X Platformunda İstenmeyen Gönderilerin Makine Öğrenmesi, Geniş Dil Modelleri ve Bilgisayarlı Görü ile Tespiti

Araştırma Makalesi/Research Article

Ebutalha CAMADAN[1], Mehmet ŞİMŞEK[2]

[1]Millî Savunma Üniversitesi, İstanbul, Türkiye
[2]Millî Savunma Üniversitesi, Kara Harp Okulu, Bilgisayar Mühendisliği Bölümü, Ankara, Türkiye

ebutalha.camadan@msu.edu.tr, mesimsek@kho.msu.edu.tr
(Geliş/Received:05.12.2024; Kabul/Accepted:11.03.2025)
DOI: 10.17671/gazibtd.1596990

Özet— Sosyal medya platformları, günümüzde bilgi paylaşımı ve iletişimde önemli araçlar haline gelirken, aynı zamanda istenmeyen gönderilerin (spam) yayılması da büyük bir sorun teşkil etmektedir. Bu çalışma, X sosyal medya platformundaki (eski adıyla Twitter) istenmeyen gönderilerin tespitine yönelik, makine öğrenmesi, geniş dil modelleri ve bilgisayarlı görü tekniklerini birleştiren yeni bir yaklaşım önermektedir. Türkiye'de popüler olan konulara dair görsel içeren gönderilerden bir veri kümesi oluşturularak, spam tespitinde en etkili makine öğrenmesi algoritmaları belirlenmeye çalışılmıştır. Gönderi içeriğinin etiketlerle ilişkisi ve birden fazla etiketin birbiriyle ilgisi gibi sosyal medya etkileşimini belirleyen öznitelikler geliştirilmiştir. Ayrıca, görsel içeriğin analizi için, görselin X platformunda ilk paylaşıldığı tarih ile internet üzerindeki diğer sayfalarda geçtiği metinle benzerliği gibi görsel odaklı öznitelikler de dahil edilmiştir. Bu öznitelikler, Google Gemini ve Cloud Vision AI araçları kullanılarak geliştirilmiştir. Beş farklı makine öğrenmesi algoritması (Karar Ağaçları, Rastgele Orman, SVM, Lojistik Regresyon, Çok Katmanlı Algılayıcı) ile yapılan deneylerde, Rastgele Orman algoritması en yüksek doğruluk ve F1 skoru değerlerine ulaşmıştır. Bu çalışma, X platformunda istenmeyen gönderi tespiti için makine öğrenmesi yöntemlerinin etkinliğini göstermiş ve Google Gemini ile Cloud Vision AI araçlarının etkin kullanımına dair yeni yöntemler sunmuştur. Ayrıca geliştirilen öznitelikler, spam içeriklerin doğru bir şekilde sınıflandırılmasında güçlü bir temel oluşturmaktadır.

Anahtar Kelimeler— X sosyal medya platformu, istenmeyen gönderi, makine öğrenmesi, geniş dil modelleri, bilgisayarlı görme

Detecting Spams on the X Platform with Machine Learning, Large Language Models, and Computer Vision

Abstract— While social media platforms have become crucial tools for information sharing and communication, the spread of unwanted content (spam) has also become a significant problem. This paper proposes a novel approach for spam detection on the social media platform X (formerly Twitter) by integrating machine learning, large language models, and computer vision techniques. A dataset containing posts with visual content on popular Turkish topics was created, aiming to identify the most effective machine learning algorithms for spam detection. Feature engineering was conducted to capture key aspects of social media interaction, including the relationship between post content and hashtags, as well as the relevance between multiple hashtags. Additionally, image-based features were introduced, such as the initial posting date of an image on X and its textual similarity to other web pages, to enhance visual content analysis. These features were developed using Google Gemini and Cloud Vision AI. Experimental evaluations with five machine learning algorithms (Decision Trees, Random Forest, SVM, Logistic Regression, and Multilayer Perceptron) demonstrated that the Random Forest algorithm achieved the highest accuracy and F1 score. This paper highlights the effectiveness of machine learning methods in spam detection on X and introduces new methodologies for leveraging Google Gemini and Cloud Vision AI. Furthermore, the engineered features provide a strong foundation for accurately classifying spam content.

Keywords— X social media platform, spam message, machine learning, large language models, computer vision

1. GİRİŞ (INTRODUCTION)

Bu bölümde, Online Sosyal Ağlar (OSA'lar) özellikle X platformu üzerinden iletişim ve etkileşimdeki rolünü inceleyerek, istenmeyen gönderilerin (spam) tespitine yönelik güncel yaklaşımları ve bu alandaki zorlukları ele alacağız. Ayrıca, bu çalışmanın temel hedefi, metin ve görsel içeriklerin birleşik analizine dayalı yeni bir spam tespit yaklaşımını tanıtmaktır.

Online Sosyal Ağlar (OSA'lar), özellikle geniş kullanıcı kitlesine sahip X platformu aracılığıyla, günümüzde iletişim ve etkileşimde vazgeçilmez bir araç haline gelmiştir [1-2]. Milyonlarca kullanıcı, X platformunu anılarını paylaşmak, gözlemlerini yayınlamak ve "Gönderi (Tweet)" adı verilen kısa mesajlarla etkileşimde bulunmak için kullanarak platformun toplumsal etkisini artırmaktadır [3].

Bu çalışmanın temel motivasyonu, sağlıklı bir online sosyal ağ ortamı oluşturmak için istenmeyen gönderilerin (spam) tespitinin kritik önem taşımasıdır. İstenmeyen gönderiler, X gibi platformlarda kullanıcı deneyimini olumsuz etkileyebilir. Ayrıca, dolandırıcılık ve kimlik avı gibi güvenlik risklerine yol açarak platformun güvenilirliğini zedeler [4-5]. İstenmeyen gönderiler, genellikle ticari veya kötü niyetli amaçlarla toplu halde paylaşılan, alıcı tarafından istenmeyen ve zararlı içerikler içerebilen mesajlardır. X platformunda bu içerikler, kullanıcıları tehlikeli web sitelerine yönlendiren URL'ler içerebilen kısa mesajlar ("Gönderiler/Tweetler") şeklinde yayılmaktadır [6]. E-postayla kıyaslandığında, X üzerindeki istenmeyen gönderilerin daha yüksek tıklanma oranına sahip olması [7], bu sorunun platform için ciddi bir tehdit olduğunu göstermektedir.

X platformunda istenmeyen gönderi tespitine yönelik güncel yaklaşımlar, Web of Science ve ScienceDirect veritabanları üzerinden 2017 sonrası yayınlar taranarak incelenmiştir. Arama terimi olarak "X/Twitter Spam Detection (X İstenmeyen Gönderi Tespiti)" kullanılmış ve derin öğrenme (DL), makine öğrenmesi, bulanık mantık gibi çağdaş yöntemlere odaklanan çalışmalar önceliklendirilmiştir. Literatürdeki çalışmalar, bu algoritmaların genellikle sözdizimi analizi, özellik çıkarımı ve kara listeleme yöntemlerine dayandığını göstermektedir.

X platformunda istenmeyen gönderi tespiti, mevcut makine öğrenmesi ve kara listeleme yöntemlerinin yetersiz kalması nedeniyle zorlu bir problem teşkil etmektedir. Bu sorunu çözmek için bazı araştırmacılar derin öğrenme tekniklerinden yararlanarak yeni yaklaşımlar önermiştir. Derin öğrenme tabanlı yaklaşımlar (YSA, CNN), geleneksel yöntemlere göre daha başarılı sonuçlar göstermiştir. Ancak, bu modellerin karmaşıklığı ve yüksek veri ihtiyacı, pratik uygulamalarda zorluklar yaratmaktadır [8]. Diğer araştırmacılar [9], kullanıcıların istenmeyen gönderi tespit sistemlerinden kaçınma stratejilerine yönelik hibrit bir teknik geliştirmiştir. Başka bir grup ise gönderi tabanlı analizlerin yetersiz olduğunu ve kullanıcı hesapları ile sosyal grafiğin önemini vurgulamıştır [10].

Bir grup araştırmacı [11], kullanıcı hesaplarına dayalı bir yaklaşım önererek gönderi sıklığının bir gösterge olabileceğini belirtmiştir. Ancak, bu yöntem LSA gibi tekniklere dayansa da, içeriğin anlamını tam olarak yakalayamamaktadır. Madisetty ve Desarkar'ın ensemble yöntemi ve kelime gömme tekniklerini kullandıkları çalışma umut verici olsa da [12], diğer araştırmacıların "Honey Profiles" metodundaki gibi veri toplama zorlukları devam etmektedir [13]. URL ve kara listeleme gibi basit yöntemlerin sınırlılıkları [14] ve sınıflandırma ile kümeleme yöntemlerinin değişken başarı oranları [15-16], tek bir çözümün yetersiz olduğunu göstermektedir.

Araştırmacıların doğal dil işleme ve makine öğrenmesi sınıflandırmaları üzerine çalışmaları [17] ve başka bir grup araştırmacının Naive Bayes önerisi [18], metin analizi ve olasılık hesaplamalarının önemini vurgulamaktadır. Ancak, Naive Bayes'in doğruluk oranı veri kalitesine bağlıdır. Öte yandan, dengesiz veri setlerinin yarattığı zorluklara ve SMOTE gibi tekniklerin önemine dikkat çeken araştırmalar da bulunmaktadır [19-20]. Sumathi ve Raja'nın [21] ETC ve VC kullandıkları yöntemi ile Thomas ve Meshram'ın [22] DNFNet yaklaşımı, karmaşık modellerin potansiyelini göstermektedir. Bununla birlikte, diğer bir grup araştırmacı [23], ensemble tekniklerinin, özellikle RF'nin, bireysel modellere göre daha iyi sonuçlar verdiğini belirtmiştir. Son yıllarda ise derin öğrenme tabanlı yaklaşımlar uygulanmış ve daha önceki yıllarda belirtilen veri ihtiyacı ve pratik uygulama zorluklarına değinilmiştir [24-25]. Diğer bir çalışmada ise araştırmacılar kuantum hesaplama prensiplerini kullanarak ikili kütleçekim arama algoritmasını rastgele orman modeliyle birleştirerek daha yüksek doğruluk oranı yakalamıştır. [26].

Sonuç olarak, mevcut çalışmalar, X platformunda istenmeyen gönderi tespitinin dinamik yapısı nedeniyle tek bir ideal çözümün bulunmadığını ve sürekli gelişen yeni yaklaşımlara ihtiyaç duyulduğunu ortaya koymaktadır. Mevcut çalışmaların çoğunluğu metin tabanlı gönderilere odaklansa da görsel içeriklerin, özellikle manipüle edilmiş, yanıltıcı veya uyumsuz görsellerin yaygınlaşmasıyla birlikte, istenmeyen gönderi tespitinde giderek daha önemli bir rol oynadığı unutulmamalıdır. Görsel içerikler, metin tabanlı gönderilere kıyasla kullanıcıların dikkatini daha kolay çekmektedir [27]. Bu nedenle, görsel içeren gönderilerin tespiti, istenmeyen içeriklerin engellenmesi açısından kritik bir öneme sahiptir.

Bu çalışmada, X sosyal medya platformundaki istenmeyen gönderilerin tespitine yönelik makine öğrenmesi, geniş dil modelleri ve bilgisayarlı görü tekniklerini birleştirerek yeni bir yaklaşım geliştirmeyi amaçladık. Bu doğrultuda, X platformundan veri topladık ve spam tespiti için zenginleştirilmiş bir veri seti oluşturduk. Literatürdeki mevcut çalışmalardan farklı olarak, bu veri seti hem metin hem de görsel içeriklerin analizini bir arada sunan özgün bir yapıya sahiptir. Mevcut çalışmalar genellikle yalnızca

metin tabanlı veya sınırlı görsel analiz tekniklerine odaklanırken, bizim yaklaşımımız, geniş dil modelleri ve bilgisayarlı görü tekniklerini entegre ederek istenmeyen içerik tespitine yeni bir bakış açısı getirmektedir. Ayrıca, sadece seçtiğimiz makine öğrenmesi algoritmalarının performansını kıyaslayarak, literatürdeki diğer çalışmaların sonuçlarıyla doğrudan bir karşılaştırma yapmamaktayız.

Çalışmamızın literatüre katkıları şunlardır:

- Metin ve görsel içeriklerin birleşik bir şekilde analiz edildiği özgün bir veri seti oluşturulması.
- Google Gemini ve Cloud Vision AI kullanarak geniş dil modelleri ile görsel analiz tekniklerinin entegre edilmesi.
- Türkiye'de trend olan konulara dayalı spam içeriklerinin tespiti için özgün bir veri seti ve analiz yaklaşımı geliştirilmesi.

Veri setimizi zenginleştirmek için, sosyal medya gönderilerinin içerik analizini gerçekleştirmek amacıyla Google'ın üretken dil modeli Gemini ve görsel analiz hizmeti Cloud Vision AI kullanılmıştır. Bu teknolojiler, geleneksel yöntemlerden farklı olarak metin ve görsel içerikleri bütünsel bir şekilde analiz etmeyi mümkün kılmaktadır. Bu sayede, gönderi içeriklerinin etiketlerle olan ilişkisini daha iyi anlayıp, görsel öğeleri de analiz edebildik. Türkiye'de trend olan konulara ait görsel içeren gönderileri inceleyerek, istenmeyen içeriklerin tespitine yönelik sağlam bir temel oluşturduk.

Veri setimizin sınıflandırılmasında beş farklı makine öğrenmesi algoritması kullanarak deneyler gerçekleştirdik. Denediğimiz yöntemler arasında Karar Ağaçları, Rastgele Orman, SVM, Lojistik Regresyon ve Çok Katmanlı Algılayıcı yer almaktadır. Deneyler sonucunda, Gemini'nin etiket analizi konusundaki güçlü performansı ve Cloud Vision AI'nın görsel analiz yetenekleri, Rastgele Orman algoritmasının veri seti üzerinde en yüksek doğruluğa ulaşmasını sağlamıştır.

Çalışmamızın kısıtları şunlardır:

1. Kullanılan veri seti yalnızca belirli bir sosyal medya platformu ile sınırlıdır ve dolayısıyla genelleştirilebilirliği başka platformlar için test edilmemiştir.
2. Çalışma sadece belirli makine öğrenmesi algoritmalarında test edilmiştir; derin öğrenme modelleri kapsam dışında bırakılmıştır.
3. Görsel analiz araçlarının sınırlamaları nedeniyle bazı görsellerin yanlış sınıflandırılması mümkündür.

Bu çalışma, X platformunda istenmeyen gönderi tespiti için makine öğrenmesi yöntemlerinin uygulanabilirliğini kanıtlamanın yanı sıra, bu alanda yeni ve etkili yöntemlerin geliştirilmesine de katkı sağlamaktadır.

Bu bölümde, X platformunda istenmeyen gönderi tespiti sorununu ve bu alandaki mevcut çözümleri inceledik. Ayrıca, bu çalışmanın metin ve görsel içerikleri birleştiren özgün yaklaşımını tanıttık. Bir sonraki bölümde, kullanılan veri seti ve yöntemlerin detaylarına odaklanacağız.

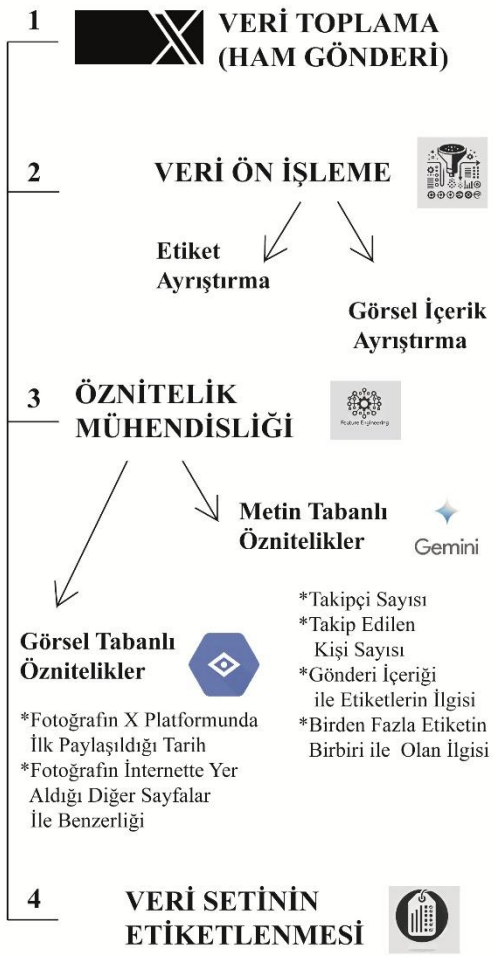
2. MATERYAL VE YÖNTEM (MATERIALS AND METHODS)

Bu bölümde, çalışmada kullanılan veri setinin oluşturulma süreci, özniteliklerin belirlenmesi ve bu özniteliklerin spam tespiti üzerindeki etkisi ele alınmaktadır. Fotoğraf içeren gönderilere odaklanılmasının nedenleri açıklanarak, bu tercihlerin analiz sürecine katkısı vurgulanmaktadır. Ayrıca, spam tespiti için kullanılan Karar Ağacı, Rastgele Orman, Destek Vektör Makineleri (SVM), Lojistik Regresyon ve Çok Katmanlı Algılayıcı (MLP) gibi makine öğrenmesi algoritmaları tanıtılmakta ve bu yöntemlerin veri setine nasıl uygulandığı detaylandırılmaktadır.

2.1. Üretilen Veri Seti

Twitter veri setlerine dayanan önceki çalışmalar, sosyal medya analizlerinde temel eksiklikler barındırmakta ve kapsamlı analizler için yetersiz kalmaktadır. Bu çalışmada kullanılan veri seti, 6 Şubat 2023 ile 17 Temmuz 2023 tarihleri arasında Türkiye'de trend olan 12 konuya ait 2481 gönderiden oluşmakta olup, tüm gönderiler fotoğraf içermektedir. Fotoğraf içeren gönderilere odaklanılmasının nedeni, istenmeyen içeriklerin çoğunlukla görsel medya üzerinden yayıldığının önceki çalışmalarda gösterilmesidir [28]. Ancak, önceki çalışmalar genellikle metin tabanlı analizlerle sınırlı kalmış ve görsel içeriklerin etkisi göz ardı edilmiştir. Bununla birlikte, görsel içerikler sosyal medya kullanıcı davranışları ve trend analizleri açısından önemli bilgiler sunmaktadır. Bu bağlamda, çalışmamızda fotoğraf içeren gönderilere yoğunlaşarak, spam içerikleri tespit etmek için yeni öznitelikler geliştirdik.

Öncelikle, takipçi sayısı, takip edilen kişi sayısı, gönderi içeriğinin etiketlerle ilişkisi, birden fazla etiketin birbirleriyle olan ilgisi gibi sosyal medya etkileşimini belirleyen önemli öznitelikler belirledik. Ayrıca, görsel içeriği analiz etmek için görselin X platformunda ilk paylaşıldığı tarih ve görselin internet üzerindeki diğer sayfalarda geçtiği metinle benzerliği gibi görsel odaklı öznitelikler de dahil ettik. Bu süreçte, X platformundan elde edilen ham verilerin (gönderilerin) özniteliklere dönüştürülmesi adımları, Şekil 1'de özetlenmektedir.



Şekil 1. Ham Veriden Öznitelik Çıkarımı Süreci
(Process of Feature Extraction from Raw Data)

Türkiye'ye özgü veri setleri üzerine yapılan çalışmalar ise sınırlıdır; Türkiye'deki yerel trendlerin, özellikle spam içerik gibi olgular üzerindeki etkisini analiz etmek için bu çalışmadaki gibi geniş kapsamlı bir veri setine ihtiyaç duyulmaktadır. Ayrıca, Google Gemini ve Cloud Vision AI gibi yeni teknolojiler, sosyal medya analizlerini daha derinlemesine inceleme fırsatı sunmasına rağmen, mevcut çalışmalarda bu araçlardan yeterince yararlanılmamıştır. Bu tür araçların entegrasyonu, daha zengin analizler yapılmasına olanak sağlayabilir. Son olarak, önceki çalışmalarda veri çeşitliliği sınırlı kalmış ve bu durum analizlerin kapsamını daraltmıştır. Çeşitli veri kaynaklarından elde edilen zenginleştirilmiş veri setleri, sosyal medya analizlerinde daha güvenilir ve ayrıntılı sonuçlar sunacaktır. Bu eksiklikler, mevcut X platformunda veri setlerinin neden yeterli olmadığını açıkça ortaya koymaktadır.

2.1.1.Veri Seti Üretiminde Kullanılan Yapay Zeka Araçları ve Alternatif Yöntemler

Bu çalışmada, metin ve görsel analizler için Google Gemini ve Google Cloud Vision AI araçları tercih edilmiştir. Bu araçlar, geniş dil ve görüntü işleme yetenekleri, API erişimi ve düşük maliyetli entegrasyon avantajları nedeniyle seçilmiştir. Ancak, Google Gemini bazı dillerde ve yerel konularda düşük doğruluk oranları sergileyebilmekte, ayrıca politikaları gereği bazı içerikleri filtreleyerek analiz dışı bırakabilmektedir. Google Cloud Vision AI ise görsel tanımada güçlü olmakla birlikte, sosyal medya spam tespiti için özel olarak tasarlanmadığından sınırlı bir performans göstermektedir.

Bu sınırlamaları aşmak için alternatif araçlar değerlendirilmiştir. OpenAI GPT gibi dil modelleri ve Microsoft Azure Computer Vision ile Amazon Rekognition gibi görsel analiz araçları potansiyel alternatifler olarak düşünülmüş, ancak çalışma kapsamına dahil edilmemiştir. Nihai tercih, Google araçlarının pratik avantajları ve yüksek doğruluk oranları nedeniyle bu yönde kullanılmıştır.

Sonuç olarak, çalışmada Google Gemini ve Cloud Vision AI'nın avantajlarından yararlanılmış, diğer araçlar literatür ışığında değerlendirilmiştir. Bu yaklaşım, spam içerik tespiti için kapsamlı bir analiz çerçevesi sunmuştur.

2.1.2.Takipçi Sayısı Özniteliği

Takipçi sayısı, X sosyal medya platformundaki hesapların etkileşim düzeyini ve popülerliğini anlamak için kullanılan önemli bir göstergedir. İstenmeyen gönderi paylaşan hesaplar, genellikle sahte veya bot hesaplar olduğundan, bu hesapların takipçi sayıları gerçek hesaplardan farklılık gösterebilir. Bu nedenle, takipçi sayısı, sahte hesapları tespit etmek için önemli bir öznitelik olarak kullanılmıştır. Bu veri, X API'si kullanılarak hesapların takipçi sayısına göre toplanmıştır.

2.1.3.Takip Edilen Kişi Sayısı Özniteliği

Takip edilen kişi sayısı, bir X sosyal medya hesabının etkileşim yapısını anlamak için önemli bir göstergedir. İstenmeyen gönderi paylaşan hesaplar genellikle rastgele veya çok sayıda kişiyi takip etme eğilimindedir. Bu durum, hesapları daha yaygın kullanıcı hesaplarından ayırabilir. Bu nedenle, takip edilen kişi sayısı, istenmeyen içerik paylaşan hesapları tespit etmede önemli bir gösterge sunar. Bu öznitelik, X API'si aracılığıyla elde edilmiştir.

2.1.4.Gönderi İçeriğinin Etiketler ile İlgisi Özniteliği

Gönderi içeriği ile etiketlerin ilişkisi, her bir gönderinin içeriği ile kullanılan etiketlerin uyumunu analiz ederek, hesapların sahte ya da bot olup olmadığına dair ipuçları sunar. İstenmeyen içerik yayan hesaplar, genellikle gönderilerini daha geniş bir kitleye ulaştırmak amacıyla popüler veya trend etiketleri kullanma eğilimindedir. Ancak, bu etiketlerin gönderi içeriğiyle uyumsuz veya alakasız olması, hesapları tespit etmek için belirleyici bir faktör olabilir.

Gönderi içeriği ile etiketlerin ilişkisinin analiz edilebilmesi için, gönderi metinleri ve etiketler ayrıştırılmış ve veriler

analiz edilebilir bir formata getirilmiştir. Bu adımda, her bir gönderi metni ve içerikteki etiketler ayrı sütunlara yerleştirilmiş, etiketlerin yoğunluğu analiz edilerek veri setinin yapısı hakkında bilgi edinilmiştir.

Gemini API, gönderinin içeriği ile ona atanan etiketlerin uyum derecesini belirlemek için kullanılmıştır. Bu API, gönderi içeriği ile ona atanan etiketlerin uyumunu 'Çok İlgisiz', 'İlgisiz', 'Nötr', 'İlgili' ve 'Çok İlgili' gibi derecelendirilmiş bir ölçekle sunar. Bu analiz, etiket ve içerik uyumunu belirleyerek, istenmeyen içerik paylaşan hesapları tespit etmeye yardımcı olur. Örneğin, tamamen alakasız etiketlere sahip bir gönderi "Çok İlgisiz" olarak sınıflandırılırken, konuya uygun etiketler taşıyan bir gönderi "Çok İlgili" olarak değerlendirilir.

Gemini API ayrıca, politikalarına aykırı içerikleri otomatik olarak filtreler ve şiddet, nefret söylemi veya cinsel içerik gibi unsurları engeller. Bu tür içeriklerin filtrelenmesi, veri setinin güvenilirliğini artırırken, engellenen içeriklerin yerine alternatif kategori etiketleri atanarak veri seti zenginleştirilmiştir.

2.1.5.Birden Fazla Etiketin Birbirleri ile Olan İlgisi Özniteliği

Etiketlerin birbirleriyle olan ilgisi özniteliği, X sosyal medya platformundaki gönderilerde yer alan etiketlerin içerik uyumunu analiz etmek amacıyla oluşturulmuştur. İstenmeyen gönderi paylaşan hesaplar genellikle çok sayıda alakasız veya popüler etiketi birlikte kullanarak dikkat çekmeye çalıştığından, etiketlerin birbiriyle olan ilgisi, bu tür hesapları tespit etmekte önemli bir göstergedir.

Bu öznitelik, Gemini API kullanılarak oluşturulmuştur. İlk olarak, her gönderideki etiket sayısı belirlenmiş, ardından etiketler ayrıştırılarak uyum analizine hazırlanmıştır. Gemini API, her gönderinin etiket uyumunu değerlendirirken 'Çok İlgisiz', 'İlgisiz', 'Nötr', 'İlgili' ve 'Çok İlgili' olmak üzere beş dereceden oluşan bir ölçek kullanmıştır. Bu ölçek, etiketlerin içerik ile ne kadar uyumlu olduğunu belirleyerek, gönderinin istenmeyen içerik olma olasılığını anlamaya yardımcı olmuştur.

2.1.6.Gönderi İçeriğinde Yer Alan Fotoğrafın X Platformu Üzerinde Paylaşılma Durumu Özniteliği

Gönderi içeriğinde yer alan fotoğrafın X üzerindeki ilk paylaşıldığı tarih, istenmeyen gönderi paylaşan hesapları tespit etmek için kullanılan önemli bir özniteliktir. Bu öznitelik, bir fotoğrafın başka bir kullanıcı tarafından daha önce paylaşılmış olup olmadığını ve yeniden paylaşılma oranını belirler. İstenmeyen hesaplar genellikle özgün içerik üretmek yerine mevcut içerikleri yeniden paylaşır, bu da fotoğrafın eski tarihlerde paylaşıldığını gösterir.

Veri setindeki fotoğraflar, Google Cloud Vision API aracılığıyla taranarak, her bir fotoğrafın X üzerindeki ilk paylaşıldığı tarih belirlenmiştir. İlk olarak, fotoğraf içeren gönderilerden URL'ler toplanmış ve ardından fotoğrafların önceki paylaşımlarına dair bilgiler elde edilmiştir. Eğer bir fotoğraf daha önce başka bir hesap tarafından paylaşılmışsa, bu durum fotoğrafın istenmeyen içerik olma olasılığını artırır. Son olarak, fotoğrafın paylaşım tarihi eskiyse 'Kaynak Kendisi-1', yeni ise 'Kaynak Başkası-2' etiketi verilmiştir.

Çalışmada kullanılan veri seti, 'Gönderi İçeriğinde Yer Alan Fotoğrafın X Platformu Üzerinde Paylaşılma Durumu' özniteliğinin Google Cloud Vision API ile oluşturulma durumuna göre ikiye ayrılmıştır: 293 gönderiden oluşan Veri Seti-I ve 2188 gönderiden oluşan Veri Seti-II.

Bir gönderideki görsel, X platformunda daha önce başka bir hesap tarafından paylaşılmamışsa, bu görselin kaynağı doğrudan "kendisi" olarak kabul edilir. Bu durum, Veri Seti-II'deki tüm görseller için geçerli olduğundan, 'Kaynak Kendisi-1' özniteliğinin eklenmesine gerek kalmamış ve bu öznitelik Veri Seti-II'den çıkarılmıştır.

2.1.7.Gönderi İçeriğinde Yer Alan Fotoğrafın İnternette Yer Alan Diğer Sayfalarda Birlikte Geçtiği Metin Arasındaki Benzerlik Özniteliği

Gönderi içeriğinde yer alan fotoğrafın internet üzerindeki diğer sayfalarda geçtiği metinle benzerliği özniteliği, X sosyal medya platformundaki içeriklerin anlamını daha iyi anlamak amacıyla kullanılmıştır. Bu öznitelik, bir fotoğrafın internet üzerindeki başka sayfalarda hangi metinlerle birlikte yer aldığını inceleyerek, gönderinin içeriği hakkında önemli bilgiler sağlar. Örneğin, bir fotoğrafın haber sitesinde aynı başlıkla yer alması, gönderinin haberle ilgili olduğunu düşündürebilir.

Bu öznitelik, Google Cloud Vision API'nin "Pages with matching images" özelliği ile oluşturulmuştur. API, görsellerin internet üzerindeki diğer sayfalarda hangi başlıklarla eşleştiğini belirler ve bu başlıkların metin içerikleriyle gönderi metnini Gemini API aracılığıyla karşılaştırır. Bu süreç, gönderi içeriği ve görsellerinin anlamını daha derinlemesine incelemeye olanak tanır. İlk aşamada, Vision API ile görsellerin eşleştiği sayfalardan "pageTitle" bilgisi toplanmış ve bu başlıkların metin içerikleri, Gemini API kullanılarak gönderi metniyle karşılaştırılmıştır.

Ardından, gönderi içeriği ile etiketlerin uyumunu belirlemek için kullanılan ölçek, 'Hiç Benzer Olmayan', 'Benzer Olmayan', 'Nötr', 'Benzer' ve 'Çok Benzer' olmak üzere beş dereceden oluşmaktadır. Bu dereceler, içerik ile etiketlerin uyumunu sırasıyla tanımlar: 'Hiç Benzer Olmayan' en düşük uyum seviyesini, 'Çok Benzer' ise en yüksek uyum seviyesini ifade eder.

2.1.8. Öznitelik Önem Analizi

Bu çalışmada kullanılan özniteliklerin spam tespitindeki önemini değerlendirmek amacıyla, Mutual Information

(MI) yöntemi kullanılarak özniteliklerin hedef değişken (spam veya spam değil) ile olan ilişkisi analiz edilmiştir. Mutual Information skoru, bir özniteliğin bağımsız değişken olarak hedef değişkeni (spam olup olmama durumu) ne kadar açıkladığını göstermektedir. Veri Seti-I için elde edilen Mutual Information skorları Tablo 1'de sunulmaktadır.

Tablo 1. Veri Seti-I için Mutual Information Skorları
(Mutual Information Scores for Dataset-II)

Öznitelik Adı	MI Skoru
Takipçi Sayısı	0.0234
Takip Edilen Kişi Sayısı	0.0146
Gönderi İçeriğinin Etiketler ile İlgisi	0.0591
Birden Fazla Etiketin Birbirleri ile Olan İlgisi	0.0569
Gönderi İçeriğinde Yer Alan Fotoğrafın X Platformu Üzerinde Paylaşılma Durumu	0.0634
Gönderi İçeriğinde Yer Alan Fotoğrafın İnternette Yer Alan Diğer Sayfalarda Birlikte Geçtiği Metin Arasındaki Benzerlik	0.0000

Veri Seti-I için yapılan MI analizi, modelin hangi değişkenlerden daha fazla bilgi edindiğini belirlemek açısından önemli bir katkı sunmuştur. Sonuçlara göre, "Gönderi İçeriğinde Yer Alan Fotoğrafın X Platformu Üzerinde Paylaşılma Durumu" (0.0634) en yüksek MI skoruna sahip olup, modelin tahmin gücüne en fazla katkıyı sağlamaktadır. Bunu "Gönderi İçeriğinin Etiketler ile İlgisi" (0.0591) ve "Birden Fazla Etiketin Birbirleri ile Olan İlgisi" (0.0569) değişkenleri takip etmektedir. Bu bulgular, içerik ve etiketler arasındaki ilişkinin modelin karar alma sürecinde kritik bir rol oynadığını göstermektedir. "Takipçi Sayısı" (0.0234) ve "Takip Edilen Kişi Sayısı" (0.0146) ise daha düşük bilgi katkısına sahip öznitelikler olarak belirlenmiştir.

Öte yandan, "Gönderi İçeriğinde Yer Alan Fotoğrafın İnternette Yer Alan Diğer Sayfalarda Birlikte Geçtiği Metin Arasındaki Benzerlik" (0.0000) değişkeninin model açısından anlamlı bir bilgi içermediği tespit edilmiştir. Bu nedenle, Veri Seti-I'den çıkarılarak modelin daha verimli hale getirilmesi sağlanmıştır. Bu tür düşük katkılı değişkenlerin elimine edilmesi, modelin sadece anlamlı değişkenlere odaklanmasını ve gereksiz hesaplama yükünü azaltmasını sağlamaktadır.

Veri Seti-II'ye yönelik yapılmış olan Mutual Information skorları ise Tablo 2'de gösterilmektedir.

Tablo 2. Veri Seti-II için Mutual Information Skorları
(Mutual Information Scores for Dataset-II)

Öznitelik Adı	MI Skoru
Takipçi Sayısı	0.0087
Takip Edilen Kişi Sayısı	0.0000
Gönderi İçeriğinin Etiketler ile İlgisi	0.0838
Birden Fazla Etiketin Birbirleri ile Olan İlgisi	0.0907
Gönderi İçeriğinde Yer Alan Fotoğrafın İnternette Yer Alan Diğer Sayfalarda Birlikte Geçtiği Metin Arasındaki Benzerlik	0.0179

Veri Seti-II için yapılan analizlerde, en yüksek MI skoruna sahip olan "Birden Fazla Etiketin Birbirleri ile Olan İlgisi" (0.0907) ve "Gönderi İçeriğinin Etiketler ile İlgisi" (0.0838) öznitelikleri, modelin öğrenme sürecinde en belirleyici faktörler olarak öne çıkmaktadır. Bu durum, gönderi içeriğinin etiketlerle ve etiketlerin birbirleriyle olan ilişkilerinin, modelin tahmin gücünü artırdığını göstermektedir. Buna karşılık, "Takipçi Sayısı" (0.0087) ve "Gönderi İçeriğinde Yer Alan Fotoğrafın İnternette Yer Alan Diğer Sayfalarda Birlikte Geçtiği Metin Arasındaki Benzerlik" (0.0179) gibi değişkenlerin etkisi daha sınırlı kalmıştır.

Özellikle "Takip Edilen Kişi Sayısı" (MI = 0.0000) özniteliğinin model için hiçbir bilgi sağlamadığı görülmüş ve Veri Seti-II'den çıkarılmıştır. Bu durum, ilgili değişkenin hedef değişken ile anlamlı bir ilişkiye sahip olmadığını ve modelin tahmin gücüne herhangi bir katkı sunmadığını ortaya koymaktadır.

2.1.9.Veri setinin Etiketlenmesi

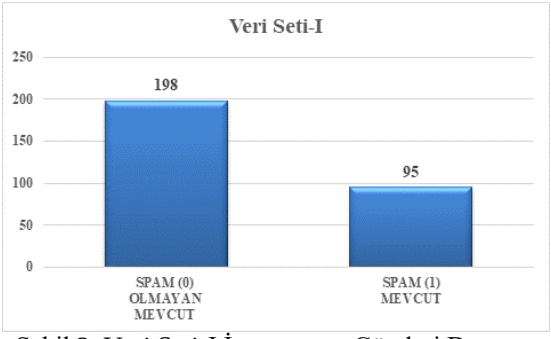
Veri seti, 'Gönderi İçeriğinde Yer Alan Fotoğrafın X Platformu Üzerinde Paylaşılma Durumu' özniteliğinin Google Cloud Vision API ile oluşturulma durumuna göre ikiye ayrılmıştır: 293 gönderiden oluşan Veri Seti-I ve 2188 gönderiden oluşan Veri Seti-II. Bu işlem sonucunda, Veri Seti-I'de 5 öznitelik, Veri Seti-II'de ise 4 öznitelik yer almaktadır.

Veri setinin etiketlenmesi sürecinde, gönderilerin istenmeyen gönderi olup olmadığını belirlemek için detaylı bir inceleme yapılmıştır. Bu inceleme, Cumhurbaşkanlığı Dijital Dönüşüm Ofisi'nin Sosyal Medya Kullanım Kılavuzu'ndaki 'spamcılar' tanımına dayandırılmıştır. Kılavuzda, spamcılar; *'Her türlü içeriğe ilgili ilgisiz yorum yapan, kendi sayfasını veya profilini ziyarete davet eden, takipçi kazanmak için takip eden, genellikle sadece beğeni ve repost yapıp içerik üretmeyen, paylaşımlarının altına konuyla tamamen ilgisiz iletiler yazan ve gündem hashtag'lerini kullanarak, bu hashtag'lerle alakasız paylaşımlar yapan'* kullanıcılar olarak tanımlanmaktadır [29]. Bu tanıma göre, hashtag içerikleriyle alakasız paylaşımlar istenmeyen gönderi olarak değerlendirilmiş ve bu tür paylaşımlar istenmeyen gönderi olarak sınıflandırılmıştır.

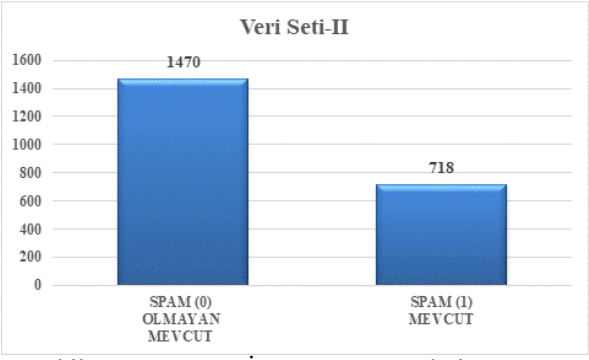
İkinci aşamada, veri setindeki tüm gönderiler detaylı bir şekilde incelenmiş ve her biri kılavuz kriterlerine göre etiketlenmiştir. Bu süreçte, gönderilerin içerdiği metinler ve fotoğraflar analiz edilmiş, özellikle hashtaglerin gönderi içeriğiyle uyumuna dikkat edilmiştir. Gönderinin hashtag ile doğrudan alakalı olmaması durumunda, gönderi istenmeyen olarak etiketlenmiştir.

Son olarak, etiketlenen veriler, makine öğrenmesi modellerinin eğitim ve test aşamalarında kullanılmak üzere düzenlenmiştir. İstenmeyen gönderiler (1) ve istenmeyen olmayan gönderiler (0) şeklinde iki ana kategoriye ayrılmıştır. Bu düzenleme, makine öğrenmesi modellerinin

istenmeyen gönderi tespitinde daha doğru sonuçlar elde etmesine yardımcı olmuştur. Veri Seti-I ve Veri Seti-II'ye yönelik spam ve spam olmayan gönderilerin dağılımı, Şekil 2 ve Şekil 3'te gösterilmektedir. Ayrıca, her iki veri setine ilişkin özet bilgiler Tablo 3'te sunulmaktadır.



Şekil 2. Veri Seti-I İstenmeyen Gönderi Durumu
(Data Set-I Spam Status)



Şekil 3. Veri Seti-II İstenmeyen Gönderi Durumu
(Data Set-II Spam Status)

Tablo 3. Veri Setleri Özet Durum
(Summary Status of Datasets)

Veri Seti Adı	Öznitelik Sayısı	İstenmeyen Gönderi (1)	İstenmeyen Olmayan Gönderi (0)	Genel Toplam
Veri Seti-I	5	95	198	293
Veri Seti-II	4	718	1470	2188

2.2. Makine Öğrenmesi Yöntemleri

Bu çalışmada Karar Ağacı, Rastgele Orman, Destek Vektör Makineleri (SVM - Support Vector Machine), Lojistik Regresyon ve Çok Katmanlı Algılayıcı (MLP) algoritmaları kullanılmıştır.

2.2.1.Karar Ağacı Algoritması

Karar Ağacı algoritması, makine öğrenmesinde sıkça kullanılan ve açıklanabilirliği yüksek olan bir yöntemdir. Verileri belirli özelliklere göre dallara ayırarak karar verme sürecini basitleştiren bu algoritma, sınıflandırma ve regresyon gibi çeşitli problemlerde yaygın olarak kullanılmaktadır. Karar ağaçları, veri setlerindeki en üst (kök) düğümden başlayarak her düğümde belirli bir özelliğe göre karar alır ve yaprak düğümlere ulaşana kadar dallanır. Yaprak düğümler, nihai bir sınıf veya değerin tahminini temsil eder. Açıklanabilirliği yüksek olan bu algoritma, özellikle sınıflandırma problemlerinde basit ve etkili bir çözüm sunar. Bu çalışmada, özellikle metin ve görsel özniteliklerin etkileşimini anlamak için tercih edilmiştir.

2.2.2.Rastgele Orman Algoritması

Rastgele Orman algoritması, makine öğrenmesinde yaygın olarak kullanılan bir ensemble (topluluk) sınıflandırıcıdır [30]. Her bir ağaç, veri kümesinin rastgele alt örnekleri ve özellikleri üzerine odaklanır. Tahmin sonuçları bir araya getirilerek, ensemble ortalaması alınır ve bu yöntemle daha doğru ve tutarlı bir sınıflandırma sağlanır [31]. Rastgele Orman algoritması, aşırı uyuma karşı dayanıklıdır ve diğer modellerden daha az hiperparametre ayarı gerektirir. Büyük ve karmaşık veri setlerinde etkilidir, bu nedenle tıbbi teşhislerden finansal tahminlere kadar çeşitli alanlarda başarıyla kullanılmaktadır. Gürültüyü azaltarak model performansını artırdığı için yüksek doğruluk gerektiren veri setlerinde sıklıkla tercih edilir. Bu nedenle, dengesiz veri setlerinde daha kararlı sonuçlar elde etmek için çalışmamızda tercih edilmiştir.

2.2.3.SVM Algoritması

SVM algoritması, makine öğrenmesinde özellikle sınıflandırma problemlerinde sıkça kullanılan önemli bir algoritmadır. SVM, veri noktalarını bir hiper düzlem ile ayırarak en iyi sınıflandırmayı sağlamak amacıyla destek vektörlerinden yararlanır. Sürekli özelliklerin olduğu veri setlerinde etkili olup, sınıflandırma sırasında veri noktalarını ayırmak için özel bir optimizasyon yöntemi kullanır. SVM, metin sınıflandırma, görüntü tanıma, biyomedikal veri analizi ve finansal analiz gibi alanlarda yaygın olarak kullanılmaktadır. Esnekliği ve yüksek performansı, bu algoritmayı geniş bir uygulama alanına sahip kılar. Özellikle yüksek boyutlu veri setlerinde etkili olan SVM, metin ve görsel özniteliklerin birleşik analizinde iyi bir performans sergilemesi beklenerek, bu çalışmada tercih edilmiştir.

2.2.4.Lojistik Regresyon Algoritması

Lojistik regresyon, ikili sınıflandırma problemlerinde yaygın olarak kullanılan güçlü ve açıklanabilir bir makine öğrenmesi algoritmasıdır. Bu yöntem, bağımsız değişkenlerle bağımlı değişken arasında doğrusal bir ilişki kurarak sonuçları 0 ile 1 arasında bir olasılık değeri olarak ifade eder. Model, bir veri noktasının özelliklerinin ağırlıklı toplamını hesaplayarak bu toplamı sigmoid fonksiyonuna uygular; böylece, pozitif sınıfa ait olma olasılığını tahmin eder. Lojistik regresyonun avantajlarından biri, katsayıların da yorumlanabilir olmasıdır; her katsayı, bağımsız değişkenin bağımlı değişken üzerindeki etkisi hakkında bilgi verir. Ancak, doğrusal olarak ayrılabilir olmayan veri setlerinde performansı sınırlıdır. Basit ve etkili bir algoritma olarak

lojistik regresyon, birçok alanda kullanılabilecek şekilde tasarlanmıştır. Bu çalışmada özellikle metin tabanlı özniteliklerin analizinde etkili olması beklendiği için tercih edilmiştir.

2.2.5.Çok Katmanlı Algılayıcı Algoritması

Çok Katmanlı Algılayıcı (MLP), makine öğrenmesinde kullanılan bir yapay sinir ağı türüdür ve karmaşık sınıflandırma problemlerini çözmekte oldukça etkilidir. MLP, girdi katmanı, bir veya daha fazla gizli katman ve çıktı katmanından oluşur. Girdi katmanındaki her nöron bağımsız değişkenleri temsil ederken, bu değerler gizli katmanlardaki nöronlara iletilir. Gizli katmanlar, girdileri işleyip aktivasyon fonksiyonlarını uygulayarak daha yüksek seviyede özellikler çıkarır. Son olarak, çıktı katmanı, sınıflandırma veya regresyon problemlerine yönelik nihai tahminleri üretir. Metin ve görsel özniteliklerin birlikte analiz edilmesi sürecinde daha derin öğrenme yetenekleri sunması nedeniyle bu çalışmada tercih edilmiştir.

2.3. Makine Öğrenmesi Yöntemlerinin Performans Değerlendirme Kriterleri

Bir algoritmanın başarısını değerlendirilebilmesi için eğitim aşamasından sonra, sınıflandırma modelinin tahmin doğruluğunun ölçülmesi gerekmektedir. Bu çalışmada, hata matrisi oluşturularak, doğruluk, duyarlılık, kesinlik ve F1 skoru gibi kriterler ile sınıflandırma modellerinin performansları karşılaştırılmaktadır.

2.3.1.Hata Matrisi

Hata matrisi, makine öğrenmesi ve özellikle sınıflandırma problemlerinde model performansını değerlendirmek için önemli bir araç olarak kullanılmaktadır. Bir hata matrisi, modelin doğru ve yanlış tahminlerini tablo şeklinde özetlemektedir ve bu sayede modelin başarısı hakkında bilgi sağlamaktadır. Hata matrisi, dört temel bileşenden oluşarak Tablo 4'te gösterilmektedir.

Tablo 4. Hata Matrisi
(Confusion Matrix)

Prensip	Tahmin Edilen (Pozitif)	Tahmin Edilen (Negatif)
Gerçek Pozitif	Doğru Pozitif (TP)	Yanlış Negatif (FN)
Gerçek Negatif	Yanlış Pozitif (FP)	Doğru Negatif (TN)

2.3.2.Doğruluk

Doğruluk, bir sınıflandırma modelinin tüm tahminlerinin ne kadar doğru olduğunu ölçen temel bir performans metriği olmaktadır. Doğruluk, doğru tahminlerin (hem TP hem de TN) toplam tahminlere oranı olarak hesaplanmaktadır. Yüksek doğruluk, modelin genel olarak iyi performans gösterdiğini belirtmektedir. Bu noktada doğruluk, diğer metriklerle birlikte desteklenerek değerlendirilmektedir. Doğruluk ölçütünün denklemi Eşitlik (1)'de ifade edilmiştir:

$$Doğruluk = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

2.3.3.Duyarlılık

Duyarlılık, aynı zamanda "hassasiyet" veya "true positive rate" olarak da isimlendirilir ve modelin, pozitif sınıfları ne kadar iyi tespit ettiğini ölçmektedir. Yüksek duyarlılık, modelin yanlış negatifleri (FN) minimize ettiğini göstermektedir. Duyarlılık ölçütünün denklemi Eşitlik (2)'de ifade edilmiştir:

$$Duyarlılık = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \quad (2)$$

2.3.4.Kesinlik

Kesinlik, modelin pozitif tahminlerinin ne kadar doğru olduğunu göstermektedir. Yani, modelin pozitif olarak sınıflandırdığı örneklerin gerçekten pozitif olma oranını ifade etmektedir. Yüksek kesinlik, modelin yanlış pozitifleri (FP) minimize ettiğini ifade etmektedir ve özellikle istenmeyen gönderi filtreleme veya hastalık teşhisi gibi yanlış pozitiflerin maliyetli olduğu durumlarda kullanılmaktadır. Kesinlik ölçütünün denklemi Eşitlik (3)'te ifade edilmiştir:

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

2.3.5.F1 Skor

F1 Skor, kesinlik ve duyarlılığı tek bir metrikte birleştirerek modelin genel performansını özetlemeyi hedeflemektedir. F1 Skor, özellikle sınıflar arasında dengesizlik olduğunda faydalı olmakla birlikte kesinlik ve duyarlılık arasındaki dengeyi sağlamaktadır. F1 Skor ölçütünün denklemi Eşitlik (4)'te ifade edilmiştir:

$$F1\ Skor = 2 \times \frac{TP}{2\times TP+FP+FN} \quad (4)$$

2.3.6.ROC AUC (ROC Eğrisi Altındaki Alan)

ROC AUC (Receiver Operating Characteristic Area Under the Curve), bir modelin sınıflandırma performansını değerlendirmek için kullanılan önemli bir metrik olmaktadır. ROC Eğrisi, modelin farklı eşik değerlerinde duyarlılık (true positive rate) ve yanlış pozitif oranını (false positive rate) göstermektedir. ROC Eğrisi altındaki alan (AUC), modelin genel performansını özetlemekle birlikte 0.5 ile 1 arasında bir değer almaktadır. 0.5 değeri, modelin rastgele tahmin yaptığını gösterirken, 1 değeri mükemmel ayırma kapasitesini belirtmektedir.

Bu bölümde, veri setinin oluşturulma süreci, kullanılan makine öğrenmesi yöntemleri ve performans değerlendirme kriterleri açıklanmıştır. Elde edilen veri seti ve uygulanan algoritmalar, spam tespiti ve sosyal medya analizlerinde önemli bir katkı sağlamaktadır. Bir sonraki bölümde, bu yöntemlerle elde edilen sonuçlar ve modellerin karşılaştırmalı analizi sunulacaktır.

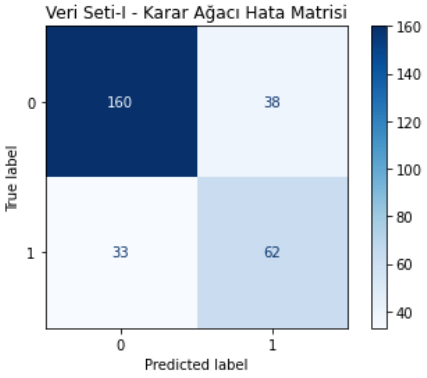
3. DENEYSEL SONUÇLAR (EXPERIMENTAL RESULTS)

Bu bölümde, makine öğrenmesi algoritmalarının performansını değerlendirmek amacıyla oluşturulan Veri Seti-I ve Veri Seti-II üzerinde gerçekleştirilen deneysel analizler sunulmaktadır.

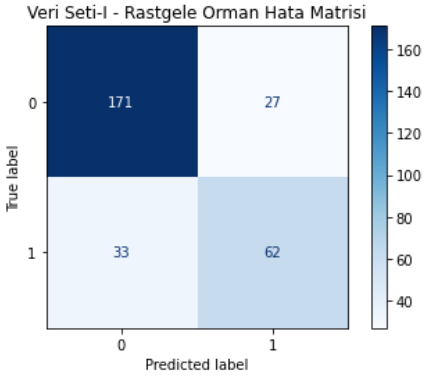
Veri seti, "Gönderi İçeriğinde Yer Alan Fotoğrafın X Platformu Üzerinde Paylaşılma Durumu" özniteliğinin Google Cloud Vision API ile oluşturulma durumuna göre ikiye ayrıldığından, bu bölümde Veri Seti-I ve Veri Seti-II olarak ayrı ayrı incelenmiştir.

3.1. Veri Seti-I

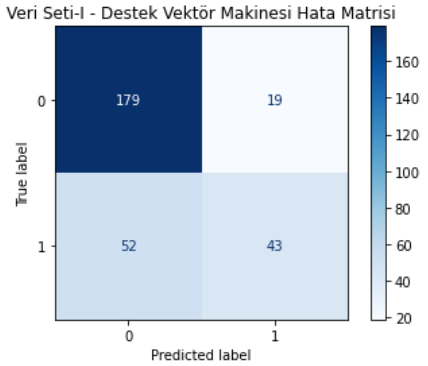
Makine öğrenmesi algoritmaları kullanılarak oluşturulan modelin test tahminleri, test verileri ile karşılaştırıldığında, Veri Seti-I için Şekil 4-8 aralığında gösterilmekte olan hata matrisi elde edilmektedir.



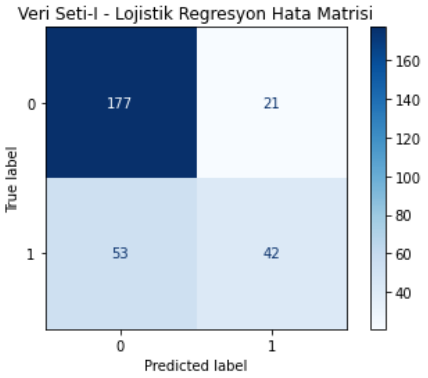
Şekil 4. Veri Seti-I Karar Ağacı Hata Matrisi
(Dataset-I Decision Tree Confusion Matrix)



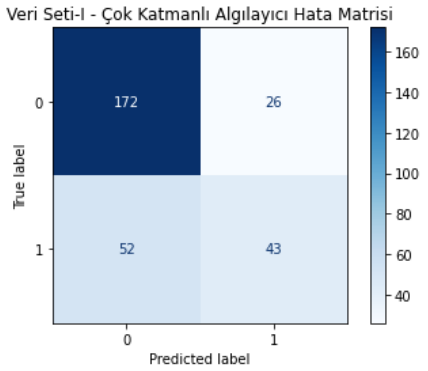
Şekil 5. Veri Seti-I Rastgele Orman Hata Matrisi
(Dataset-I Random Forest Confusion Matrix)



Şekil 6. Veri Seti-I SVM Hata Matrisi
(Dataset-I SVM Confusion Matrix)



Şekil 7. Veri Seti-I Lojistik Regresyon Hata Matrisi
(Dataset-I Logistic Regression Confusion Matrix)

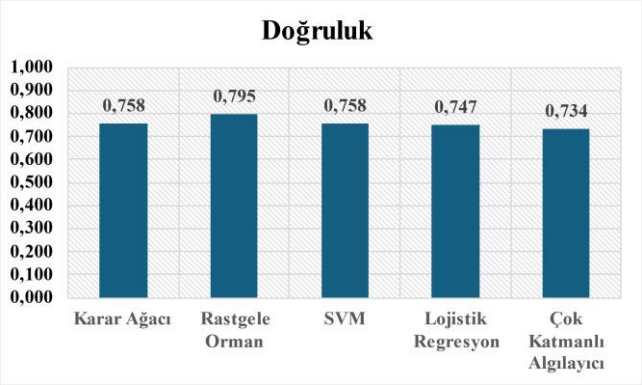


Şekil 8. Veri Seti-I Çok Katmanlı Algılayıcı Hata Matrisi
(Dataset-I MLP Confusion Matrix)

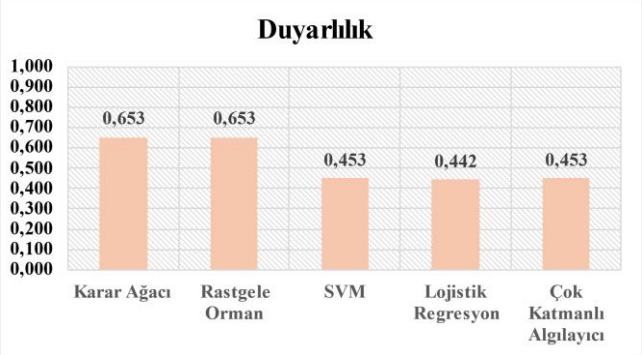
Veri Seti-I'e yönelik performans metrikleri Tablo 5'te ve Şekil 9-13 aralığında gösterilmektedir.

Tablo 5. Veri Seti-I Performans Metrikleri
(Dataset-I Performance Metrics)

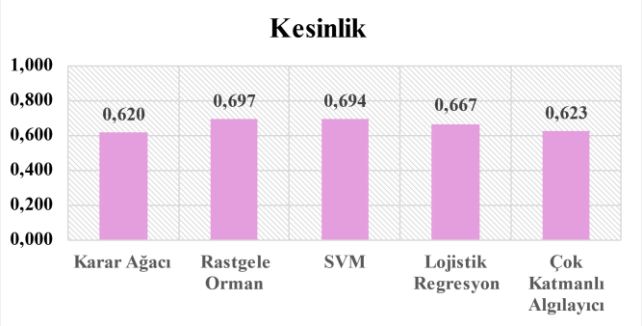
Algoritma	Doğruluk	Duyarlılık	Kesinlik	F1 Skor	ROC AUC
Karar Ağacı	0,758	0,653	0,620	0,632	0,727
Rastgele Orman	0,795	0,653	0,697	0,667	0,838
SVM	0,758	0,453	0,694	0,536	0,692
Lojistik Regresyon	0,747	0,442	0,667	0,526	0,772
MLP	0,734	0,453	0,623	0,519	0,754



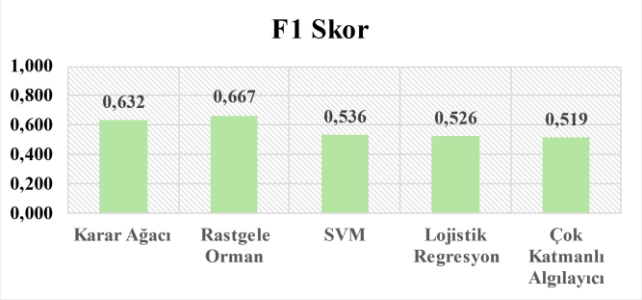
Şekil 9. Veri Seti-I Doğruluk Metrikleri
(Dataset-I Accuracy Metrics)



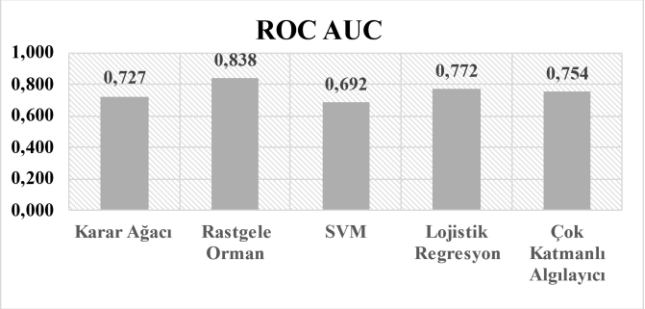
Şekil 10. Veri Seti-I Duyarlılık Metrikleri
(Dataset-I Recall Metrics)



Şekil 11. Veri Seti-I Kesinlik Metrikleri
(Dataset-I Precision Metrics)



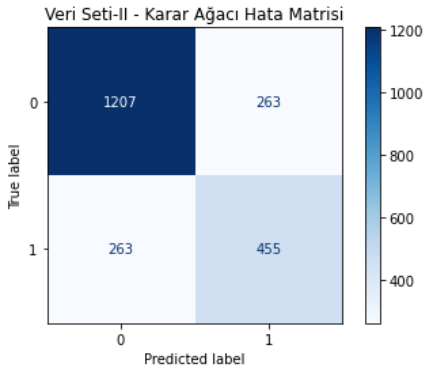
Şekil 12. Veri Seti-I F1 Skor Metrikleri
(Dataset-I F1 Score Metrics)



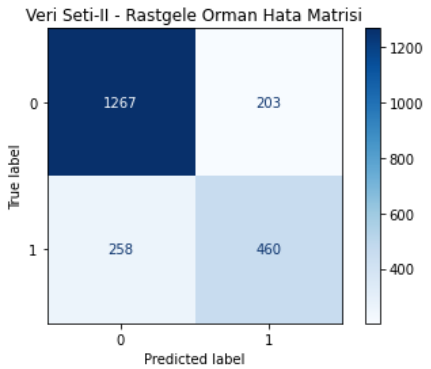
Şekil 13. Veri Seti-I ROC AUC Metrikleri
(Dataset-I ROC AUC Metrics)

3.2. Veri Seti-II

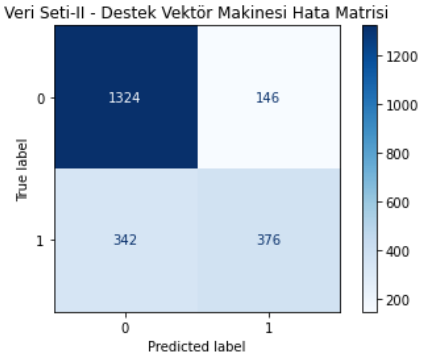
Makine öğrenmesi algoritmaları kullanılarak oluşturulan modelin test tahminleri, test verileri ile karşılaştırıldığında, Veri Seti-II için Şekil 14-18 aralığında gösterilmekte olan hata matrisi elde edilmektedir.



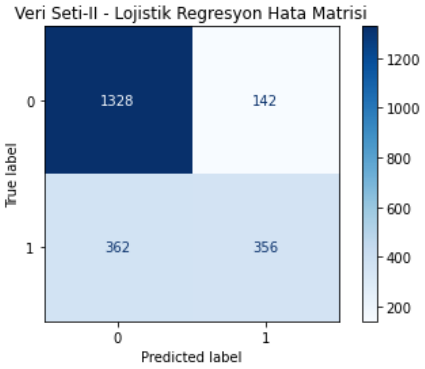
Şekil 14. Veri Seti-II Karar Ağacı Hata Matrisi
(Dataset-II Decision Tree Confusion Matrix)



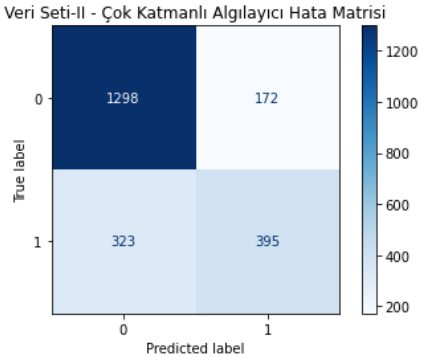
Şekil 15. Veri Seti-II Rastgele Orman Hata Matrisi
(Dataset-II Random Forest Confusion Matrix)



Şekil 16. Veri Seti-II SVM Hata Matrisi
(Dataset-II SVM Confusion Matrix)



Şekil 17. Veri Seti-II Lojistik Regresyon Hata Matrisi
(Dataset-II Logistic Regression Confusion Matrix)

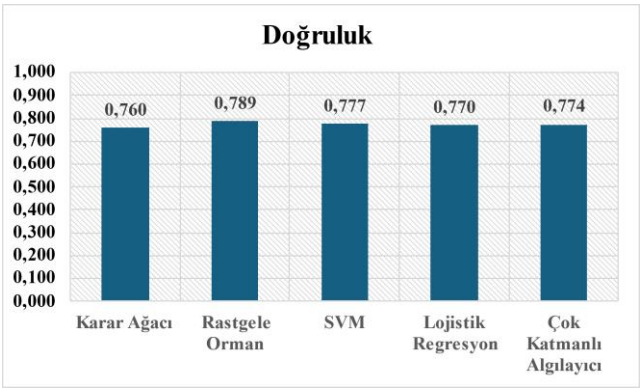


Şekil 18. Veri Seti-II Çok Katmanlı Algılayıcı Hata Matrisi
(Dataset-II MLP Confusion Matrix)

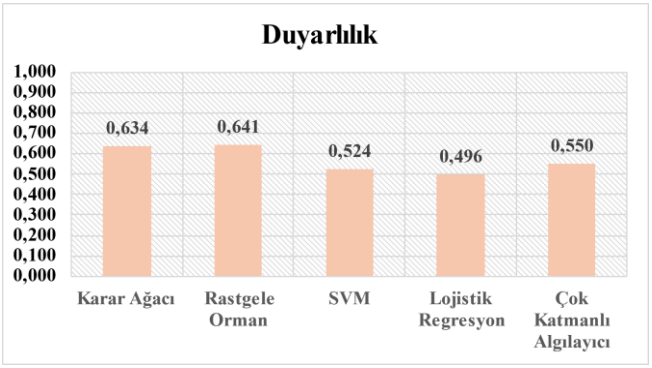
Veri Seti-II'e yönelik performans metrikleri Tablo 6'da ve Şekil 19-23 aralığında gösterilmektedir.

Tablo 6. Veri Seti-II Performans Metrikleri
(Dataset-II Performance Metrics)

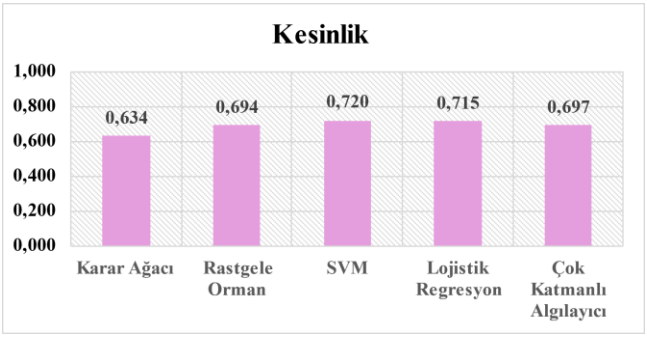
Algoritma	Doğruluk	Duyarlılık	Kesinlik	F1 Skor	ROC AUC
Karar Ağacı	0,760	0,634	0,634	0,635	0,728
Rastgele Orman	0,789	0,641	0,694	0,667	0,842
SVM	0,777	0,524	0,720	0,606	0,757
Lojistik Regresyon	0,770	0,496	0,715	0,585	0,763
MLP	0,774	0,550	0,697	0,615	0,802



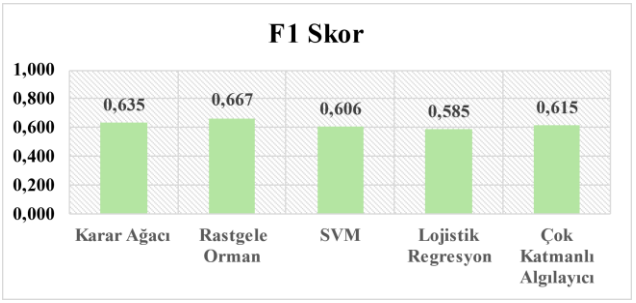
Şekil 19. Veri Seti-II Doğruluk Metrikleri
(Dataset-II Accuracy Metrics)



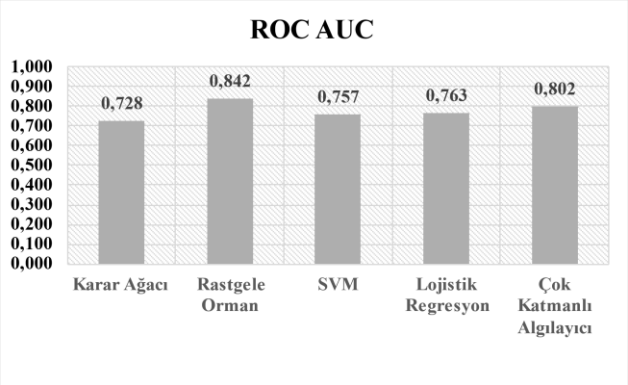
Şekil 20. Veri Seti-II Duyarlılık Metrikleri
(Dataset-II Recall Metrics)



Şekil 21. Veri Seti-II Kesinlik Metrikleri
(Dataset-II Precision Metrics)



Şekil 22. Veri Seti-II F1 Skor Metrikleri
(Dataset-II F1 Score Metrics)



Şekil 23. Veri Seti-II ROC AUC Metrikleri
(Dataset-II ROC AUC Metrics)

3.3. Algoritmaların Veri Setleri Üzerindeki Performans Farklılıkları ve Nedenleri

Bu çalışmada kullanılan iki farklı veri seti, içerik ve büyüklük açısından farklılık göstermektedir. Veri Seti-I (293 gönderi) daha küçük ve dengesiz bir yapıya sahipken, Veri Seti-II (2188 gönderi) daha büyük ve dengesiz bir dağılıma sahiptir. Yapılan analizler, makine öğrenmesi algoritmalarının bu farklı veri setleri üzerinde farklı performans gösterdiğini ortaya koymuştur.

Özellikle Rastgele Orman algoritması, her iki veri setinde de en yüksek genel performansa ulaşmış, doğruluk, duyarlılık ve F1 skoru açısından en iyi sonuçları elde etmiştir. Bunun temel nedenleri şunlar olabilir:

- Veri Seti-II'nin daha büyük olması, bazı algoritmaların (SVM, Lojistik Regresyon, MLP) daha iyi genelleme yapmasını sağlamış, ancak Rastgele Orman ve Karar Ağacı için belirgin bir fark yaratmamıştır.
- Veri dengesizliği, özellikle Veri Seti-I üzerinde SVM ve Lojistik Regresyon'un performansını olumsuz etkilemiştir. Veri Seti-II'de bu algoritmaların F1 skorlarında iyileşme görülmüştür.
- Karar Ağaçları ve Rastgele Orman gibi ağaç tabanlı yöntemler, veri dengesizliği ile daha iyi başa çıkabilmiştir. Ağaç tabanlı yöntemler, her sınıfa özel ayrım noktaları belirleyerek dengesiz veri setlerinde daha kararlı sonuçlar üretebilir.
- Veri Seti-I'de F1 skorlarının düşük olması, azınlık sınıfının (spam) yeterince iyi öğrenilememesinden kaynaklanmıştır. Bu durum, Veri Seti-I için SMOTE (Synthetic Minority Over-sampling Technique) gibi veri dengeleme yöntemlerinin kullanılmasını gerektirebilir.

Bu bölümde elde edilen sonuçlar doğrultusunda, farklı makine öğrenmesi algoritmalarının spam tespiti üzerindeki performansları karşılaştırılmıştır. Özellikle Destek Vektör Makineleri, Lojistik Regresyon ve Çok Katmanlı Algılayıcı (MLP), Veri Seti-II üzerinde daha yüksek F1 skoru ve duyarlılık göstermiştir. Rastgele Orman ve Karar Ağacı algoritmalarında ise veri setleri arasında belirgin bir üstünlük gözlenmemiştir. Bir sonraki bölümde, elde edilen bulguların genel bir değerlendirmesi yapılarak sonuçlar tartışılacaktır.

4. SONUÇ VE TARTIŞMA (RESULTS AND DISCUSSION)

Bu çalışmada, X sosyal medya platformundaki istenmeyen gönderilerin tespitine yönelik makine öğrenmesi, geniş dil modelleri ve bilgisayarlı görü tekniklerini birleştiren yeni bir yaklaşım geliştirilmiştir. Yapılan analizlerde, dengesiz yapıya sahip iki farklı veri seti (Veri Seti-I ve Veri Seti-II) üzerinde çeşitli makine öğrenmesi algoritmalarının performansları değerlendirilmiştir. Veri Seti-I 293 örnekten (95 istenmeyen gönderi, 198 istenmeyen olmayan gönderi) oluşurken, Veri Seti-II 2188 örnekten (718 istenmeyen gönderi, 1470 istenmeyen olmayan gönderi) oluşmaktadır. Performans değerlendirmelerinde, özellikle spam tespiti gibi dengesiz veri setlerinde, yalnızca doğruluğa (accuracy) odaklanmak yanıltıcı olabileceğinden, sınıflar arası dengesizliklere duyarlı olan F1 skoru, ROC AUC, duyarlılık (recall) ve kesinlik (precision) gibi kapsamlı ölçütler tercih edilmiştir. Bu yaklaşım, sınıflar arasındaki dengesizliklerin yanlış yorumlanmasını engellemeyi ve sonuçların daha tutarlı bir şekilde değerlendirilmesini sağlamayı hedeflemiştir.

Her iki veri seti üzerinde yapılan analizler sonucunda, Rastgele Orman algoritmasının en dengeli ve güçlü sonuçları sunduğu görülmüştür. Veri Seti-I'de ROC AUC değeri (0,838), F1 skoru (0,667) ve kesinlik değeri (0,697) ile en iyi sonuçları elde ederken, Veri Seti-II'de ROC AUC değeri (0,842), F1 skoru (0,667) ve kesinlik değeri (0,694) ile diğer modelleri geride bırakmıştır. Bu bulgular, Rastgele Orman algoritmasının hem spam tespiti (duyarlılık) hem de yanlış pozitifleri (FP) minimize etme konusunda dengeli bir performans sunduğunu göstermektedir. Veri Seti-I'de duyarlılık değeri (0,653), Veri Seti-II'de ise (0,641) olarak ölçülmüş olup, her iki veri setinde de FN oranlarının daha da azaltılması için iyileştirme potansiyeli bulunmaktadır. Ayrıca, düşük FP oranları sayesinde iletişim uzmanlarının yanlış pozitifleri inceleme iş yükü azalmakta ve zamanlarını daha verimli kullanmaları sağlanmaktadır.

Karar Ağacı algoritması, her iki veri setinde de ortalama bir performans sergilemiştir. Veri Seti-I'de 0,758 doğruluk ve 0,632 F1 skoru elde edilirken, Veri Seti-II'de 0,760 doğruluk ve 0,635 F1 skoru ile benzer bir başarı yakalanmıştır. Ancak, ROC AUC değerleri (Veri Seti-I: 0,727; Veri Seti-II: 0,728) ve duyarlılık değerleri (Veri Seti-I: 0,653; Veri Seti-II: 0,634), bu algoritmanın dengesiz veri setlerinde spam tespiti açısından yeterince güçlü olmadığını göstermektedir. Özellikle azınlık sınıfına (istenmeyen gönderi) yönelik öğrenme kapasitesinin sınırlı olması nedeniyle spam içerikleri belirleme konusunda zayıf kalmıştır. Bu durum, Karar Ağacı algoritmasının FN oranlarını düşürme konusunda yetersiz kaldığını göstermektedir.

Destek Vektör Makineleri (SVM), her iki veri setinde de düşük duyarlılık değerleri nedeniyle spam tespitinde yetersiz kalmıştır. Veri Seti-I'de duyarlılık değeri (0,453), Veri Seti-II'de ise (0,524) olarak ölçülmüş olup, diğer modellere kıyasla belirgin şekilde daha düşük performans göstermiştir. Bu durum, SVM'nin dengesiz veri setlerinde azınlık sınıfını ayırt etmede zorlandığını ve FN oranlarını artırdığını göstermektedir. Ayrıca, yüksek FP oranları iletişim uzmanlarının yanlış pozitifleri inceleme sürecini uzatarak iş yükünü artırabilir.

Lojistik Regresyon, her iki veri setinde de orta düzeyde bir performans sergilemiştir. Veri Seti-I'de 0,747 doğruluk ve 0,526 F1 skoru, Veri Seti-II'de ise 0,770 doğruluk ve 0,585 F1 skoru elde edilmiştir. Ancak, duyarlılık değerleri (Veri Seti-I: 0,442; Veri Seti-II: 0,496) algoritmanın spam tespitinde zayıf kaldığını göstermektedir. Bu durum, Lojistik Regresyon'un FN oranlarını azaltma konusunda etkisiz olduğunu ve iletişim uzmanlarının iş yükünü artırabilecek yanlış pozitifler (FP) ürettiğini göstermektedir.

Çok Katmanlı Algılayıcı (MLP), Veri Seti-I'de 0,734 doğruluk ve 0,519 F1 skoru ile en düşük performansı sergilerken, Veri Seti-II'de 0,774 doğruluk ve 0,615 F1 skoru ile nispeten daha iyi bir performans göstermiştir. ROC AUC değerleri (Veri Seti-I: 0,754; Veri Seti-II: 0,802) ve duyarlılık değerleri (Veri Seti-I: 0,453; Veri Seti-II: 0,550), bu algoritmanın dengesiz veri setlerinde spam tespiti açısından daha başarılı olduğunu göstermektedir. Bununla birlikte, FP oranlarının daha da optimize edilmesi gerekmektedir.

Elde edilen bulgular değerlendirildiğinde, Rastgele Orman algoritmasının dengesiz veri yapıları için en etkili çözümü sunduğu görülmektedir. Bu algoritma, duyarlılık ve kesinlik arasında optimal bir denge sağlayarak spam tespitinde yüksek performans göstermiştir. Özellikle spam içerikleri kaçırmama hedefi doğrultusunda, Rastgele Orman'ın yüksek duyarlılık (recall) değerleri FN oranlarını minimize etme konusunda etkili olmuştur. Ayrıca, düşük FP oranları sayesinde iletişim uzmanlarının yanlış pozitifleri inceleme yükü azalarak iş süreçlerinde verimlilik sağlanmaktadır.

Bu çalışma, makine öğrenmesi ve bilgisayarlı görü yöntemlerinin entegrasyonu sayesinde, spam tespiti alanında hem teorik hem de pratik önemli katkılar sunmaktadır. Teorik açıdan, metin ve görsel içeriklerin birlikte analiz edilmesi, geleneksel yaklaşımlara kıyasla daha kapsamlı bir çözüm önermektedir.

Pratikte ise, özellikle sosyal medya platformlarında otomatik spam tespit sistemlerinin geliştirilmesine yönelik önemli çıktılar sağlamaktadır. Ancak, modelin genelleştirilebilirliği konusunda sınırlamalar bulunmaktadır. Çalışmanın veri seti yalnızca Türkiye'de belirli bir zaman aralığında paylaşılan görselleri içermekte ve farklı platformlar, diller ve kültürel bağlamlarda test edilmemiştir. Bu durum, modelin yerel bağlamda etkili sonuçlar vermesini sağlamakla birlikte, uluslararası platformlarda uygulanabilirliği konusunda bazı sınırlamalar ortaya çıkarmaktadır. Özellikle, farklı kültürel bağlamlar, dil yapıları ve sosyal medya kullanım alışkanlıkları, modelin performansını etkileyebilir. Ancak, geliştirilen yöntemlerin (metin ve görsel analiz tekniklerinin entegrasyonu gibi) evrensel olarak uygulanabilir olduğu düşünülmektedir. Bu nedenle, geniş ölçekli uygulamalar için çapraz platform analizleri ve uzun vadeli verilerle doğrulama yapılması gerekmektedir.

Gelecekteki çalışmalarda, daha derin sinir ağı tabanlı yaklaşımlar (örneğin, Transformer modelleri veya CNN tabanlı görsel analiz teknikleri) ile mevcut yöntemin karşılaştırılması önemli bir adım olabilir. Özellikle BERT veya CLIP gibi modellerin metin ve görsel içerikleri birlikte analiz etme yetenekleri, spam tespit sistemlerinin doğruluğunu artırabilir. Ayrıca, gerçek zamanlı spam tespiti için işlem süresi ve hesaplama maliyetlerini optimize edecek stratejiler geliştirilmesi, bu tür sistemlerin pratik uygulanabilirliğini artıracaktır. Daha büyük ölçekli ve dinamik olarak güncellenen veri setleri üzerinde çapraz doğrulama yapılması, modelin genelleştirilebilirliğini ve güvenilirliğini artırabilir. Bunun yanı sıra, farklı platformlardan (Instagram, Facebook gibi) toplanacak çeşitli veri setleri üzerinde karşılaştırmalı analizler yapılarak, platformlar arası spam davranışlarının benzerlikleri ve farklılıkları incelenebilir. Bu tür çalışmalar, spam tespit sistemlerinin farklı sosyal medya ortamlarında daha etkili ve tutarlı bir şekilde çalışmasını sağlayabilir.

Ayrıca çalışmada kullanılan Google Gemini ve Cloud Vision AI gibi araçların alternatifleriyle (örneğin OpenAI GPT, Microsoft Azure Computer Vision, Amazon Rekognition) karşılaştırmalı incelemeler yapılması da önemli bir katkı sağlayabilir. Farklı modellerin performanslarının aynı veri setleri üzerinde test edilmesi, hangi araçların belirli senaryolarda daha etkili olduğunu ortaya koyabilir. Bu tür karşılaştırmalar, spam tespiti için en uygun araç seçimini kolaylaştırabilir ve daha geniş bir perspektif sunabilir.

Sonuç olarak, bu çalışma istenmeyen içerik tespiti alanında önemli bir temel oluşturarak, daha kapsamlı araştırmalar için bir yol haritası sunmuştur. Gelişmiş algoritmalar, daha zengin veri setleri, gerçek zamanlı analiz yöntemleri ve alternatif araçların karşılaştırmalı değerlendirilmesiyle, bu alanda daha etkili çözümler geliştirilmesi mümkün olacaktır.

KAYNAKLAR (REFERENCES)

[1] P. Sharma, T. Nagpal, G. Shrivastava, J. D. Kumar, "A Systematic Review on Social Bots Account Detection Using Machine Learning," *2023 5th International Conference on Advances in Computing, Communication Control and Networking* (ICAC3N), pp. 862–866, 2023.

[2] S. Ghosh, B. Viswanath, F. Kooti, N. K. Sharma, K. P. Gummadi, B. Krishnamurthy, "Understanding and combating link farming in

the twitter social network," *The Web Conference* (WWW), pp. 61–70, 2012.

[3] N. L. Jin, N. Y. Chen, N. T. Wang, N. P. Hui, A. V. Vasilakos, "Understanding user behavior in online social networks: a survey," *IEEE Communications Magazine*, vol. 51, no. 9, pp. 144–150, 2013.

[4] K. Thomas, C. Grier, D. Song, V. Paxson, "Suspended accounts in retrospect," 12th ACM Workshop on Hot Topics in Networks (HotNets), pp. 1–7, 2011.

[5] A. Aggarwal, A. Rajadesingan, P. Kumaraguru, "PhishAri: Automatic real-time phishing detection on twitter," eCrime Researchers Summit, pp. 1–12, 2012.

[6] D. Wang, S. Navathe, L. Liu, D. Irani, A. Tamersoy, C. Pu, "Click Traffic Analysis of Short URL Spam on Twitter," **IEEE International Conference on Big Data, pp**. 1–8, 2013.

[7] C. Grier, K. Thomas, V. Paxson, M. Zhang, "@spam," **Proceedings of the 2010 ACM Conference on Computer and Communications Security**, pp. 27–37, 2010.

[8] T. Wu, S. Liu, J. Zhang, Y. Xiang, "Twitter spam detection based on deep learning," Proceedings of the Australasian Computer Science Week Multiconference (ACSW), pp. 1–9, 2016.

[9] M. Mateen, M. A. Iqbal, M. Aleem, M. A. Islam, "A hybrid approach for spam detection for Twitter," **2022 19th International Bhurban Conference on Applied Sciences and Technology** (IBCAST), pp. 106–111, 2017.

[10] S. Sedhai, A. Sun, "Semi-Supervised Spam Detection in Twitter Stream," IEEE Transactions on Computational Social Systems, vol. 5, no. 1, pp. 169–175, 2017.

[11] I. Inuwa-Dutse, M. Liptrott, I. Korkontzelos, "Detection of spam-posting accounts on Twitter," Neurocomputing, vol. 315, pp. 496–511, 2018.

[12] S. Madisetty, M. S. Desarkar, "A Neural Network-Based Ensemble Approach for Spam Detection in Twitter," IEEE Transactions on Computational Social Systems, vol. 5, no. 4, pp. 973–984, 2018.

[13] R. Kaur, S. Singh, H. Kumar, "Rise of spam and compromised accounts in online social networks: A state-of-the-art review of different combating approaches," *Journal of Network and Computer Applications, vol.* 112, pp. 53–88, 2018.

[14] K. Binsaeed, G. Stringhini, A. Eleyan, "Detecting Spam in Twitter Microblogging Services: A Novel Machine Learning Approach based on Domain Popularity," *International Journal of Advanced Computer Science and Applications, vol*. 11, no. 11, pp. 1–7, 2020.

[15] K. S. Adewole, T. Han, W. Wu, H. Song, A. K. Sangaiah, "Twitter spam account detection based on clustering and classification methods," *The Journal of Supercomputing, vol.* 76, no. 7, pp. 4802–4837, 2018.

[16] N. El-Mawass, P. Honeine, L. Vercouter, "SimilCatch: Enhanced social spammers detection on Twitter using Markov Random Fields," Information Processing & Management, vol. 57, no. 6, p. 102317, 2020.

[17] Y. Kontsewaya, E. Antonov, A. Artamonov, "Evaluating the effectiveness of machine learning methods for spam detection," Procedia Computer Science, vol. 190, pp. 479–486, 2021.

[18] K. U. Santoshi, S. S. Bhavya, Y. B. Sri, B. Venkateswarlu, "Twitter Spam Detection Using Naïve Bayes Classifier," *International Conference on Inventive Computation Technologies, pp*. 773–777, 2021.

[19] S. B. Abkenar, E. Mahdipour, S. M. Jameii, M. H. Kashani, "A hybrid classification method for Twitter spam detection based on differential evolution and random forest," Concurrency and Computation: Practice and Experience, vol. 33, no. 21, 2021.

[20] C. Kumar, T. S. Bharti, S. Prakash, "A hybrid Data-Driven framework for Spam detection in Online Social Network," Procedia Computer Science, vol. 218, pp. 124–132, 2023.

[21] M. Sumathi, S. P. Raja, "Machine learning algorithm-based spam detection in social networks," Social Network Analysis and Mining, vol. 13, no. 1, 2023.

[22] M. Thomas, B. B. Meshram, "ChSO-DNFNet: Spam detection in Twitter using feature fusion and optimized Deep Neuro Fuzzy Network," Advances in Engineering Software, vol. 175, p. 103333, 2022.

[23] S. J. Alsunaidi, R. T. Alraddadi, H. Aljamaan, "Twitter spam accounts detection using machine learning models," **2022 14th International Conference on Computational Intelligence and Communication Networks (CICN)**, vol. 15, pp. 525–531, 2022.

[24] S. Kaddoura, S. Henno, "Dataset of Arabic spam and ham tweets," Data in Brief, vol. 52, p. 109904, 2023.

[25] D. SivaKrishna, G. Srinivas, "StopSpamX: A multi-modal fusion approach for spam detection in social networking," MethodsX, vol. 12, p. 103227, 2025.

[26] K. P. Sharma, M. Gupta, A. Mishra, "Quantum Behaved Binary Gravitational Search Algorithm with Random Forest for Twitter Spammer Detection," Results in Engineering, vol. 20, p. 103993, 2025.

[27] A. Nekrasov, S. H. Teoh, S. Wu, "Visuals and attention to earnings news on Twitter," *SSRN Electronic Journal*, 2019.

[28] C. Boididou, S. Papadopoulos, Y. Kompatsiaris, J. Schifferes, N. Newman, "Verifying information with multimedia content on Twitter," Multimedia Tools and Applications, vol. 77, no. 12, pp. 15545–15571, 2017.

[29] Cumhurbaşkanlığı İletişim Başkanlığı, Sosyal Medya Kullanım Kılavuzu, Türkiye, 2020.

[30] Q. Ren, H. Cheng, H. Han, "Research on machine learning framework based on random forest algorithm," **AIP Conference Proceedings, vol.** 1864, no. 1, p. 020164, 2017.

[31] A. Gupte, S. Joshi, P. Gadgul, A. Kadam, "Comparative Study of Classification Algorithms used in Sentiment Analysis," *International Journal of Computer Science and Information Technologies, vol.* 5, no. 5, pp. 6261–6264, 2014.