merkezi Bilişim Dergisi*, c. 1, s. 1., ss. 1-4, 2019.

[13] P. K. Roy, J. P. Singh, S. Banerjee, "Deep learning to filter SMS Spam", *Future Generation Computer Systems*, vol. 102, pp. 524-533, 2020.

[14] V. Chaudhary, A. Sureka, "Contextual feature based one-class classifier approach for detecting video response spam on youtube", *Eleventh Annual Conference on Privacy, Security and Trust*, 2013, pp. 195-204.

[15] T. Abd, H. Altabrawee, S.Q. Ajmi, "YouTube Spam Comments Detection Using Artificial Neural Network", *Journal of Engineering and Applied Sciences*, vol. 13, no. 22, pp. 9638-9642, 2018.

[16] T. C. Alberto, J.V. Lochter, T.A. Almeida, "Tubespam: Comment spam filtering on YouTube". *Machine Learning and Applications (ICMLA)*, 2015, pp. 138-143.

[17] N.A.M. Samsudin, C. F. B. M. Foozy, N. Alias, P. Shamala, N. F. Othman, W. I. S. W. Din, "Youtube spam detection framework using naïve bayes and logistic regression", *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 14, no. 3, pp. 1508-1517, 2019.

[18] A.K. Uysal, "Feature selection for comment spam filtering on YouTube", *Data Science and Applications*, vol. 1, no. 1, pp. 4-8, 2018.

[19] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, "The WEKA data mining software: an update", *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10-18, 2009.

[20] B. Oralhan, "Veri madenciliği yaklaşımı ile telekomunikasyon sektöründe arıza giderme analizi", *Business & Management Studies: An International Journal* , vol. 8, no. 1, pp. 1026-1043, 2020.

[21] M. Amrehn, F. Mualla, E. Angelopoulou, S. Steidl, A. Maier. (2018, 19 Aralık). The random forest classifier in WEKA: Discussion and new developments for imbalanced data. [Online]. Erişim: https://arxiv.org/abs/1812.08102v1.

[22] S. Kalmegh, "Analysis of weka data mining algorithm reptree, simple cart and randomtree for classification of indian news", *International Journal of Innovative Science, Engineering & Technology*, vol. 2, no. 2, pp. 438-446, 2015.

[23] S. Kalmegh, M.S. Ghogare, "Performance comparison of rule based classifier: Jrip and decisiontable using weka data mining tool on car reviews", *International Engineering Journal For Research & Development*, vol. 4, no. 5, pp. 5-5, 2019.

[24] J. Brownlee, "Machine learning mastery with Weka", *E-book*, vol.1, 2019.

[25] A. Nakra, M. Duhan, "Comparative Analysis of Bayes Net Classifier, Naive Bayes Classifier and Combination of both Classifiers using WEKA", *IJ Inf. Technol. Comput. Sci*, vol. 3, pp. 38-45, 2019.

[26] A. Gümüşçü, İ.B. Aydilek, R. Taşaltın, "Mikro-dizilim Veri Sınıflandırmasında Öznitelik Seçme Algoritmalarının Karşılaştırılması", *Harran Üniversitesi Mühendislik Dergisi*, c. 1, s. 1, ss. 1-7, 2016.

[27] M. Z. Alam, M.S. Rahman, M.S. Rahman, "A Random Forest based predictor for medical data classification using feature ranking", *Informatics in Medicine Unlocked*, vol. 15, pp. 1-12, 2019.

[28] H. Schütze, C. D. Manning, P. Raghavan, *Introduction to information retrieval*, Cambridge: Cambridge University Press, c. 39, ss. 234-265, 2018.

[29] A. A. Akın, M. D. Akın, “Zemberek, an open source NLP framework for Turkic Languages”, *Structure*, 2007, pp. 1-5.