Türkçe Yaramaz E-postaların Farklı Öznitelik Seçim Yöntemleri Kullanılarak Makine Öğrenmesi Algoritmaları ile Tespit Edilmesi

Detection of Turkish Spam Emails with Machine Learning Algorithms Using Different Feature Selection Methods

Ersin Enes ERYILMAZ
Ondokuz Mayıs Üniversitesi, Bilgisayar Mühendisliği Bölümü, Samsun, Türkiye
enes.eryilmaz@bil.omu.edu.tr
ORCID: 0000-0003-1163-970X

Durmuş Özkan ŞAHİN
Ondokuz Mayıs Üniversitesi, Bilgisayar Mühendisliği Bölümü, Samsun, Türkiye
durmus.sahin@bil.omu.edu.tr
ORCID: 0000-0002-0831-7825

Erdal KILIÇ
Ondokuz Mayıs Üniversitesi, Bilgisayar Mühendisliği Bölümü, Samsun, Türkiye
erdal.kilic@bil.omu.edu.tr
ORCID: 0000-0003-1585-0991

Öz

Elektronik postalar, kullanımının kolaylığı, maliyetlerinin ucuz olmasından dolayı propaganda, reklam, oltalama yapmak isteyen kişi veya topluluklar tarafından etkin bir biçimde kullanılmaktadır. Amaçlarını gerçekleştirmek isteyen kişi veya topluluklar hiç tanımadıkları e-posta hesaplarına gereksiz ve yaramaz postalar gönderirler. Bu postalar internet kullanıcılarına maddi ve manevi ciddi zararlar vermekte ayrıca internet trafiğini de meşgul etmektedirler. Yaramaz e-postalar alıcıya rızası dışında gönderilen ve genellikle kötü niyetli veya tanıtım amaçlı olan kişilerin başvurduğu bir yöntemdir. Bu çalışmada iki farklı Türkçe e-posta veri kümesi üzerinde yedi farklı makine öğrenmesi algoritması kullanılarak yaramaz e-postalar tespit edilmeye çalışılmıştır. Bu algoritmaları kullanmadan önce veri kümesi üzerinde ön işlem adımları gerçekleştirilmiştir. Daha sonrasında ise öznitelik çıkarımı ve öznitelik seçimi yapılmıştır. Öznitelik seçimleri sonrasında özellik vektörü oluşturarak makinenin anlayacağı formatta değerler elde edilmiştir. Özellik vektörü makine öğrenmesi algoritmaları ile test edilerek yaramaz e-posta filtreleme işlemiyle elde edilen başarım sonuçları değerlendirilmiştir. Metin sınıflandırma çalışmalarında sıkça kullanılan filtreleme tabanlı Ki-kare (CHI), Bilgi Kazancı (IG), Doküman Frekansı Eşikleme (DF), Odds Oranı (OR) ve ACC öznitelik seçme yöntemleri kullanılmaktadır. İki Türkçe e-posta veri kümesi ile CHI, IG, ACC, OR, DF öznitelik seçme yöntemlerinin çeşitli makine öğrenmesi sınıflandırma algoritmaları üzerinde verdiği sonuçlar incelendiğinde en başarılı sonuç Ki-Kare öznitelik seçimi ile görülmüştür. "TurkishEmail" veri kümesi ile Destek Vektör Makinesi tabanlı SMO algoritması ve CHI öznitelik seçimi ile 0,985 F-ölçütü başarım sonucu elde edilmiştir. "TRHamSpamEmailv1.0" veri kümesi ile CHI öznitelik seçim yöntemi Rastgele Orman (RF) ve Naive Bayes (NB) algoritması ile 0,748 F-ölçütü başarıma ulaşmıştır. Herhangi bir öznitelik seçimi yapılmadan tüm özniteliklerin kullanılması ile elde edilen sınıflandırma başarıları da verilmiştir. Öznitelik seçimi yapılmadan "TurkishEmail" veri kümesi üzerinde RF algoritması ile başarım sonucu 0,514 F-ölçütü, "TRHamSpamEmailv1.0" veri kümesi üzerinde RF algoritması ile başarım sonucu 0,535 F-ölçütü olarak elde edilmiştir.

Anahtar Sözcükler: *e-posta sınıflandırma, öznitelik çıkarımı, özellik seçimi, yaramaz e-posta, spam filtreleme, makine öğrenmesi, Türkçe e-posta sınıflandırma, Türkçe spam filtreleme, metin sınıflandırma.*

Abstract

Electronic mails are used effectively by people or communities who want to make propaganda,

Gönderme ve kabul tarihi: 14.10.2020 - 03.12.2020
Makale türü: Araştırma

advertising, phishing because of its ease of use and low cost. People or communities who want to achieve their goals send junk and spam emails to e-mail accounts they do not know. These mails cause serious material and moral damages to internet users and also engage internet traffic. Spam e-mails are a method that is sent to the recipient without their consent and are often used by malicious or promotional people. In this study, it was tried to detect spam e-mails by using seven different machine learning algorithms on two different Turkish e-mail datasets. Before using these algorithms, pre-processing steps were performed on the datasets. Afterward, feature extraction and feature selection were made. After the feature selections, the values were obtained in a format that the machine can understand by creating the feature vector. The performance results of the spam filtering process were evaluated by testing the feature vector with machine learning algorithms. Which are frequently used in text classification studies, filtering-based Chi-square (CHI), Information Gain (IG), Document Frequency Threshold (DF), Odds Ratio (OR), and ACC feature selection methods are used. When examining the results of two Turkish e-mail datasets and CHI, IG, ACC, OR, DF feature selection methods on different machine learning classification algorithms, the most successful result was seen with Chi-Square feature selection. With the "TurkishEmail" dataset, the SMO algorithm based on Support Vector Machine, and CHI feature selection, 0,985 F-measure performance result was obtained. With the "TRHamSpamEmailv1.0" dataset, the CHI feature selection method achieved a 0,748 F-measure with Random Forest (RF) and Naive Bayes (NB) algorithm. Classification successes obtained by using all features without any feature selection are also given. The performance result was obtained as a 0,514 F measure with the RF algorithm on the "TurkishEmail" dataset without the feature selection and as a 0,535 F-measure on the "TRHamSpamEmailv1.0" dataset with the RF algorithm.

Keywords: *e-mail classification, feature extraction, feature selection, spam email, spam filtering, machine learning, Turkish e-mail classification, Turkish spam filtering, text classification.*

1. Giriş

Basit, anlaşılır, maliyeti az ve herkesin kolayca kullanabilmesinden dolayı pek çok şirketin reklamını yapmak için popüler hale gelen e-postalar günümüzün en çok kullanılan elektronik iletişim araçlarındandır. İnternet erişiminin her yere yayılması, herkes tarafından internetin ulaşılabilir olması insanların bir şekilde e-posta ile tanışmasını sağlamaktadır. E-posta ile kişi veya kurumların doğruluğu kontrol edilebilmektedir. E-posta sayesinde insanlar istedikleri zaman dünyanın farklı yerlerinde bulunan bir alıcıyla kolayca iletişime geçebilmektedir. Ucuzluğu ve kolaylığı nedeniyle tercih edilen e-posta ile her gün dünyada milyarlarca e-posta gönderilip alınmaktadır [1].

Günümüzde Yahoo, Outlook, Gmail, Yandex vb. gibi oldukça kullanışlı e-posta arayüzleri geliştirilmiş olup ücretsiz olan e-posta adresleri milyonlarca insan tarafından kullanılmaktadır. E-posta hesapları, bu hizmeti veren çeşitli sitelerden ücretsiz veya belirli bir ücret karşılığında açılabilir. E-posta, çok fazla zaman ve para tasarrufu sağladığı için etkili bir iletişim yoludur, bu da e-postayı kişisel ve profesyonel iletişimde favori iletişim aracı haline getirir. E-postalar internet kullanıcılarının küresel olarak kolayca bilgi aktarmalarını sağlar. İlk önemli e-posta standardı basit mesaj aktarma protokolü (SMTP: Simple Mail Transfer Protocol) olarak adlandırılmıştır. SMTP mesaj göndermeyi iddia eden kişinin iddia ettiği kişi olup olmadığını anlamaya çalışmaz. Bu nedenle e-posta iletişiminde sahtecilik yaygınlaşmıştır. Protokoldeki bu basitlik virüs, solucan, güvenlik sahtekârları ve yaramaz e-posta yayınlayıcılar tarafından kullanılmaktadır.

Yaramaz e-posta, talep etmeyen çok sayıda alıcıya bir reklam veya alakasız içeriği olan bir mesajın gönderilmesi anlamına gelir. Kişi veya topluluklar herhangi bir amaç için tanımadıkları kişi veya kurumlara gereksiz, önemsiz, yaramaz, yaramaz e-posta denilen e-postalar gönderebilir. Bu e-postalar internet trafiğini meşgul etmekle birlikte zaman zaman e-posta muhatabı olana ciddi zararlar verebilir.

Bilgisayarların karmaşık problemleri çözmek için insan beyninin davranışlarından ilham alarak taklit etme, öğrenme, kavrama, yorumlama gibi yeteneklerin makinelere kazandırılması çalışılmaktadır. Yaramaz e-postaları tespit etmek için farklı teknik ve yöntemler kullanılmakla birlikte

bu çalışmalar genellikle İngilizce veri kümeleri üzerindedir. Türkçe yaramaz e-postalar üzerinde çalışmalar yeterli düzeyde değildir. Bunun genel sebebi veri kümelerinin azlığı veya yetersizliğidir.

Yaramaz e-postaları ayıklamak için birçok yöntem bulunmaktadır. Bunlar genel olarak iki grupta toplanabilir. Bunlardan birincisi yapay zekâya dayanmayan sistemlerdir. Bir diğeri ise bu çalışmanın içeriğinde kullanılan makine öğrenmesi tekniklerini de içinde bulunduran yapay zekâ tabanlı sistemlerdir [2].

Makine öğrenmesinin en yaygın şekli denetimli öğrenmedir. Bir e-postanın yaramaz e-posta veya yaramaz e-posta olup olmadığını sınıflandırabilmek için önce, her biri kategorisiyle etiketlenmiş yaramaz e-posta ve yaramaz e-posta olmayan geniş bir veri kümesinin toplanması gerekir. Eğitim sırasında, makineye bir e-posta gösterilir ve her kategori için bir tane olmak üzere bir skor vektörü şeklinde bir çıktı üretilir. İstenilen kategorinin tüm kategorilerde en yüksek puana sahip olması istenir, ancak bunun eğitimden önce gerçekleşmesi olası değildir. Burada çıktı puanları ile istenen puan deseni arasındaki hatayı (veya mesafeyi) ölçen nesnel bir fonksiyon kullanılmaktadır. Makine daha sonra bu hatayı azaltmak için dâhili ayarlanabilir parametrelerini değiştirir. Genellikle ağırlık olarak adlandırılan bu ayarlanabilir parametreler, makinenin giriş-çıkış fonksiyonunu tanımlayan gerçek ayar düğmeleri gibidir. Ağırlık vektörünü uygun şekilde ayarlamak için öğrenme algoritması, ağırlık küçük bir miktar arttırılınca hatanın ne kadar artacağını veya azalacağını gösteren bir gradyan vektörü hesaplar. Ağırlık vektörü daha sonra gradyan vektörüne zıt yönde ayarlanır [3].

Bu makale çalışmasında yapay zekâya dayalı sistemlerden olan metin madenciliği ve makine öğrenmesi yöntemleri ile Türkçe yaramaz e-posta algılama sistemi geliştirilmiştir. Bu sistem için farklı özellik seçim yöntemleri ve algoritmalar kullanılmıştır. Literatürde makine öğrenmesi teknikleri kullanılarak yaramaz e-posta tespiti yapan bazı çalışmalar aşağıda özetlenmiştir.

Ateş çalışmasında hem Türkçe hem de İngilizce veri kümesi kullanmaktadır [4]. Ergin ve ark. [21] tarafından oluşturulan Türkçe veri kümesinde; 800 e-posta kullanılırken, Karşılıklı Bilgi algoritması ile en yüksek değere sahip 49 adet terim öznitelik olarak seçilmektedir. Naïve Bayes (NB) ile %99, Destek Vektör Makinesi (DVM) ile %95, Gauss Karışım Modeli ile %93 doğrulukla sınıflandırma başarımı elde edilmektedir. Normal e-posta tespitinde doğrusal DVM ile %99 başarım sonucu raporlanmaktadır.

Sharma ve ark. [5], TREC07 veri kümesi üzerinde Çok Katmanlı Algılayıcı (MLP: Multi Layer Perceptron) ve NB algoritmalarını kullanmaktadır. En başarılı sonuç MLP algoritmasıyla elde edilmektedir. Elde edilen sonuçlar incelendiğinde, doğruluk ve tutturma metriğine göre %93, bulma metriğine göre ise %93,2 sınıflandırma başarımları tespit edilmektedir.

Karthika ve Visalakshi [6], Spambase veri kümesi üzerinde k-En Yakın Komşu (KNN: K-Nearest Neighbors), NB, DVM ve Hibrit ACO-DVM algoritmalarını kullanmaktadır. En başarılı sonuç ACO-DVM melez algoritmasıyla elde edilirken, bu sonuç doğruluk metriğine göre %81,25, tutturma metriğine göre %87,02 ve bulma metriğine göre %75,1'dir.

Renuka ve ark. [7], Spambase veri kümesi üzerinde GA-Naive Bayes, ACO-Naive Bayes gibi melez algoritmaları kullanmaktadır. En başarılı sonuç ACO-Naive Bayes melez algoritmasıyla elde edilmektedir. Bu sonuç doğruluk, tutturma, bulma ve F-ölçütü metriklerine göre sırasıyla %84, %89, %78 ve %87'dir.

Palanisamy ve ark. [8], Lingspam veri kümesi üzerinde Negatif Seçim Algoritması (NSA) tabanlı Parçacık Sürü Optimizasyonu (PSO: Particle Swarm Optimization), DVM, NB ve DFS-DVM gibi melez algoritmaları kullanmaktadır. En başarılı sonuç Negatif Seçim Algoritması (NSA) tabanlı melez PSO algoritması ile elde edilmektedir. Bu sonuç doğruluk metriğine göre %93,2'dir.

Zavvar ve ark. [9] Spambase veri kümesini kullanarak PSO, Öz Düzenleyici Haritalar (SOM: Self Organizing Map), k-ortalama ve DVM algoritmaları ile sınıflandırma yapmaktadır. En başarılı sonuç DVM algoritması ile AUC ölçütüne göre %93,07 olarak raporlanmaktadır.

Foqaha [10], Spambase veri kümesini Yarıçapsal Temelli Ağ (RBF: Radial Basis Function), MLP ve melez HC-RBFPSO algoritmaları ile deney etmektedir. Çalışmada elde edilen en başarılı sonucu MLP algoritması vermektedir. Bu sonuç, doğruluk metriğine göre %93,28 olarak bulunmaktadır.

Sharma ve Suryawanshi [11], Spambase veri kümesi üzerinde Bayes, KNN ve DVM algoritmalarını kullanmaktadır. En başarılı sonuç KNN algoritması ile tespit edilmektedir. Bu sonuç doğruluk metriğine göre %97,54, kesinlik metriğine göre %97,72, bulma metriğine göre %93,52 ve F-ölçütü metriğine göre %95,6 olarak raporlanmaktadır.

Alkaht ve Al Khatib [12], çalışmalarında üç farklı veri kümesi kullanmaktadır. Bunlar; İngilizce veri kümeleri CSmining CSDMC 2010 ve SpamAssassin ile Arapça ve İngilizce karışık e-posta içeren Tarassul veri kümeleridir. İleri Beslemeli Ağ ve SOM algoritmaları, Çeşitli Aşamalı Sinir Ağı (SNN) adı verilen yöntemle kullanılmıştır. Farklı aktivasyon fonksiyonu ve özellik boyutu kullanılarak birçok tutturma, bulma ve F-ölçütü sonucu elde edilmiştir. Sonuçlar, ileri beslemenin farklı konu ve alana sahip e-postaları sınıflandırmak için uygun olduğunu göstermiştir. Öz düzenleyici haritaların birçok alan içeren e-postaları sınıflandırmak için uygun olduğunu göstermiştir. Çalışma ile, İngilizce ve Arapça dillerinin bir arada kullanıldığı e-postalarla ilgili sınıflandırıcı özelliğin düşük olduğu belirtilmiştir.

Rajamohana ve ark. [13], Ott ve ark. [14] tarafından oluşturulan veri kümesi üzerinde Uyarlanabilir İkili Çiçek Tozlaşma algoritmasını (ABFPA: Adaptive Binary Flower Pollination Algorithm) kullanarak yaramaz e-postaların sınıflandırılmasını amaçlamaktadır. Çalışmada elde edilen sınıflandırma başarımı doğruluk metriğine göre %91,42 olarak tespit edilmektedir.

Akinyelu ve Adevumi [15], 2000 adet kimlik avı ve normal e-postalardan oluşan veri kümesi üzerinde Rastgele Orman (RF: Random Forest) algoritmasıyla sınıflandırma yapmaktadır. Çalışmada elde edilen sonuçlar doğruluk metriğine göre %99,7, tutturma metriğine göre %99,47, bulma metriğine göre %97,5 ve F-ölçütü metriğine göre %98,45'dir.

Yıldız [16], kurumsal verileri dış ağlara paylaşmadan yerel ağda Türkçe gerçek verilerle sınıflandırma yapabilen bir masaüstü uygulaması önermektedir. 310 adet Türkçe e-posta verisi Zemberek ile köklerine ayrılmaktadır. En başarılı sonuç Çok Terimli Naïve Bayes algoritması ile elde edilmektedir. Bu sonuç kappa ölçütüne göre %94 olurken, doğruluk ölçütüne göre %96,31 olarak tespit edilmektedir. Ayrıca tutturma ve bulma metrikleri de kullanılmaktadır. Bu metriklere göre sırasıyla %91 ve %100 sınıflandırma başarımları elde edilmektedir. Yerel ağda Türkçe gerçek verilerle sınıflandırma yapabilen bir masaüstü uygulama önerilmesi literatüre katkı olarak değerlendirilebilir ama az sayıda e-posta kullanılması çalışmanın eksik yanı olarak göze çarpmaktadır.

Şahin [17], e-posta içeriğinde yer alan bağlantı linklerinin metinlerini kullanmaktadır. Kelime Çantası Tekniği (BOW) ile yaramaz e-postaların sınıflandırılması yapılmaktadır. Farklı n-gram modellerinin sınıflandırma başarımına etkisi incelenmektedir. 5-gramlı modelin %95 başarımla sınıflandırma performansına etkisinin en fazla olduğu belirtilmektedir. Çalışmada on iki klasik makine öğrenmesi algoritması kullanılmaktadır. Bunlar farklı Naïve Bayes, DVM, çok katmanlı yapay sinir ağı, k-en yakın komşu ve karar ağaçları algoritmalarıdır. Karar ağaçları haricindeki bütün modeller %98'in üzerinde sınıflandırma başarımı verirken karar ağaçlarına dayalı algoritmalarının yaramaz e-posta sınıflandırmada düşük başarı gösterdiği vurgulanmaktadır. Karar ağaçlarından elde edilen başarım ise %65'dir. Çekirdek Naive Bayes ve Doğrusal DVM algoritmaları ile doğruluk metriğine göre %99,89, F-ölçütü metriğine göre %99,81 sınıflandırma başarımı tespit edilmektedir.

Kale [18], 2013 yılında oluşturulmuş Louis Dorard'a ait 4709 adet e-posta verisi ile Karar Ağaçları, Derin öğrenme, Gradient Boosted Tree (GBT), KNN, NB, RF ve Lojistik Regresyon (LR) algoritmalarını kullanmaktadır. Çok Terimli NB ile en başarılı sonuç elde edilmektedir. Çalışmasının performans ölçütlerinde %95,5 doğruluk, %100 tutturma, %91 bulma ve %95,8 F-ölçütü başarım sonuçları görülmektedir.

Nazlı [19], yaramaz ve normal e-postaların olduğu Enron veri kümesini kullanarak yaramaz e-postaları sınıflandırmayı hedeflemektedir. Çalışmada, WEKA [44] veri madenciliği aracında yer alan DVM modelleri, Naïve Bayes modelleri ve karar ağaçlarından olan C4.5 ve RF algoritması kullanılmaktadır. Word2Vec modeli kullanılarak öznitelik vektörü oluşturulmaktadır. Oluşturulan vektör, makine öğrenmesi algoritmalarına verildiğinde en yüksek başarım DVM Polinom çekirdek algoritması ile elde edilmektedir. Bu başarım 300 e-posta için %98,33 olmaktadır. Farklı makine öğrenmesi teknikleri ve küçük veri kümeleri üzerinde yüksek doğrulukla sınıflandırma

yapılırken, veri kümesi arttıkça F-ölçütü değerinin %50'ye kadar düştüğü görülmektedir.

Al-Azzawi [20], Spambase veri kümesi üzerinde Kaotik ateş böceği algoritmasına dayanan sarmal öznitelik seçimli NB melez algoritmasını kullanmaktadır. Çalışmada doğruluk metriğine göre %95,14 sınıflandırma başarımı elde edilmektedir.

Ablel-Rheem ve ark. [21], Spambase veri kümesi üzerinde NB, Karar Ağaçları ve Kolektif (Ensemble) öğrenme algoritmaları 10 katlı çapraz doğrulama ve Bilgi Kazancı (IG) öznitelik seçimi kullanılarak karmaşıklık matrisi hesaplanmıştır. Melez Kolektif yöntemlerle %94,4 tutturma, bulma ve F-ölçütü başarımına ulaşmıştır.

Zamir ve ark. [22], veri kümesinin mevcut öznitelikleri yanında içerik, duygu, anlambilim, kullanıcı ve yaramaz e-posta sözlüğü gibi çeşitli özellikleri kullanmıştır. Elde edilen özelliklere bilgi kazancı, kazanç oranı ve Relief-F gibi özellik seçme teknikleri uygulanarak derin sinir ağına verilmiştir. Çalışmada %97,2 oranında sınıflandırma başarımı elde edilmiştir.

Mohammad [23], ömür boyu süreceği iddia edilen bir yaramaz e-posta sınıflandırma modeli olan, "Ayarlanabilir Veri Kümesi Bölümleme Kullanan Topluluk Tabanlı Yaşam Boyu Sınıflandırma" modeli önermiştir. "ham" olarak işaretlenen 16.545 e-posta iletisi ve "yaramaz posta" olarak işaretlenen 17.171 e-posta iletisi olmak üzere toplam 33.716 e-posta iletisinden oluşan "Enron-Spam" veri kümesi kullanılmıştır. Doğruluk, tutturma, bulma ve F-ölçütü metriklerine göre sırasıyla % 95,80, %94,40, %95,80 ve % 95,10 oranında sınıflandırma başarımı tespit edilmiştir. Kolektif yaramaz e-posta sınıflandırma yaklaşımının yaramaz e-postaları tanımlamak için daha uygun olduğu gözlemlenmektedir.

Kumar ve Sonowal [24], Kaggle web sitesinden "spam.csv" isimli veri kümesini 7 farklı makine öğrenmesi algoritması üzerinde deney etmiştir. Python programlama dili ile oluşturulan "sklearn" kütüphanesi sınıflandırma algoritmaları kullanılarak modeller eğitilmiştir. En yüksek başarım çok terimli NB algoritmasıyla %98 doğrulukla tespit edilmiştir. Metin ön işleme adımında, durak sözcükler ve noktalama işaretleri kaldırılarak temiz sözcükler elde edilmiştir. Kullanılan veri kümesinin %70'i eğitim, %30'u test kümesi seçilerek sözcük dağarcığı ile hiperparametre ayarları yapılmıştır.

Deniz ve ark. [25] çalışmasında, TurkishEmail e-posta veri kümesinden ile özellik çıkarımı yapılmış ve Doc2Vec kütüphanesine ait algoritmalar kullanılmıştır. Doc2Vec iki farklı algoritma içermektedir. Bu algoritmalar sırası ile Distributed Bag of Word (DBoW) ve Distributed Memory (DM) olarak adlandırılır. Altı farklı makine öğrenmesi algoritması ile sınıflandırma yapılmıştır. En başarılı sonuç DBoW+DM modeli özellik seçimi ve destek vektör makinesi algoritması ile %78,75 olarak elde edilmiştir.

Karamollaoğlu ve ark. [26], TurkishEmail veri kümesini üzerinde Naive Bayes algoritmasıyla %95,5 başarım, Vektör Uzay Modeliyle %93,5 başarım tespit etmişlerdir.

Kaynar ve ark. [27] çalışmasında, TurkishEmail veri kümesi üzerinde derin öğrenme yöntemlerinden oto kodlayıcılar ile iki farklı deney gerçekleştirmiştir. İlk deney ince ayar (fine tuning) öncesi olurken ikinci deney ise ince ayar sonrasıdır. Deneylerden sırasıyla %98 ve %97 doğrulukla sınıflandırma başarımları elde edilmektedir. Derin öğrenme modelinde softmax katmanı içeren bir sınıflandırıcı, sırasıyla 231 ve 115 gizli nörondan oluşan bir ağ yapısı kullanılmıştır. Ağın ezberlemesini önlemek için her iki modelde de L2 ve Sparse düzenleyicisi kullanılmıştır.

Eryılmaz ve ark. [1] tarafından yapılan çalışmada TurkishEmail veri kümesi üzerinde derin öğrenme kütüphanesi Keras kullanılmıştır. Farklı aktivasyon ve eniyileme yöntemleri farklı oranlarda eğitim ve test veri kümeleri üzerinde denenmiştir. LSTM modeli ile birçok yöntem üzerinde %100 sınıflandırma başarım oranına ulaşılmıştır.

Eryılmaz ve ark. [29] tarafından yapılan çalışmada TurkishEmail veri kümesi üzerinde yedi farklı makine algoritması ve iki farklı öznitelik seçme yöntemi karşılaştırılmıştır. Normal e-postalar ile yaramaz e-postalar arasında ayrım gücü en yüksek 250 sözcük seçilmiştir. CHI öznitelik seçme yöntemi ve destek vektör makinesine dayalı sıralı minimum optimizasyon (SMO) algoritmasıyla %98,5, IG öznitelik seçme yöntemi ve MLP algoritmasıyla %98,4 oranında sınıflandırma başarımı elde edilmiştir.

Ergin ve ark. [28] tarafından sunulan bildiride TurkishEmail veri kümesi üzerinde iki farklı Bayes modeli ile yaramaz postaların filtrelenmesi amaçlanmıştır. Bu modeller olasılıklı ve ikili Bayes modelleridir. Olasılıklı Bayes modele göre %89 performans elde edilmiştir. İkili Bayes model sınıflandırmasına göre ise başarım %93'tür.

Yapılan literatür araştırmasına göre genel bir değerlendirme yapılacak olunursa, Türkçe e-posta veri kümesi boyutunun, sayısının ve çalışmaların az olduğu görülmektedir. Kullanılan bazı e-posta veri kümelerinde yaramaz ve meşru elektronik postaların dengeli dağıldığı, bazılarında ise bu dengenin bozulduğu gözlemlenmektedir. Kullanılan veri kümelerinde bazı makine öğrenmesi algoritmaları yüksek başarım oranlarına ulaşırken, bazı algoritmalar ise düşük sınıflandırma başarımı göstermektedir. Ön işlem adımları ve öznitelik seçimlerinin başarım sonuçlarını etkilediği tespit edilmiştir. Yaramaz e-postaların filtrelenmesi çalışmalarında öznitelik seçimlerine göre başarım oranlarının yüksek veya düşük çıktığı görülmüştür. İngilizce Enron, Spambase ve Lingspam e-posta veri kümeleri üzerinde makine öğrenmesi tekniklerinin çoğunlukla kullanıldığı görülmüştür. Çalışmalarda sınıflandırma performanslarını değerlendirmek için genellikle doğruluk, tutturma, bulma ve F-ölçütü metrikleri kullanılmakla birlikte Kappa ve AUC değerleri de görülmüştür.

Yaramaz e-posta tespit çalışmalarında, MLP gibi yapay sinir ağlarına (YSA) dayanan model eğitimlerinin zaman aldığı, Bayes, Naive Bayes, DVM ve karar ağaçlarına dayalı yaklaşımlarda eğitim sürelerinin nispeten daha hızlı gerçekleştiği gözlemlenmektedir. Bu yöntemlerin yanında sınıflandırma başarımlarını arttırmak için melez yaklaşımlar da kullanılmaktadır. Hatta birçok çalışmada melez algoritmalar kullanılarak tasarlanan yaramaz e-posta tespit sistemlerinin sınıflandırma başarımlarının yüksek olduğu görülmektedir.

Nispeten daha küçük veri kümeleri üzerinde çalışan sistemlerde daha iyi sonuçlar verebilen DVM, NB, RF, MLP ve YSA gibi makine öğrenmesi yöntemlerinin kullanılması uygun olmaktadır. Makine öğrenmesinde (öznitelik çıkarımı, özellik seçimi, doğal dil işleme, eğitim vb.) en iyi sonuçları veren yöntemlerin bir arada kullanılması önerilmektedir.

Makine öğrenmesi teknikleri uygulanmadan önce, ön işlem adımlarının, özellik vektörü oluşturmanın ve öznitelik seçiminin dikkatli bir şekilde yapılması sınıflandırma performansını artıracağı gibi geliştirilen yaramaz e-posta filtreleme sisteminin hızlı çalışmasını da sağlayacaktır. Bu etmenler göz önüne alınarak bu çalışmada, metin sınıflandırmada kullanılan önemli teknikler Türkçe e-postaların sınıflandırılmasına uygulanarak sınıflandırma başarımı yüksek bir yaramaz e-posta tespit sistemi önerilmektedir.

Bu çalışma ile RF [37], C4.5 [38], SMO [39], KNN [40], LR [41], NB [42] ve MLP [43] algoritmaları CHI, IG, ACC, OR ve DF öznitelik seçme yöntemleri ile denenmiştir. Çalışma kapsamında iki farklı Türkçe e-posta veri kümesi kullanılmıştır. Ön işlem adımları ve öznitelik seçimlerinin sınıflandırma başarımına etkisi yedi farklı makine öğrenmesi algoritması ile test edilmiştir. Ön işlem adımları ve öznitelik seçimleri yapılmadan sınıflandırma başarımları çok düşük başarım oranlarında kalmıştır. Yedi farklı sınıflandırıcıdan elde edilen sonuçlar karşılaştırıldığında en başarılı yaramaz e-posta filtreleme algoritması SMO olmaktadır. CHI öznitelik seçiminin SMO algoritması ile kullanımında %98,5 oranında sınıflandırma başarımına ulaşılmıştır.

1.1 Motivasyon ve Katkı

Mevcut çözümler çoğunlukla yaramaz e-posta göndericilerin sürekli olarak getirdiği yenilikçiliğin gerisinde kalmakta, bundan dolayı da makine öğrenmesi tabanlı yaramaz e-posta tespit sistemlerine ihtiyaç gün geçtikçe artmaktadır. Makine öğrenmesi yöntemlerinin yaramaz e-posta algılamasındaki başarımı bu çalışmanın temel motivasyon nedenlerinden bir tanesidir.

Yaramaz e-posta tespiti için farklı teknik ve yöntemler kullanılmakla birlikte bu çalışmalar genelde İngilizce veri kümeleri üzerinde yapılmıştır. Makale kapsamında yeni bir Türkçe e-posta veri kümesi oluşturulmuştur. Bu çalışmada, Türkçe yaramaz e-posta filtreleme probleminin çözümüne katkı yapılması için iki farklı veri kümesi kullanılmıştır. Yaramaz e-posta tespitinde öznitelik seçiminin sınıflandırma performansına etkisi gösterilmiştir. Çalışmada çok sayıda metin sınıflandırma ve makine öğrenmesi tekniği bir arada

kullanıldığı için bu alanda çalışacak araştırmacılara yol gösterici olacaktır.

1.2 Organizasyon

Bu makalenin ilk bölümünde e-posta ve yaramaz e-posta tanımları ve literatür araştırması yapılmıştır. Ardından makalenin motivasyonu ve katkısı verilmiştir. Bölüm 2'de, kullanılan veri kümeleri, bu veri kümeleri üzerinde yapılan ön işlem adımları, öznitelik çıkarımı ve özellik seçimi anlatılacaktır. Ayrıca kullanılan makine öğrenmesi algoritmaları ve performans ölçütleri de verilecektir. Bölüm 3'te farklı öznitelik seçim yöntemlerine göre makine öğrenmesi algoritmalarından elde edilen sonuçlar verilip yorumlanacaktır. Bununla birlikte; öznitelik seçimi yapılmadan makine öğrenmesi başarımları da verilerek öznitelik seçiminin başarıma etkisi gösterilecektir. Bölüm 4'te ise çalışmanın genel bir değerlendirmesi yapılıp yaramaz e-posta tespitinde yer alan açık problemler dâhilinde gelecekte yapılması planlanan çalışmalar hakkında bilgi verilecektir.

2. Yöntem

Bu bölümde önerilen yaramaz e-posta tespit sisteminin alt yapısına değinilecektir. Çalışmada önerilen yaramaz e-posta filtreleme sisteminin her bir bileşeni alt bölümler halinde anlatılacaktır.

2.1 Kullanılan Veri Kümeleri

Çalışmada iki adet Türkçe veri kümesi kullanılmaktadır. Birincisi Ergin ve ark. tarafından oluşturulan "TurkishEmail" veri kümesidir [28]. Normal ve yaramaz e-postalardan oluşan bu veri kümesinde toplam 800 adet e-posta yer almaktadır. İkinci veri kümesi ise bu makale kapsamında yazarlar tarafından hazırlanan "TrHamSpamEmailv1.0" e-posta veri kümesidir. Normal ve yaramaz e-postalardan oluşan bu veri kümesinde toplam 350 adet e-posta yer almaktadır. Veri kümelerine ait dağılım Çizelge-1'de verilmektedir. Kullanılan veri kümelerinin her ikisi de dengeli veri kümeleridir.

Çizelge-1: Kullanılan Veri Kümeleri Dağılımı

Veri Kümesi	Normal e-posta	Yaramaz posta	Toplam
TurkishEmail [28]	400	400	800
TrHamSpamEmailv1.0	175	175	350

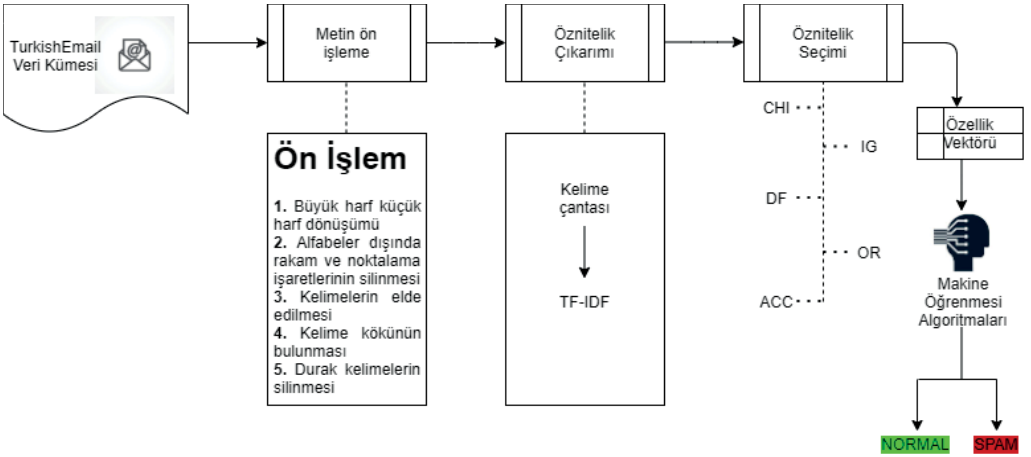
"TurkishEmail" veri kümesinde 400 adet normal e-posta, 400 adet de yaramaz e-posta bulunmaktadır. Veri kümesi üzerinde sözcük dağılımları incelendiğinde; Şekil-1 veri kümesinde bulunma sıklığına göre yaramaz e-posta kümesine ait sözcük bulutunu göstermektedir.

Şekil-1: "TurkishEmail" YaramazE-posta Kelime Bulutu

"TrHamSpamEmailv1.0" veri kümesinde 175 adet normal e-posta, 175 adet de yaramaz e-posta bulunmaktadır. Veri kümesi üzerinde sözcük dağılımları incelendiğinde; Şekil-2 veri kümesinde bulunma sıklığına göre yaramaz e-posta kümesine ait sözcük bulutunu göstermektedir.

Şekil-2: "TrHamSpamEmailv1.0" Yaramaz E-posta Kelime Bulutu

"TrHamSpamEmailv1.0" e-posta veri kümesi yeni oluşturulan bir veri kümesi olduğu için doğal olarak bu veri kümesi üzerinde herhangi bir çalışma yapılmamıştır. "TurkishEmail" veri kümesi üzerinde



Şekil-3: Önerilen YaramazE-posta Filtreleme Sistemi

yapılan birkaç çalışma Çizelge-2 ile özetlenmiştir. Farklı algoritma ve modeller bu çalışmalarda değerlendirilmiştir. Öznitelik seçme yöntemleri ve ön işlem adımlarının başarım sonuçlarına etkilerinin çok fazla olduğu görülmektedir. Bunun yanında, çalışmada önerilen yaramaz e-posta filtreleme sistemi Şekil-3'te verilmektedir.

Çizelge-2: "TurkishEmail" Veri Kümesi Üzerinde Yapılan Çalışmalar

Çalışma	Kullanılan Yöntemler	Açıklama
Ergin ve ark. [28]	Olasılıklı ve İkili (Binary) Bayes	İkili Bayes %93 başarım
Ateş [4]	NB, DVM ile Gauss Karışım Modeli	Doğrusal DVM ile %99 başarım
Deniz ve ark. [25]	Distributed Bag of Word (DBoW) ve Distributed Memory (DM) öznitelik seçim yöntemleri	DBoW+DM modeli üzerinde uygulanan destek vektör makinesi algoritması ile %78,75 başarım
Karamollaoğlu ve ark. [26]	NB, Vektör Uzay Modeli	NB algoritmasıyla %95,5, Vektör Uzay Modeliyle %93,5 başarım
Kaynar ve ark. [27]	Derin öğrenme tekniklerinden oto kodlayıcı	İnce ayar (fine tuning) öncesi %98
		ince ayar sonrası %97 başarım
Eryılmaz ve ark. [1]	Derin öğrenme kütüphanesi Keras ile LSTM modeli	Farklı aktivasyon ve en iyileme yöntemleri ile LSTM %100 başarım
Eryılmaz ve ark. [29]	CHI ve IG öznitelik seçme yöntemiyle ayrım gücü en yüksek 250 sözcük seçilerek, yedi farklı makine öğrenmesi algoritması kullanılmış	CHI öznitelik seçme yöntemi ve Destek vektör makinesine dayalı SMO algoritmasıyla %98,5, IG öznitelik seçme yöntemiyle MLP %98,4 başarım

2.2 Ön İşlem Adımları

Elektronik postalar da metin verilerindendir. Metin verileri yapısal olmayan veri kümeleri arasında yer almaktadır [30]. Her ne kadar metin belgeleri bilgisayarlar tarafından okunulup görüntülense de bu belgeler üzerinde makine öğrenmesi algoritmalarının çalıştırılabilmesi için çeşitli işlemlerden geçirilerek yapısal veri kümelerine dönüştürülmelidir. Bu adımlar, metnin insan dilinden makine tarafından anlaşılabilir formata aktarmak için gerekmektedir. Bu adımların en başında ön işlem adımı gelmektedir. Çalışmanın ön işlem aşamasında uygulanan adımlar şöyledir:

- Tüm harfler büyük harften küçük harfe dönüştürülür.
- Türkçe ve İngilizce dilinde yer alan alfabeler dışında tüm karakterler ve noktalama işaretleri silinir.
- Rakamlar çıkartılır.
- Metin içerisinde yer alan belli başlı kısaltmalar genişletilir.
- Kelimeler boşluk karakterine göre ayrılır.
- Sonuçta elde edilen her sözcük Türkçe doğal dil işleme kütüphanesi olan Zemberek [31] yazılımı vasıtasıyla köklerine ayrıştırılır.
- Durak sözcükler çıkartılır.

Ön işlem adımının uygulanmasıyla e-postalar içerisinde yer alan tüm sözcüklerin standartlaştırılması sağlanmaktadır. Örneğin "otel" sözcüğü ile "OTEL" sözcüğü aynı öneme sahiptir. Eğer ön işlem adımı uygulanmazsa, bu iki sözcük bilgisayar tarafından farklı sözcükler gibi değerlendirilecektir. Bunun sonucu olarak özellik vektörü farklı olacak ve makine öğrenmesi algoritması farklı sonuç verecektir. Bu gibi nedenlerden dolayı ön işlem adımında uygulanan her bir aşama sınıflandırma başarımı için oldukça önemlidir.

2.3 Özellik Vektörünün Oluşturulması

Ön işlem adımının ardından makine öğrenmesi algoritmalarının anlayabileceği sayısal girdiyi elde edebilmek için özellik vektörü oluşturulmaktadır. Geliştirilen yaramaz e-posta filtreleme sisteminin özellik vektörü oluşturulması aşamasında, sözcük çantası ve terim frekansı-ters doküman frekansı (TF-IDF) yaklaşımları uygulanmaktadır. Kelime çantası yaklaşımıyla her bir e-posta dosyası sahip olduğu sözcüklere göre bir vektör gibi temsil edilmektedir. Ardından TF-IDF tekniğine göre her sözcük ağırlıklandırılarak vektör sayısal biçime dönüştürülmektedir.

2.3.1 Öznitelik Çıkarımı

Kelime Çantası (BOW), metin belgelerinden özellikleri ayıklama yöntemidir. Ayrıca bu özellikler, makine öğrenimi algoritmalarını eğitmek için kullanılabilir. Kelime Çantası, Eğitim veri kümesindeki tüm belgede bulunan tüm benzersiz sözcüklerin sözlüğünü oluşturur [24]. TF-IDF algoritması iki çarpandan oluşmaktadır. Bu çarpanlardan biri TF olurken diğeri ise IDF çarpanıdır. Literatürde TF çarpanının hesaplanmasında çok sayıda yaklaşım uygulanmaktadır [32]. Bu çalışmada kullanılan TF çarpanı, ilgili terimin bir e-posta dosyasında kaç kez tekrarlandığının sayısıdır. TF çarpanı Eşitlik 1'e göre hesaplanmaktadır. Burada j. e-posta dosyasında geçen i. terimin TF hesabı yapılmaktadır.

$$TF(t_i, d_j) = frekans(t_i, d_j) \quad (1)$$

IDF ise Eşitlik 2'ye göre hesap edilmektedir. Burada i. terimin IDF hesabı yapılmaktadır. Denklemde N toplam e-posta dosya sayısını gösterirken, df_{t_i} ise i. terimin kaç tane e-posta dosyasında görüldüğünün sayısı olmaktadır. Eşitlik 1 ve Eşitlik 2 çarpılarak i. terime ait TF-IDF değeri hesaplanmaktadır.

$$IDF_{t_i} = \log\left(\frac{N}{df_{t_i}}\right) \quad (2)$$

2.3.2 Öznitelik Seçimi

Metin sınıflandırma probleminde olduğu gibi yaramaz e-postaların sınıflandırılması probleminde de çok sayıda terim ortaya çıkmaktadır. Bu terimlerin çoğu normal e-postalar ile yaramaz e-postalar arasında ayrım yapacak özellikte değildir. Bu nedenle makine öğrenmesi algoritmalarının hem cevap süresi uzamakta hem de sınıflandırma başarımları düşmektedir. Bu sözcüklerin kullanılmaması geliştirilen sistem için önem arz etmektedir. Bütün sözcükleri kullanmak yerine ilgili en iyi alt kümenin seçilmesi işlemine öznitelik seçimi denilmektedir [33]. Bu çalışmada metin sınıflandırma çalışmalarında sıkça kullanılan filtreleme tabanlı Ki-kare, Bilgi Kazancı, Doküman Frekansı Eşikleme, Odds Oranı ve ACC öznitelik

seçme yöntemleri kullanılmaktadır [34, 35, 36]. Filtreleme tabanlı öznitelik seçme yöntemlerinde ilgili öznitelik seçme yöntemin matematiksel denkleminden yararlanılarak her bir özniteliğe ait ilişki skoru hesaplanmaktadır. Bu ilişki skorları büyükten küçüğe doğru sıralanarak en anlamlı özniteliklerin üst sıralarda kalması sağlanmaktadır. Böylece skoru en yüksek n tane öznitelikten oluşan özellik vektörü alt kümesiyle sınıflandırma yapılması gerçekleştirilecektir. Burada $n = 250$ seçilerek skoru en yüksek 250 terim ile sınıflandırma yapılmaktadır. Çalışmada kullanılan filtreleme tabanlı öznitelik seçme yöntemlerinin matematiksel altyapılarında kullanılan parametreler Çizelge 3'te verilmektedir.

Çizelge-3: Filtreleme Tabanlı Öznitelik Seçme Yöntemlerinin Matematiksel Altyapısında Kullanılan Parametreler

Parametreler	Açıklaması
a	t terimini içeren c_i sınıfındaki e-posta dosyalarının sayısı
b	t terimini içermeyen c_i sınıfındaki e-posta dosyalarının sayısı
c	t terimini içeren ve c_i sınıfına ait olmayan e-posta dosyalarının sayısı
d	t terimini içermeyen ve c_i sınıfına ait olmayan e-posta dosyalarının sayısı
N	toplam e-posta dosyalarının sayısını başka bir ifadeyle $a+b+c+d$ toplamı
M	toplam sınıf sayısı
$P(c_i)$	bir e-posta dosyasının c_i sınıfına ait olma olasılığı
$P(t)$	t teriminin külliyat içerisinde bir belgeye dâhil edilme olasılığı
$P(t')$	t teriminin külliyat içerisinde bir belgeye dâhil edilmeme olasılığı
$P(c_i \mid t)$	t teriminin c_i sınıftaki belgelerden birinde en az bir kez geçme olasılığı
$P(c_i \mid t')$	t teriminin c_i sınıfındaki belgelerin hiçbirinde görülmeme olasılığı

Bu yöntemlere ait matematiksel gösterimler şöyledir:

Eşitlik 3'te CHI metriğinin matematiksel gösterimi verilmektedir.

$$CHI(t, c_i) = N \frac{(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)} \quad (3)$$

Eşitlik 3'te c_i sınıfında yer alan t terimine ait CHI skoru hesaplanmaktadır.

Eşitlik 4'te c_i sınıfında yer alan t terimine ait DF skoru hesaplanmaktadır.

$$DF(t, c_i) = a + c \quad (4)$$

Eşitlik 5'te c_i sınıfında yer alan t terimine ait ACC skoru hesaplanmaktadır.

$$ACC(t, c_i) = a - c \quad (5)$$

Eşitlik 6'da c_i sınıfında yer alan t terimine ait OR skoru hesaplanmaktadır.

$$OR(t, c_i) = \frac{a*d}{b*c} \quad (6)$$

Eşitlik 7'de c_i sınıfında yer alan t terimine ait IG skoru hesaplanmaktadır.

$$\begin{aligned} IG(t, c_i) &= -\sum_{i=1}^{M} P(c_i)\, log\, P(c_i) \\ &+ P(t) \sum_{i=1}^{M} P(c_i \mid t)\, log\, P(c_i \mid t) \\ &+ P(t') \sum_{i=1}^{M} P(c_i \mid t')\, log\, P(c_i \mid t') \end{aligned} \quad (7)$$

Bu öznitelik seçme yöntemleri kullanılarak normal e-postalar ile yaramaz e-postalar arasında ayrım gücü en yüksek 250 sözcük seçilerek özellik vektörünün boyutu indirgenmektedir.

2.4. Kullanılan Makine Öğrenmesi Algoritmaları

Önerilen sistemin başarımını ölçmek için yedi farklı makine öğrenmesi algoritması kullanılmaktadır. Bunlar RF, C4.5 karar ağacı, SMO, KNN, LR, NB ve MLP algoritmalarıdır. Algoritmalar için WEKA [44] kütüphanesinden faydalanılmaktadır. KNN algoritmasında k değeri 1 seçilerek sınıflandırma işlemi yapılmaktadır. Diğer algoritmalarda ise WEKA aracındaki ön tanımlı ayarlar kullanılmaktadır Literatür incelendiğinde, bu algoritmaların en çok kullanılan makine öğrenme algoritmaları olması bu çalışmada temel kullanım nedenidir. Kullanılan makine öğrenmesi algoritmaları ile ilgili açıklamalar alt bölümlerde verilmektedir.

2.4.1 Rastgele Orman (RF)

Rastgele orman sınıflandırıcısı, farklı şekil ve boyutlara sahip farklı türdeki karar ağaçlarından oluşan bir topluluk ağacı modelidir. Standart ağaçlarda her düğüm, tüm değişkenler arasında en iyi bölünme kullanılarak gerçekleştirilir. Buna karşın rastgele orman algoritmasında, her düğüm, o düğümde rastgele seçilen bir tahminin alt kümesi arasından en iyisi seçilerek bölünür [37].

2.4.2 C4.5 Karar Ağacı

C4.5, bilgi teorisi kavramını kullanarak, ID3 ile aynı şekilde bir dizi eğitim verisinden karar ağaçları oluşturan istatistiksel sınıflandırıcıdır. Bilgi kuramından yararlanarak karar ağacını eniyilemeyi amaçlar. Bunun için de değişkenlerin entropi değerlerini kullanır. Bugüne kadar uygulamada makine öğreniminde bir dönüm noktası olup yaygın olarak kullanılan karar ağacı algoritmasıdır. En yüksek bilgi kazanımı sağlayan tahmin edici değişken tespit edilir ve ağaç bu değişkenden itibaren dallandırılmaya başlanır. Böylece her bir dalın altında veriler dengeli bir biçimde dağılacaktır. İlk tahmin edici değişken tespit edildikten sonra aynı işlem bu defa toplam entropi üzerinden değil, bu belirlenen tahmin edici değişkenin bilgi değeri üzerinden tekrarlanarak geriye kalan tahmin edici değişkenlerden hangisiyle bu belirlenen değişkenin bölümlenmesinin daha fazla bilgi kazanımı sağlayacağı hesaplanır. Bu işlem tüm tahmin edici değişkenler ağaca yerleştirilinceye kadar devam eder. [38].

2.4.3 DVM tabanlı SMO

DVM tasarımının püf noktası, doğrusal kısıtlamalara sahip ikinci dereceden bir programlama (QP) problemini çözmektir. Eğitim verilerinin boyutundaki büyük artışla birlikte, çekirdek matrisini depolamak için bellek alanı, $O(N^2)$ düzeyi ile artmaktadır. Burada N, eğitim verilerinin sayısıdır. Geleneksel teknikler büyük boyutlu problemler için uygun olmamaktadır. Son zamanlarda araştırmacılar bu sorunu çözmek için eğitim verimliliğini artırmada DVM algoritmasını önermişlerdir. Bu yöntemler arasında SMO en popüler olanıdır. Kavramsal olarak basittir, genellikle daha hızlıdır ve diğer DVM algoritmalarından daha iyi ölçeklendirme özelliklerine sahiptir. [39].

2.4.4 k-En Yakın Komşu (KNN)

KNN parametrik olmayan, tembel bir öğrenme algoritmasıdır. Eğitim verilerini öğrenmez, bunun yerine eğitim veri kümesini "ezberler". Bir tahmin yapmak istediğimizde, tüm veri kümesinde en yakın komşuları arar. Algoritmanın çalışmasında bir k değeri belirlenir. Bu k değerinin anlamı bakılacak eleman sayısıdır. Bir değer geldiğinde en yakın k kadar eleman alınarak gelen değer arasındaki uzaklık hesaplanır. Uzaklık hesaplama işleminde genelde Öklid fonksiyonu kullanılır. Öklid fonksiyonuna alternatif olarak Manhattan, Minkowski ve Hamming fonksiyonları da kullanılabilmektedir. Uzaklık hesaplandıktan sonra sıralanır ve gelen değer uygun olan sınıfa atanır [40].

2.4.5 Lojistik Regresyon (LR)

Lojistik regresyon ile en az değişkenin kullanılmasıyla en iyi uyuma sahip olacak şekilde biyolojik olarak kabul edilebilir bağımlı ve bağımsız değişkenler arasındaki ilişkiyi tanımlayabilen bir model oluşturulmaktadır. Bu teknik ile sonucu belirleyen bir veya daha fazla bağımsız değişkeni bulunan veri kümesinin analizi gerçekleştirilir. Karekök hataların toplamını en aza indirgeyen parametreleri seçmek yerine (sıradan doğrusal regresyon gibi), lojistik regresyonda tahmin, örnek değerlerin gözlem olasılığını en yükseğe çıkaran parametrelerin seçilmesiyle gerçekleştirilmektedir [41].

2.4.6 Naive Bayes (NB)

Naive Bayes sınıflandırıcısının temeli Bayes teoremine dayanır. Tembel bir öğrenme algoritmasıdır. Algoritma bir eleman için her durumun olasılığını hesaplar ve olasılık değeri en yüksek olana göre sınıflandırır. Naive Bayes algoritması, metin sınıflandırma alanındaki en etkili yöntemlerden biridir, ancak yalnızca büyük eğitim örnek kümesinde daha doğru bir sonuç elde edebilir. Test kümesindeki bir değerin eğitim kümesinde gözlemlenemeyen bir değeri varsa olasılık değeri olarak 0 verir yani tahmin yapamaz. Bu durum Sıfır Frekans olarak bilinir. Naïve Bayes sınıflandırıcısı, hedef değer verildiğinde öznitelik değerlerinin koşullu olarak bağımsız olduğu şeklindeki basitleştirici varsayıma dayanmaktadır [42].

Naïve Bayes sınıflandırıcı algoritması, denetimli öğrenme için kullanılan bir algoritmadır. Bayes

sınıflandırıcı, bağımlı olaylar üzerinde çalışır. Daha önce meydana gelen aynı olay tespiti veya gelecekte meydana gelecek olayın olasılığını tespit etme üzerinde çalışır. Naïve Bayes, özelliklerin birbirlerinden bağımsız olduğunu varsayan Bayes teoremi üzerinde yapılmıştır. Naïve Bayes sınıflandırıcı tekniği, yaramaz e-postaları sözcük olasılığı ana rol oynadığından sınıflandırmak için kullanılabilir. Spam olarak sıkça geçen ancak normal olarak bulunmayan herhangi bir sözcük varsa, o zaman bu e-posta yaramazdır.

2.4.7 Çok Katmanlı Algılayıcı (MLP)

XOR problemini çözmek için yapılan çalışmalar sonucu ortaya çıkmıştır. İleri beslemeli yapay sinir ağının bir türüdür. Birçok giriş için bir nöron yeterli olmayabilir. Paralel işlem yapan birden fazla nörona ihtiyaç duyulduğunda katman kavramı devreye girer. Katman sayısı en az bir olmak üzere probleme göre değişir ve ihtiyaca göre ayarlanır. Her katmanın çıkışı bir sonraki katmanın girişi olmaktadır. Bu sayede çıkışa ulaşılmaktadır. Her işlem elemanı yani nöron bir sonraki katmanda bulunan bütün nöronlara bağlıdır. Ayrıca katmandaki nöron sayısı da probleme göre belirlenir. Çıkış katmanı önceki katmanlardan gelen verileri işleyerek ağın çıkışını belirler. Sistemin çıkış sayısı çıkış katmanında bulunan eleman sayısına eşittir. Modelde aktivasyon fonksiyonu olarak herhangi bir matematiksel fonksiyon kullanılabilir [43]. Eğitim süresi diğer makine öğrenme algoritmalarına göre uzasa da MLP algoritması yaramaz e-posta tespitinde tercih edilen algoritmalardandır.

2.5 Performans Ölçütleri

Model performansını değerlendirmede kullanılan temel kavramlar tutturma, bulma ve F-ölçütüdür. Modelin başarısı, doğru sınıfa atanan örnek sayısı ve yanlış sınıfa atılan örnek sayısı ile ilgilidir. F-ölçütünün alt yapısı şöyledir:

Doğru-Pozitif (DP): Modelin yaramaz e-posta olarak etiketli mesajı yaramaz olarak tahmin etmesi durumudur.

Doğru-Negatif (DN): Modelin normal e-posta olarak etiketli mesajı normal e-posta olarak tahmin etmesi durumudur.

Yanlış-Pozitif (YP): Modelin yaramaz e-posta olarak etiketli mesajı normal e-posta olarak tahmin etmesi durumudur.

Yanlış-Negatif (YN): Modelin normal e-posta olarak etiketli mesajı yaramaz e-posta olarak tahmin etmesi durumudur.

Çizelge-4 ile ise bu değerlerle oluşturulan karmaşıklık matrisi verilmiştir. Karmaşıklık matrisi tahminlerin doğruluğu hakkında bilgi veren bir ölçüm aracıdır.

$$\text{Tutturma} = \frac{DP}{DP+YP} \quad (8)$$

$$Bulma = \frac{DP}{DP+YN} \quad (9)$$

Tutturma ve bulma ölçütleri kendi başlarına anlamlı bir karşılaştırma sonucu vermek için yeterli değildir. F-ölçütü bu amaç için tanımlanmıştır. Doğru sınıfa atanan örnek sayısı ve yanlış sınıfa atılan örnek sayısını veren F-ölçütü performans metriği kullanılmıştır. Yani F-ölçütü içinde hem tutturma hem de bulma performans metriklerini içinde bulundurur. F-ölçütü tutturma ve bulma değerlerindeki aşırılıkları cezalandırır. Kısaca; F-ölçütü, tutturma ve bulmanın harmonik ortalamasıdır.

F-ölçütü Eşitlik 8 ve Eşitlik 9'dan yararlanılarak Eşitlik 10'da verildiği gibi hesaplanmaktadır.

$$F-\text{ölçütü} = \frac{2*\text{Tutturma}*\text{Bulma}}{\text{Tutturma}+Bulma} \quad (10)$$

Çizelge-4: Karmaşıklık Matrisi

		Gerçek Sınıf	
		Pozitif	Negatif
Tahmin edilen Sınıf	Pozitif	DP	YP
	Negatif	YN	DN

3. Bulgular ve Tartışma

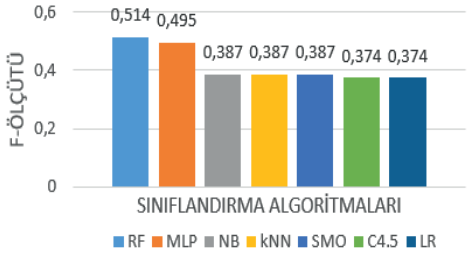
Bu bölümde, Türkçe e-posta veri kümelerinden elde edilen bulgular verilerek yorumlanacaktır.

3.1 Öznitelik Seçimi Yapılmadan Başarım Sonuçları

Elektronik postaların yaramaz olup olmadığının tespiti hala güncel bir konu olup üzerine çalışmalar

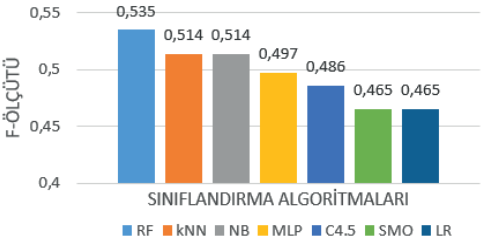
devam etmektedir. Çalışma ile amaçlanan çok sık kullanılan özellik seçme yöntemlerinin e-posta veri kümesine uygulanarak performans artımının olup olmayacağının görülmesidir. Ancak böyle bir karşılaştırma yapabilmek için orijinal özellikler ile yapılan sınıflandırmadan elde edilen sonuçların mutlaka verilmesi gerekmektedir. Aksi halde öznitelik seçiminin başarıya etkisi gözlemlenemez. Bu sebeple bu bölümde öznitelik çıkarılmadan tüm özniteliklerle başarım sonuçları verilmiştir. Yöntemin başarımı 10 kat çapraz doğrulama ile test edilmiştir.

"TurkishEmail" veri kümesinde öznitelik seçimi olmadan 17435 özellik (sözcük sayısı) bulunmaktadır. "TurkishEmail" veri kümesi ile Şekil-4'te öznitelik seçimi olmadan elde edilen sınıflandırma sonuçları verilmektedir. Şekil-4'e göre 7 farklı sınıflandırıcıdan elde edilen sonuçlar karşılaştırıldığında en başarılı e-posta filtreleme algoritması RF olmaktadır. RF algoritmasına göre elde edilen başarım sonucu 0,514 F-ölçütüdür. Sınıflandırma algoritmalarından en kötü sonucu C4.5 ve LR algoritması vermektedir. Bu algoritmalardan elde edilen ortalama başarım sonucu 0,374 F-ölçütüdür.



Şekil-4: "TurkishEmail" Veri Kümesi ile Öznitelik Seçimi Yapılmadan Elde Edilen Sınıflandırma Başarımları

"TRHamSpamEmailv1.0" veri kümesinde öznitelik seçimi olmadan 5394 özellik (sözcük sayısı) bulunmaktadır. "TRHamSpamEmailv1.0" veri kümesi ile Şekil-5'te öznitelik seçimi olmadan elde edilen sınıflandırma sonuçları verilmektedir. Şekil-5'e göre 7 farklı sınıflandırıcıdan elde edilen sonuçlar karşılaştırıldığında en başarılı e-posta filtreleme algoritması RF olmaktadır. RF algoritmasına göre elde edilen ortalama başarım sonucu 0,535 F-ölçütüdür. Sınıflandırma algoritmalarından en kötü sonucu SMO ve LR algoritması vermektedir. Bu algoritmalardan elde edilen ortalama başarım sonucu 0,465 F-ölçütüdür.



Şekil-5: "TRHamSpamEmailv1.0" Veri Kümesi ile Öznitelik Seçimi Yapılmadan Elde Edilen Sınıflandırma Başarımları

Öznitelik seçimi yapılmadan tüm sözcükler öznitelik olarak kullanıldığında başarım sonuçlarının çok düşük olduğu görülmektedir. İki farklı Türkçe e-posta veri kümesi ile öznitelik seçimi yapılmadan RF algoritması en başarılı yöntem olmuştur. Her iki veri kümesinde de en az başarım sonucu LR algoritması ile elde edilmiştir. Makine öğrenmesi algoritmaları ile öznitelik seçimi yapılmadan düşük başarım sonuçları elde edildiği için aynı veri kümeleri ile öznitelik seçim yapılarak elde edilen sonuçların da görülmesi gerekmektedir.

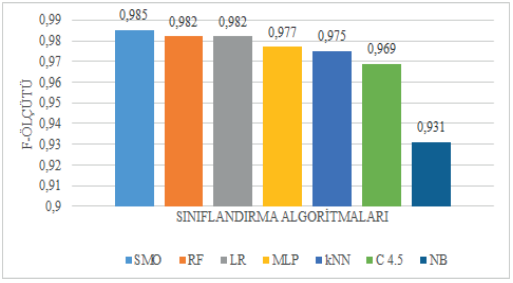
3.2 Öznitelik Seçimi ile Başarım Sonuçları

Bu çalışmada metin sınıflandırma çalışmalarında genellikle kullanılan filtreleme tabanlı CHI, IG, DF, OR ve ACC öznitelik seçme yöntemleri kullanılmıştır. Bu öznitelik seçim teknikleri kullanılarak normal e-postalar ile yaramaz e-postalar arasında ayrım yapabilmek için ayrım gücü en yüksek 250 sözcük seçilmiş ve özellik vektörü indirgenmiştir.

3.2.1 "TurkishEmail" Veri Kümesi ile Öznitelik Seçimi

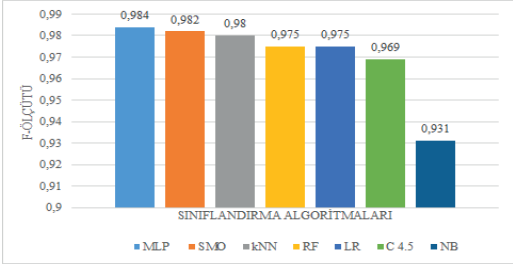
Şekil-6'da CHI öznitelik seçme yönteminin kullanılmasıyla elde edilen sınıflandırma sonuçları verilmektedir. Şekil-6'ya göre 7 farklı sınıflandırıcıdan elde edilen sonuçlar karşılaştırıldığında en başarılı yaramaz e-posta filtreleme algoritması SMO olmaktadır. SMO algoritmasına göre elde edilen sonuç 0,985'tir. SMO algoritmasından sonra elde edilen en iyi sonuç 0,982

ile LR ve RF algoritmalarından elde edilmektedir. İkili sınıflandırma problemlerinde başarılı sonuçlar veren SMO ve LR algoritmaları, Türkçe e-postaların filtrelenmesi probleminde de başarılı sonuçlar vermektedir. Ayrıca birçok ağacın birleşmesiyle meydana gelen RF algoritmasının performansı göz ardı edilmemelidir. Sınıflandırma algoritmalarından en kötü sonucu NB algoritması vermektedir. NB algoritmasından elde edilen sonuç 0,931'dir.



Şekil-6: "TurkishEmail" Veri Kümesi ile CHI Öznitelik Seçme Metriğine Göre Sınıflandırma Başarımları

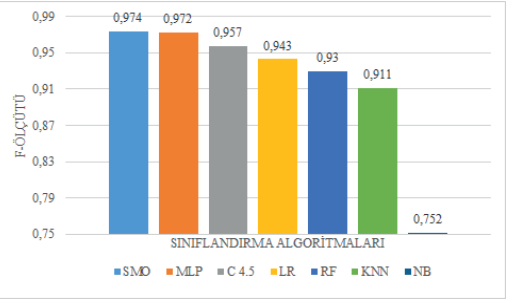
Şekil-7'de IG öznitelik seçme yönteminin kullanılmasıyla elde edilen sınıflandırma sonuçları verilmektedir. Şekil-7'ye göre 7 farklı e-posta sınıflandırma algoritmasından elde edilen sonuçlar karşılaştırıldığında en başarılı algoritma MLP olmaktadır. MLP algoritmasına göre elde edilen sonuç 0,984'tür. MLP algoritmasından sonra elde edilen en iyi sonuç 0,982 ile SMO algoritmasıdır. E-posta sınıflandırma algoritmalarından en kötü sonucu NB algoritması vermektedir. NB algoritmasından elde edilen sonuç 0,931'dir.



Şekil-7: "TurkishEmail" Veri Kümesi ile IG Öznitelik Seçme Metriğine Göre Sınıflandırma Başarımları

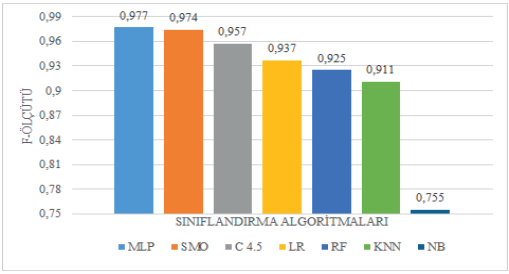
CHI ve IG öznitelik seçme yöntemleri birbirleriyle kıyaslanacak olursa her iki yöntem de birbirlerine yakın sonuçlar üretmektedir. CHI ve IG yöntemlerinden çıkarılan öznitelikler farklılık gösterseler de, C4.5 ve NB algoritmasından elde edilen sınıflandırma sonuçlarında farklılık gözlemlenmemektedir. Bu sonuç her iki yöntemden çıkarılan özniteliklerin sınıflandırma ayrımına aynı oranda katkı yaptığını göstermektedir. CHI ile elde edilen öznitelikler IG yöntemine göre RF, SMO ve LR algoritmalarının sınıflandırma performanslarını arttırmaktadır. Buna karşın MLP ve KNN algoritmalarının performanslarını azaltmaktadır. Başka bir ifadeyle, MLP ve KNN algoritmaları için IG yöntemiyle elde edilen öznitelikler daha anlamlıdır.

Şekil-8'de ACC öznitelik seçme yönteminin kullanılmasıyla elde edilen sınıflandırma sonuçları verilmektedir. Şekil-8'e göre 7 farklı e-posta sınıflandırma algoritmasından elde edilen sonuçlar karşılaştırıldığında en başarılı algoritma SMO olmaktadır. SMO algoritmasına göre elde edilen sonuç 0,974'tür. SMO algoritmasından sonra elde edilen en iyi sonuç 0,972 ile MLP algoritmasıdır. E-posta sınıflandırma algoritmalarından en kötü sonucu NB algoritması vermektedir. NB algoritmasından elde edilen sonuç 0,752'dir.



Şekil-8: "TurkishEmail" Veri Kümesi ile ACC Öznitelik Seçme Metriğine Göre Sınıflandırma Başarımları

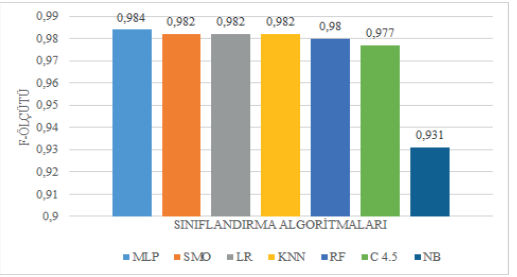
Şekil-9'da OR öznitelik seçme yönteminin kullanılmasıyla elde edilen sınıflandırma sonuçları verilmektedir. Şekil-9'a göre 7 farklı e-posta sınıflandırma algoritmasından elde edilen sonuçlar karşılaştırıldığında en başarılı algoritma MLP olmaktadır. MLP algoritmasına göre elde edilen sonuç 0,977'dir. MLP algoritmasından sonra elde edilen en iyi sonuç 0,974 ile SMO algoritmasıdır. E-posta sınıflandırma algoritmalarından en kötü sonucu NB algoritması vermektedir. NB algoritmasından elde edilen sonuç 0,755'dir.



Şekil-9: "TurkishEmail" Veri Kümesi ile OR Öznitelik Seçme Metriğine Göre Sınıflandırma Başarımları

ACC ve OR öznitelik seçme yöntemleri birbirleriyle kıyaslanacak olursa her iki yöntem de birbirlerine yakın sonuçlar üretmektedir. ACC metriğinden çıkartılan öznitelikler RF algoritmasına girdi olarak verildiğinde elde edilen sınıflandırma başarımı 0,93 olmaktadır. Aynı algoritmaya OR öznitelik seçme yönteminden çıkartılan özellik vektörü girdi olarak verilirse sınıflandırma başarımı 0,925 olmaktadır. Hem ACC hem de OR metriğinden çıkartılan öznitelikler C4.5 algoritmasına girdi olarak verildiğinde elde edilen sınıflandırma başarımları aynıdır. Bu sonuç 0,957 olmaktadır. C4.5 algoritmasına benzer bir durum SMO ve KNN algoritmalarında da vardır. KNN algoritması hem ACC hem de OR metriği ile 0,911 sınıflandırma başarımı vermektedir. SMO algoritması ise ACC ve OR metriği altında 0,974 sınıflandırma başarımına ulaşmaktadır. IG ve CHI yöntemlerinde RF algoritması C4.5 algoritmasına göre daha başarılı olurken, ACC ve OR yöntemlerinde ise C4.5 algoritması RF algoritmasına göre daha başarılı olmaktadır. Bunun nedeni ACC ve OR ile elde edilen öznitelikler C4.5 algoritması için daha ayırt edici olurken, RF algoritması için daha az ayırt edici özelliğe sahiptir. ACC metriğinden çıkartılan öznitelikler LR algoritmasına girdi olarak verildiğinde elde edilen sınıflandırma başarımı 0,943 olmaktadır. Aynı algoritmaya OR öznitelik seçme yönteminden çıkartılan vektör girdi olarak verilirse sınıflandırma başarımı 0,937 olmaktadır.

Şekil-10'da DF öznitelik seçme yönteminin kullanılmasıyla elde edilen sınıflandırma sonuçları verilmektedir. Şekil-10'a göre 7 farklı e-posta sınıflandırma algoritmasından elde edilen sonuçlar karşılaştırıldığında en başarılı algoritma MLP olmaktadır. MLP algoritmasına göre elde edilen sonuç 0,984'dür. MLP algoritmasından sonra elde edilen en iyi sonuç 0,982'dir. Bu sonuç SMO, KNN ve LR algoritmaları tarafından elde edilmektedir. E-posta sınıflandırma algoritmalarından en kötü sonucu NB algoritması vermektedir. NB algoritmasından elde edilen sonuç 0,931'dir.



Şekil-10: "TurkishEmail" Veri Kümesi ile DF Öznitelik Seçme Metriğine Göre Sınıflandırma Başarımları

DF öznitelik seçme yöntemi diğer yöntemler ile kıyaslandığında genel olarak OR ve ACC metriklerinden daha başarılı sonuç vermektedir. Bu durum NB, LR, KNN ve RF algoritmalarında görülmektedir. Bu algoritmalardan elde edilen sonuçlar incelendiğinde, DF öznitelik seçme yönteminin başarımı OR ve ACC metriklerinden daha yüksektir. DF öznitelik seçme yöntemi CHI ve IG öznitelik seçme yöntemleri ile kıyaslanacak olursa bu üç yöntem birbirlerine yakın sonuçlar üretmektedir. Bu yöntemlerden çıkarılan öznitelikler farklılık gösterse de, sınıflandırma algoritmalarından elde edilen sonuçlarda farklılık gözlemlenmemektedir. Bu sonuç her üç yöntemden çıkarılan özniteliklerin sınıflandırma ayrımına aynı oranda katkı yaptığını göstermektedir.

Naive Bayes algoritması, metin sınıflandırma alanındaki en etkili yöntemlerden biridir, ancak yalnızca büyük eğitim örnek kümesinde daha doğru bir sonuç elde edebilir. NB algoritması "TurkishEmail" veri kümesi için, beş öznitelik seçme yönteminde de en düşük başarım sonuçları vermiştir. Eğitim kümesi ve öznitelik sayısı artırıldığında muhtemelen daha yüksek başarıma ulaşacaktır. ACC ve OR öznitelik seçim yöntemi en belirgin öznitelikler yönünden kontrol edildiğinde birbirine en fazla benzeyen öznitelik kümesine sahip olduğu görülmektedir.

Şekil 6-10 arasında, Türkçe e-posta veri kümesi üzerinde ön işlem adımlarından sonra 5 farklı öznitelik seçme yöntemi kullanılarak 7 farklı makine öğrenmesi algoritmasının başarım sonuçları verilmiştir. Bu sonuçların yanında her öznitelik

seçme yönteminden elde edilen en belirgin 10 öznitelik Çizelge-5'te verilmektedir. Çizelge-5'te verilen öznitelikler yaramaz e-postaları normal e-postalardan ayıran metriklerin bulduğu en ayırt edici terimlerdir.

Beş öznitelik seçim yönteminde de en belirgin 10 öznitelik arasında **ürün, tıkla, sadece, fırsat, indir** sözcükleri vardır. CHI ve IG en belirgin 10 öznitelikleri tamamıyla aynı sözcüklerdir. ACC ve OR en belirgin öznitelikleri de tamamıyla aynı sözcüklerdir. En belirgin 10 öznitelik içinde; CHI ve IG ile ACC ve OR arasında 8, DF ile 6 öznitelik aynıdır. ACC ve OR ile DF arasında 5 öznitelik aynıdır.

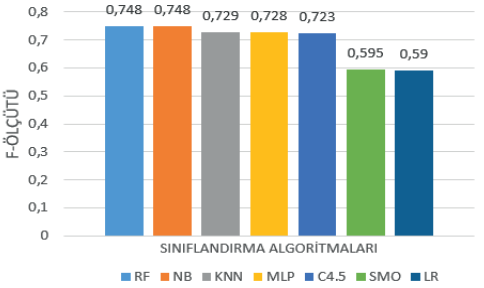
Çizelge-5: "TurkishEmail" Veri Kümesi YaramazE-postaları Normal E-postalardan Ayıran En Belirgin Öznitelikler

CHI	IG	ACC	OR	DF
tıkla	tıkla	tıkla	tıkla	tl
fırsat	fırsat	fırsat	fırsat	sadece
indir	indir	indir	indir	kullan
bülten	bülten	bülten	bülten	fırsat
sadece	konakla	sadece	sadece	tıkla
ürün	grupfon	ürün	ürün	indir
grupfon	ürün	dahil	dahil	yerin
konakla	sadece	adet	adet	grupfon
kahvaltı	kahvaltı	konakla	konakla	iste
paket	paket	kahvalt	kahvalt	ürün

Çizelge-5 incelendiğinde CHI, IG, ACC ve OR öznitelik seçme yöntemleri **tıkla, fırsat, indir** gibi istenilmeyen e-postalarda sıkça görülen terimleri ilk sırada getirmektedir. Bu sözcükler DF metriği tarafından ilk sırada seçilmese de listede yer almaktadır. DF metriğinin ilk sırasında **tl** ikinci sırasında ise **sadece** sözcüksi yer almaktadır. CHI ve IG metriklerinin buldukları ilk 10 özniteliğin tamamı aynı olurken sadece sıralamalarda farklılık görülmektedir. Bu durum ACC ve OR metriklerinde de vardır. Bu metrikler tarafından bulunan ilk 10 öznitelikte birbirlerinin aynısıdır. CHI ve IG metrikleri tarafından seçilen 10 özniteliğin 8 tanesi ACC ve OR metrikleri listesinde yer almaktadır. CHI ve IG metriklerinin öznitelik listesinde yer alan **grupfoni** ve **paket** sözcükleri ACC ve OR metrikleri tarafından en ayırt edici 10 öznitelik arasında yer almamaktadır. Bu sözcükler yerine **ürün** ve **dahil** sözcükleri seçilmektedir. DF metriğinin listesinde yer alan 6 sözcük IG ve CHI listesinde görülmektedir. Buna karşın DF metriği listesinde yer alan 5 sözcük aynı zamanda ACC ve OR metrikleri tarafından seçilmektedir. Sınıflandırma yapılırken 250 öznitelik kullanıldığı göz önünde bulundurulursa, metrikler tarafından seçilen en ayırt edici 10 sözcüğün verilmesi sınıflandırma performansı ile doğrudan ilişkilendirilemez. Fakat metrikler tarafından benzer sözcüklerin bulunması ve sınıflandırma başarımlarının genellikle yakın sonuçlar vermesi kullanılan öznitelikler arasında çok büyük farklılıkların olmadığını göstermektedir.

3.2.2 "TRHamSpamEmailv1.0" Veri Kümesi ile Öznitelik Seçimi

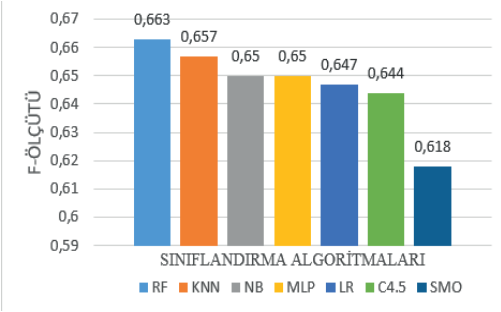
Şekil-11'de CHI öznitelik seçme yönteminin kullanılmasıyla elde edilen sınıflandırma sonuçları verilmektedir. Şekil-11'e göre 7 farklı sınıflandırıcıdan elde edilen sonuçlar karşılaştırıldığında en başarılı yaramaz e-posta filtreleme algoritması RF ve NB olmaktadır. RF ve NB algoritmalarından elde edilen sonuç 0,748'dir. Bu algoritmalardan sonra elde edilen en iyi sonuç 0,729 ile KNN algoritmasından elde edilmektedir. İkili sınıflandırma problemlerinde başarılı sonuçlar veren SMO ve LR algoritmaları, bu veri kümesi üzerinde diğer algoritmalar kadar başarılı olamamaktadır. Bu durum CHI metriği ile seçilen öznitelikler ve veri kümesi üzerindeki dağılımlardan kaynaklanmaktadır.



Şekil-11: "TRHamSpamEmailv1.0" Veri Kümesi ile CHI Öznitelik Seçme Metriğine Göre Sınıflandırma Başarımları

Şekil-12'de IG öznitelik seçme yönteminin kullanılmasıyla elde edilen sınıflandırma sonuçları verilmektedir. Şekil-12'ye göre 7 farklı e-posta sınıflandırma algoritmasından elde edilen sonuçlar karşılaştırıldığında en başarılı algoritma RF olmaktadır. RF algoritmasına göre elde edilen sonuç

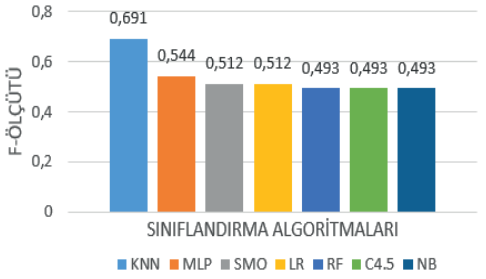
0,663'tür. RF algoritmasından sonra elde edilen en iyi sonuç 0,657 ile KNN algoritmasıdır. E-posta sınıflandırma algoritmalarından en kötü sonucu SMO algoritması vermektedir. SMO algoritmasından elde edilen sonuç 0,618'dir.



Şekil-12: "TRHamSpamEmailv1.0" Veri Kümesi ile IG Öznitelik Seçme Metriğine Göre Sınıflandırma Başarımları

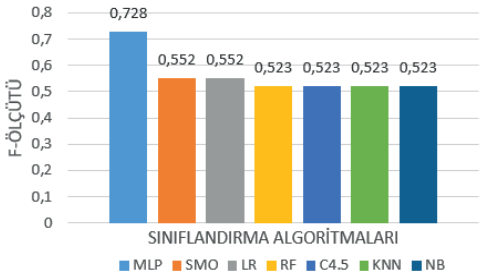
CHI ve IG öznitelik seçme yöntemleri birbirleriyle kıyaslanacak olursa CHI yöntemi ile elde edilen sonuçlar genel olarak IG yöntemine göre elde edilen sonuçlardan daha iyidir. IG ve CHI yöntemlerinin kullanılmasıyla seçilen öznitelikler SMO ve LR algoritmalarına verildiğinde, IG yöntemine göre sınıflandırma daha başarımı daha başarılı bir şekilde yapılmaktadır. Buna karşın diğer sınıflandırma algoritmalarında CHI yöntemi ile seçilen öznitelikler daha anlamlı olmaktadır. Bir başka ifadeyle IG yöntemi SMO ve LR algoritmaları için daha ayırt edici öznitelikleri bulurken, CHI diğer algoritmalar için daha ayırt edici öznitelikler bulmaktadır.

Şekil-13'de ACC öznitelik seçme yönteminin kullanılmasıyla elde edilen sınıflandırma sonuçları yer almaktadır. Şekil-13'e göre 7 farklı e-posta sınıflandırma algoritmasından elde edilen sonuçlar karşılaştırıldığında en başarılı algoritma KNN olmaktadır. KNN algoritmasına göre elde edilen sonuç 0,691'dir. KNN algoritmasından sonra en iyi sonuç 0,544 ile MLP algoritmasından alınmaktadır. ACC yöntemine göre elde edilen öznitelikler RF, C4.5 ve NB algoritmalarına girdi olarak verildiğinde sınıflandırma başarımlarında dikkate değer bir azalma görülmektedir. Bu algoritmalardan elde edilen sınıflandırma başarımları 0.493'dür.



Şekil-13: "TRHamSpamEmailv1.0" Veri Kümesi ile ACC Öznitelik Seçme Metriğine Göre Sınıflandırma Başarımları

Şekil-14'de OR öznitelik seçme yönteminin kullanılmasıyla elde edilen sınıflandırma sonuçları gösterilmektedir. Şekil-14'e göre 7 farklı e-posta sınıflandırma algoritmasından elde edilen sonuçlar karşılaştırıldığında en başarılı algoritma MLP olmaktadır. MLP algoritmasına göre elde edilen sonuç 0,728'dir. MLP algoritmasından sonra en iyi sonuç 0,552 ile SMO ve LR algoritmalarından elde edilmektedir. Bu algoritmaların dışında kalan algoritmalar ile sınıflandırma yapıldığında 0,523 oranında sınıflandırma başarımı alınmaktadır.

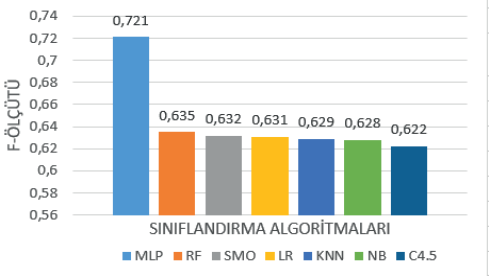


Şekil-14: "TRHamSpamEmailv1.0" Veri Kümesi ile OR Öznitelik Seçme Metriğine Göre Sınıflandırma Başarımları

ACC ve OR öznitelik seçme yöntemleri birbirleriyle kıyaslanacak olursa OR metriği genel olarak ACC metriğinden daha başarılıdır. ACC metriğinden çıkartılan öznitelikler RF algoritmasına girdi olarak verildiğinde elde edilen sınıflandırma başarımı 0,493 olmaktadır. Aynı algoritmaya OR öznitelik seçme yönteminden çıkartılan öznitelikler girdi olarak verilirse sınıflandırma başarımı 0,523 olmaktadır. OR metriği tarafından seçilen öznitelikler KNN algoritması dışında bütün algoritmalar için daha ayırt

edici olmaktadır. Çünkü ACC metriği için KNN algoritmasının sınıflandırma başarımı 0,691 olurken aynı algoritma OR metriğinde 0,523 sonucunu vermektedir.

Şekil-15'de DF öznitelik seçme yönteminin kullanılmasıyla elde edilen sınıflandırma sonuçları verilmektedir. Şekil-15'e göre 7 farklı e-posta sınıflandırma algoritmasından elde edilen sonuçlar karşılaştırıldığında en başarılı algoritma MLP olmaktadır. MLP algoritmasına göre elde edilen sonuç 0,721'dir. MLP algoritmasından sonra elde edilen en iyi sonuç 0,635'dir. Bu sonuç RF algoritması tarafından elde edilmektedir. Sınıflandırma algoritmalarından en kötü sonucu C4.5 algoritması vermektedir. C4.5 algoritmasından elde edilen sonuç 0,622'dir.



Şekil-15: "TRHamSpamEmailv1.0" Veri Kümesi ile DF Öznitelik Seçme Metriğine Göre Sınıflandırma Başarımları

DF öznitelik seçme yöntemi diğer yöntemler ile kıyaslandığında genel olarak OR ve ACC metriklerinden daha başarılı sonuç vermektedir. Bu durum NB, LR ve RF algoritmalarında görülmektedir. Bu algoritmalardan elde edilen sonuçlar incelendiğinde, DF öznitelik seçme yönteminin başarımı OR ve ACC metriklerinden daha yüksektir. Fakat ACC metriği ile seçim yapılıp KNN algoritmasına girdi olarak verildiğinde ve OR metriği ile seçim yapılıp MLP algoritmasına girdi olarak verildiğinde DF yöntemine göre daha başarılı sonuçlar elde edilmektedir. DF öznitelik seçme yöntemi CHI ve IG öznitelik seçme yöntemleri ile kıyaslanacak olursa genel olarak CHI ve IG yöntemlerinden daha kötü sonuçlar vermektedir.

Yapılan bütün deneyler göz önüne alındığında, "TRHamSpamEmailv1.0" veri kümesi üzerinde en yüksek sınıflandırma başarımı CHI metriği ile öznitelik seçimi yapılıp RF ve NB algoritmalarına girdi olarak verildiğinde elde edilmektedir.

Şekil 11-15 arasında, "TRHamSpamEmailv1.0" Türkçe e-posta veri kümesi üzerinde ön işlem adımlarından sonra 5 farklı öznitelik seçme yöntemi kullanılarak 7 farklı makine öğrenmesi algoritmasının başarım sonuçları verilmiştir. "TRHamSpamEmailv1.0" veri kümesi ile CHI öznitelik seçim yöntemi RF ve NB algoritması ile 0,748 başarıma ulaşmıştır. TurkishEmail veri kümesine göre başarımın düşük olması; Şekil-1 ve Şekil-2 yaramaz e-posta veri kümesi sözcük bulutları incelendiğinde "TRHamSpamEmailv1.0" kümesi için çok farklı sözcüklerin sıklığının birbirine yakın olması ve e-posta veri kümesinin boyutunun azlığı olarak işaret edilebilir.

"TRHamSpamEmailv1.0" veri kümesinden elde edilen sınıflandırma sonuçlarının yanında her öznitelik seçme yönteminden elde edilen en belirgin 10 öznitelik Çizelge-6'da verilmektedir. Çizelge-6'da verilen öznitelikler yaramaz e-postaları normal e-postalardan ayıran metriklerin bulduğu en ayırt edici terimlerdir.

Beş öznitelik seçim yönteminden CHI ve IG birbirlerine daha yakın sözcük gruplarını seçmektedir. Buna karşın ACC ve OR teknikleri birbirlerine yakın sözcük gruplarını seçmektedir. DF metriğinde ise genel olarak bu dört metrikten benzer sözcükler yer almaktadır. CHI, IG ve DF yöntemlerinin seçtiği sözcükler ACC ve OR yöntemlerinin seçmiş oldukları sözcüklere kıyasla istenilmeyen eposta örneklerinde olması daha muhtemeldir. Bu yüzden CHI, IG ve DF yöntemlerinin kullanılmasıyla elde edilen özellik vektörlerinin sınıflandırma başarımları ACC ve OR metriklerine kıyasla daha iyi olmaktadır. En belirgin 10 öznitelik içinde; indir sözcüğü DF metriği dışında bütün metrikler tarafından seçilmektedir. CHI ve IG metrikleri tarafından seçilen 10 sözcükten 9 tanesi ortaktır. IG metriğinin ilk 10 sözcüğü arasında sistem sözcüğü görünmüyorken, CHI metriğinin ilk 10 sözcüğü arasında ise hizmet sözcüğü görünmemektedir. ACC ve OR metrikleri tarafından seçilen 10 sözcükten 7 tanesi ortaktır. ACC metriğinin ilk 10 sözcüğü arasında iste, şimdi ve takip sözcükleri yokken, OR metriğinin ilk 10 sözcüğü arasında ise biri, premium ve yakala sözcükleri yer almamaktadır. Sınıflandırma yapılırken 250 öznitelik kullanıldığı göz önünde bulundurulursa, metrikler tarafından seçilen en ayırt edici 10 sözcüğün verilmesi sınıflandırma performansı ile doğrudan ilişkilendirilemez. Fakat

metrikler tarafından farklı sözcükler seçildiği için sınıflandırma başarımlarının farklı çıkması yorumunun yapılabilmesine imkân verir.

Çizelge-6: "TRHamSpamEmailv1.0" Veri Kümesi Yaramaz E-postaları Normal E-postalardan Ayıran En Belirgin Öznitelikler

CHI	IG	ACC	OR	DF
kullan	kullan	indir	indir	bilgi
öğren	öğren	fırsat	fırsat	kullan
bilgi	bilgi	eyewire	başla	özel
link	çalış	başla	eyewire	çalış
çalış	indir	biri	kaçır	veri
indir	fırsat	yarış	yarış	fazla
eyewire	link	kaçır	hücre	öğren
fırsat	uygula	hücre	iste	iste
sistem	eyewire	premium	şimdi	başla
uygula	hizmet	yakala	takip	alan

4. Sonuçlar ve Öneriler

Bu çalışmada metin madenciliği ve makine öğrenmesi tekniklerinin bir arada kullanılmasıyla yaramaz Türkçe e-postaların filtrelenmesini sağlayan yapı önerilmektedir. Literatürde Türkçe dilinde yaramaz e-posta filtreleme çalışmaları kısıtlı sayıdadır. Bu nedenle çok sayıda metin sınıflandırma ve makine öğrenmesi teknikleri Türkçe e-postalar üzerine uygulanarak yaramaz e-postalar ile normal e-postalar arasında ayrımın iyi yapılması sağlanmaktadır. Çalışmada elde edilen sonuçlara göre normal e-postalar ile yaramaz e-postaların ayrımının iyi yapıldığı ve bazı makine öğrenmesi algoritmalarının %98'in üzerinde sınıflandırma başarımı gösterdiği gözlenmektedir. Türkçe e-posta veri kümelerinde öznitelik seçiminin makine öğrenmesi algoritmalarında başarım sonuçlarını çok fazla etkilediği çalışma kapsamında kullanılan her iki veri kümesinde de gösterilmiştir. "TurkishEmail" veri kümesi üzerinde bütün sözcükleri öznitelik olarak kullanıp da sonuç alan herhangi bir çalışma görülmemiştir. Bu çalışmada, aynı veri kümesinde yapılan diğer çalışmalara göre, daha fazla öznitelik seçim yöntemi ve daha fazla makine öğrenmesi algoritması test edilmiştir. Bu alanda çalışacak araştırmacılar için mevcut problemlerden bir tanesi Türkçe elektronik postaların olduğu veri kümesi yetersizliğidir. Çalışma kapsamında yeni bir Türkçe e-posta veri kümesi "TRHamSpamEmailv1.0" oluşturulmuştur. Gelecek çalışmalarda Türkçe e-postalardan oluşan "TRHamSpamEmailv1.0" veri kümesinin boyutunun artırılarak geniş kapsamlı denemelerin yapılması hedeflenmektedir. Aynı zamanda oluşturulan Türkçe yaramaz e-posta içeren veri kümesinin boyutunun artırılarak derin öğrenme metotları ile de başarım sonuçlarının değerlendirilmesi planlanmaktadır.

5. Bilgilendirme

Bu çalışma 5. Uluslararası Bilgisayar Bilimleri ve Mühendisliği Konferansında (UBMK 2020) sunulan [29] nolu çalışmanın genel bir yapıya uyarlanmış ve genişletilmiş halidir. Makale kapsamında oluşturulan "TRHamSpamEmailv1.0" Türkçe e-posta veri kümesi yazarlarla iletişime geçildiğinde araştırmacılara temin edilecektir.

Kaynakça

[1] Eryılmaz, E. E., Şahin D. Ö. ve Kılıç, E. *Filtering Turkish Spam Using LSTM From Deep Learning Techniques*, 2020 8th International Symposium on Digital Forensics and Security (ISDFS), IEEE, p. 1-6, 2020.

[2] Eryılmaz, E. E., Kılıç, E. *İstenmeyen E-postaların Tespiti için Kullanılan Yöntemlerin İncelenmesi*, Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi, 11(3), 977-987, 2020.

[3] LeCun, Y., Bengio, Y. ve Hinton, G. *Deep learning,* Nature, 521:7553, 436-444, 2015.

[4] Ates, N. *Support vector machine and gauss mixture model detection of unsolicited e-mails,* Master's thesis, Süleyman Demirel Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı, 2014.

[5] Sharma, A. ve ark., *A Comparative Study Between Naive Bayes and Neural Network (MLP) Classifier for Spam Email Detection*, 2014.

[6] Karthika, R. ve Visalakshi, P. *A hybrid ACO based feature selection method for email spam classification,* WSEAS Trans. Comput 14, 171-177, 2015.

[7] Renuka, D. K., Visalakshi P ve Sankar, T., *Improving E-mail spam classification using ant colony optimization algorithm*, Int. J. Comput. Appl, 22-26, 2015.

[8] Palanisamy, C., Kumaresan, T. ve Varalakshmi S. E., *Combined techniques for detecting email spam using negative selection and particle swarm optimization*, Int. J. Adv. Res. Trends Eng. Technol., 3, 2016.

[9] Zavvar, M., Rezaei M. ve Garavand S., *Email spam detection using combination of particle swarm optimization and artificial neural network and support vector machine*, International Journal of Modern Education and Computer Science, 8(7), 68, 2016.

[10] Foqaha M. A. M., *Email spam classification using hybrid approach of RBF neural network and particle swarm optimization*, International Journal of Network Security & Its Applications, 8(4), 17-28, 2016.

[11] Sharma A. ve Suryawanshi A., *A novel method for detecting spam email using KNN classification with spearman correlation as distance measure*. International Journal of Computer Applications, 136(6), 28-35, 2016.

[12] Alkaht I. J. ve Al-Khatib B., *Filtering SPAM Using Several Stages Neural Networks*, Int. Rev. Comp. Softw., 11, 2, 2016.

[13] Rajamohana S. P., Umamaheswari K. ve Abirami B., *Adaptive binary flower pollination algorithm for feature selection in review spam detection*, 2017 International Conference on Innovations in Green Energy and Healthcare Technologies (IGEHT), pp. 1-4, IEEE, 2017.

[14] Myle O ve ark., *Finding deceptive opinion spam by any stretch of imagination*, ACM Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 309-319, 2011.

[15] Akinyelu A. A. ve Adewumi A. O., *Classification of phishing email using random forest machine learning technique*, Journal of Applied Mathematics, 2014.

[16] Yıldız A., *Kurumsal e-posta sınıflandırma sistemi.* Yüksek Lisans Tezi, Gazi Üniversitesi Fen Bilimleri Enstitüsü, 82, Ankara, 2017.

[17] Şahin E., *Makine öğrenme yöntemleri ve sözcük kümesi tekniği ile yaramaz e-posta / e-posta sınıflaması.* Yüksek Lisans Tezi, Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, 60, Ankara, 2018.

[18] Kale B., *Veri madenciliği sınıflandırma algoritmaları ile e-posta önemliliğinin belirlenmesi.* Yüksek Lisans Tezi, Çukurova Üniversitesi Fen Bilimleri Enstitüsü, 120, Adana, 2018.

[19] Nazlı N., *Analysis of machine learning – based spam filtering techniques,* Yüksek Lisans Tezi, Çankaya University The Graduate School of Natural and Applied Sciences, 79, Ankara, 2018.

[20] Al-Azzawi F., *Wrapper feature selection approach for spam e-mail filtering*, Master Thesis, Erciyes University Graduate school of natural and applied science, Kayseri, 2018.

[21] Ablel-Rheem D. M., Ibrahim A. O., Kasim S., Almazroi A. A., ve Ismail M. A., *Hybrid Feature Selection and Ensemble Learning Method for Spam Email Classification*. International Journal, *9*(1.4), 2020.

[22] Zamir A., Khan H. U., Mehmood W., Iqbal T., ve Akram A. U., *A feature-centric spam email detection model using diverse supervised machine learning algorithms*. The Electronic Library, 2020.

[23] Mohammad R. M. A., *A lifelong spam emails classification model*. Applied Computing and Informatics. 2020.

[24] Kumar N. ve Sonowal S.. *Email Spam Detection Using Machine Learning Algorithms.* In *2020* Second International Conference on Inventive Research in Computing Applications (ICIRCA) (pp. 108-113). IEEE. 2020, July.

[25] Deniz E., Erbay H., Coşar M.. *Türkçe e-postaların Doc2Vec ile sınıflandırılması.* In 2019 1st International Informatics and Software Engineering Conference (UBMYK) (pp. 1-4). IEEE, 2019.

[26] Karamollaoglu H., Dogru İ. A., Dorterler M., *Detection of Spam E-mails with Machine Learning Methods*, 2018

[27] Kaynar O., Görmez Y. ve Işık Y. E., *Oto Kodlayıcı Tabanlı Derin Öğrenme Makinaları İle Spam Tespiti. 3. Uluslararası Yönetim Bilişim Sistemleri Konferansı*, 44. 2016.

[28] Ergin S., Sora Gunal E., Yigit H. ve Aydin R.. *Turkish anti-spam filtering using binary and*

probabilistic models. *Global Journal on Technology*, *1*. 2012

[29] Eryilmaz E. E., Ozkan Şahin D. ve Kılıç E., *Machine Learning Based Spam E-mail Detection System for Turkish*, 2020 5th International Conference on Computer Science and Engineering (UBMK), Diyarbakır, Turkey, pp. 7-12, 2020

[30] Hotho A., Nürnberger A. ve Paaß G., *A brief survey of text mining*, in Ldv Forum, vol. 20, no. 1. Citeseer, pp. 19–62, 2005.

[31] Akın A. A. ve Akın M. D., *Zemberek, an open source nlp framework for turkic languages*, Structure, 10, 1-5, 2007.

[32] Domeniconi G. Ve ark. *A study on term weighting for text categorization: A novel supervised variant of tf.idf*, in DATA, pp. 26–37, 2015.

[33] Şahin D. Ö. ve Kılıç E., *Two new feature selection metrics for text classification*, Automatika, vol. 60, no. 2, pp. 162–171, 2019.

[34] Zheng Z. Ve ark., *Feature selection for text categorization on imbalanced data*, ACM Sigkdd Explorations Newsletter, vol. 6, no. 1, pp. 80-89, 2004.

[35] Forman G., *An extensive empirical study of feature selection metrics for text classification*, Journal of machine learning research, vol. 3, no. Mar, pp. 1289-1305, 2003.

[36] Şahin D. Ö., Ateş N. ve Kiliç E., Feature selection in text classification. 2016 24th signal processing and communication application conference (SIU), IEEE, pp. 1777-1780, 2016.

[37] Liaw A. Ve ark., *Classification and regression by randomforest*, R news, vol. 2, no. 3, pp. 18–22, 2002.

[38] Ruggieri S., *Efficient c4.5 classification algorithm*, IEEE transactions on knowledge and data engineering, vol. 14, no. 2, pp. 438–444, 2002.

[39] Zeng Z.-Q. Ve ark., *Fast training support vector machines using parallel sequential minimal optimization*, 2008 3rd international conference on intelligent system and knowledge engineering, vol. 1. IEEE, pp. 997–1001, 2008.

[40] Cover T. ve Hart P., *Nearest neighbor pattern classification*, IEEE transactions on information theory, vol. 13, no. 1, pp. 21–27, 1967.

[41] Dreiseitl S. ve Ohno-Machado L., *Logistic regression and artificial neural network classification models: a methodology review*, Journal of biomedical informatics, vol. 35, no. 5-6, pp. 352–359, 2002.

[42] Huang Y. ve Li L., *Naive bayes classification algorithm based on small sample set*, in 2011 IEEE International Conference on Cloud Computing and Intelligence Systems. IEEE, pp. 34–39, 2011.

[43] Ruck D. W., S. Rogers K., Kabrisky M., Oxley M.E. ve Suter B. W.. *The multilayer perceptron as an approximation to a Bayes optimal discriminant function.* IEEE Transactions on Neural Networks, *1*(4), 296-298. 1990

[44] Frank E., Mark A. Hall ve Witten I. H., *The WEKA Workbench*. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.