NÖHÜ Müh. Bilim. Derg. / NOHU J. Eng. Sci., 2025; 14(1), 175-182
Niğde Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi
Niğde Ömer Halisdemir University Journal of Engineering Sciences
Araştırma makalesi / Research article
www.dergipark.org.tr/tr/pub/ngumuh / www.dergipark.org.tr/en/pub/ngumuh

Derin öğrenme ve makine öğrenmesi yöntemleri ile sosyal medya verilerinden suç tespiti

Crime detection on social media data using deep learning and machine learning methods

Sultan Zeybek[*1], Berat Alkın[2], Yusuf Kaya[3]

[1] *Fatih Sultan Mehmet Vakıf Üniversitesi, Yapay Zeka ve Veri Mühendisliği Bölümü, 34015, İstanbul, Türkiye*
[2,3] *Fatih Sultan Mehmet Vakıf Üniversitesi, Bilgisayar Mühendisliği Bölümü, 34015, İstanbul, Türkiye*

Öz

Bu çalışmada, Türkçe sosyal medya paylaşımlarındaki tehdit ve hakaret içeriklerinin tespiti amaçlanmıştır. Doğal Dil İşleme teknikleri kullanılarak sosyal medya verileri üzerinde derin öğrenme algoritmalarıyla modeller geliştirilmiş ve bu modeller makine öğrenmesi algoritmaları ile karşılaştırılmıştır. Türkçe sosyal medya verilerinden toplanan veri kümesi etiketlenerek Uzun Kısa Süreli Bellek ve BERT derin öğrenme modelleri ile suç tespiti amacıyla kullanılmıştır. Derin öğrenme modelleri, makine öğrenmesi modellerinden Destek Vektör Makineleri, Rastgele Orman ve Gradyan Artırma modelleri ile karşılaştırılmıştır. Önerilen derin öğrenme modelleri, %90 doğruluk oranıyla tehdit ve hakaret içeriklerini başarılı bir şekilde tespit ederek makine öğrenmesi modellerine kıyasla daha üstün performans sergilemiştir.

Anahtar kelimeler: Derin öğrenme, Makine öğrenmesi, Sosyal medya, Suç tespiti, Sınıflandırma

Abstract

This study aims to detect threats and insults in Turkish social media posts. Models have been developed using Natural Language Processing techniques and deep learning algorithms, and the proposed models have been compared with machine learning algorithms. The dataset, collected from Turkish social media posts, has been labelled and used for crime detection in social media using Long Short-Term Memory and BERT deep learning models. The deep learning models have been compared with machine learning models such as Support Vector Machines, Random Forest, and Gradient Boosting. The proposed deep learning models have outperformed the machine learning models, successfully detecting threatening content with an accuracy of 90%.

Keywords: Deep learning, Machine learning, Social media, Crime detection, Classification

1 Giriş

İnternet kullanımının hızla yaygınlaşması ve sosyal medya platformlarının sayısının artmasıyla birlikte dijital ortamlarda üretilen içerik miktarı da önemli ölçüde artmıştır. Sosyal medya siteleri hem dünyada hem de Türkiye'de en çok ziyaret edilen web siteleri arasında yer almakta ve kullanıcılarına düşüncelerini özgürce paylaşma imkânı sunmaktadır. Ancak, bu platformlar aynı zamanda suç içeriklerinin, özellikle tehdit ve hakaret unsurlarının yayılmasına da zemin hazırlamaktadır. Tehdit ve hakaret gibi suçlar, bireylerin manevi dünyasına yönelik saldırılar içerir ve bu nedenle, bu içeriklerin tespit edilmesi hukuki süreçlerin işleyişi açısından büyük önem taşır [1].

Tehdit ve hakaret suçları, Türk Ceza Kanunu (TCK) kapsamında bireylerin manevi dünyasını koruyan önemli suç tiplerindendir [2]. Tehdit suçu, TCK'nın 106. maddesinde düzenlenmiş olup, bir kişiyi kendisinin ya da bir yakınının hayatına, vücut bütünlüğüne veya cinsel dokunulmazlığına zarar vereceği yönünde tehdit etmeyi kapsar. Bu suç, yalnızca gerçek kişilere karşı işlenebilir ve somut bir zarara yol açmasa da mağduru korkutmaya yönelik her türlü tehdit suç olarak kabul edilir. Hakaret suçu ise TCK'nın 125. maddesinde düzenlenmiştir ve bir kimsenin onur, şeref ve saygınlığını zedeleyecek nitelikte hakaret veya isnatlarda bulunmayı içerir. Bu suç da yalnızca gerçek kişilere karşı işlenir ve kişinin toplumsal itibarını korumayı hedefler. Sosyal medya platformlarında bu tür suçların doğru tespit edilmesi, kamu güvenliği açısından büyük önem taşımaktadır.

Son yıllarda sosyal medya platformlarındaki hakaret, tehdit ve nefret içeriklerinin tespitine yönelik yapılan çalışmalar, yapay zeka ve makine öğrenmesi tabanlı yaklaşımların bu alanda etkin çözümler sunduğunu göstermektedir. Karayiğit ve arkadaşlarının çalışmasında, Evrişimsel Sinir Ağları (CNN) gibi derin öğrenme yöntemlerinin Türkçe yorumlar üzerinde yüksek doğruluk oranıyla hakaret içeriklerini tespit edebildiği kanıtlanmıştır [3]. Benzer şekilde, Bozyiğit ve arkadaşları siber zorbalık gibi çevrimiçi şiddet içeriklerinin yapay sinir ağları ile başarıyla tespit edilebileceğini vurgulamaktadır [4]. Facebook zorbalığı ve mağduriyeti ölçeklerinin Türkçe 'ye uyarlanması çalışmasında, zorbalık ve mağduriyet konularında Türkçe 'ye uygun ölçekler geliştirilmiş; ancak bu çalışma doğrudan makine öğrenmesi yöntemlerine

* Sorumlu yazar / Corresponding author, e-posta / e-mail: szeybek@fsm.edu.tr (S. Zeybek)
Geliş / Recieved: 18.09.2024 Kabul / Accepted: 14.11.2024 Yayımlanma / Published: 15.01.2025
doi: 10.28948/ngumuh.1551734

odaklanmamıştır [5]. Kaynar vd. çeşitli öznitelik seçim yöntemleri kullanılarak saldırı tespit sistemleri için başarılı modeller geliştirilmiş ve saldırı ve tehdit içeriklerinin makine öğrenmesi yardımıyla nasıl tespit edilebileceğini göstermiştir [6]. Bu tür teknolojik çözümler, sosyal medya ve çevrimiçi platformlarda yaygın olarak kullanılan dilsel saldırıların tespitinde büyük bir ilerleme kaydetmiştir.

Türkçe sosyal medya verileri üzerinde nefret söylemi tespitine yönelik çeşitli yöntemler geliştirilmiştir. Nefret söylemi, toplumsal ayrımcılığı körükleyen ve bireylere ya da topluluklara yönelik düşmanca, aşağılayıcı ifadelerden oluşan bir suç kategorisidir. Sosyal medya platformlarının yaygın kullanımıyla birlikte bu tür içerikler hızla artmıştır. Beyhan ve arkadaşları Türkçe nefret söylemi veri kümesi ve tespit sistemi çalışması ile Türkçe sosyal medya içeriklerinde nefret söylemi tespiti için bir veri kümesi geliştirilmiş ve BERT gibi derin öğrenme tekniklerinin kullanımı incelenmiştir [7]. Çalışmada geliştirilen So-haTRed modeli ile makine öğrenmesi ile derin öğrenme yöntemlerinin hibrit bir yapı içinde birleştirilerek nefret söylemi tespitinde başarılı sonuçlar elde vereceği gösterilmiştir. Türkçe nefret söylemi tespiti için TurkishBERTweet modelinin kullanıldığı bir başka çalışma VRLLab at HSD-2Lang ortamında geliştirilmiştir. Bu model, sosyal medya platformlarında yayılan nefret içeriklerinin tespiti için düşük-rank adaptasyonu (LoRA) ile ince ayar yapılmış ve yüksek doğruluk oranlarına ulaşmıştır [8]. Bir diğer çalışmada ChatGPT modelinin Türkçe sosyal medya gönderilerinde nefret söylemi tespitindeki yeteneklerini incelenmiştir. Çalışmada, modelin özellikle dilin karmaşık yapılarında ne derece etkili olduğunu ve nefret söylemini tespit etme kapasitesini değerlendirerek derin öğrenme yöntemleri ile karşılaştırmalı bir analiz sunmaktadır [9]. Bayrak ve arkadaşları çalışmasında, BERT-Base modelinin Türkçe sosyal medya gönderilerinde nefret söylemi tespitinde %92.53 test doğruluğu ile başarılı sonuçlar elde ettiğini göstermişlerdir. [10]. Benzer bir çalışmada Hierarchical Attention Network (HAN) ve BERT tabanlı derin öğrenme modelleri ile nefret söylemi tespiti üzerinde durulmuş ve eleştirel söylem analizinden elde edilen dilsel özelliklerle yüksek doğruluk oranlarına ulaşılmıştır [11].

Bu çalışmalar, BERT tabanlı modellerin Türkçe nefret söylemi tespitinde ne derece etkili olduğunu göstermektedir. Ancak bu çalışmaların büyük bir kısmı belirli platformlara ve spesifik içerik türlerine odaklanmaktadır ve genel tehdit veya hakaret suçlarını kapsayan geniş çaplı bir model geliştirilmemiştir. Aynı zamanda, Özar'ın çalışmasında dile getirildiği gibi, Türkiye'deki nefret suçlarına yönelik yasal düzenlemelerin kapsamı da bu teknolojik gelişmelerle tam anlamıyla entegre edilmemiştir [12]. Türkçe salgırgan ifadeler içeren dil kullanımı üzerine hazırlanmış ilk veri kümesi Çöltekin [13] tarafından Twitter'dan toplanmış hakaret ve nefret söylemi içeren ifadelerle Türkçe dil yapısının ve sosyal medyadaki gayri resmi dil kullanımının özelliklerini göz önüne sermiştir. Bu veri kümelerinin genişletilerek, hakaret ve tehdit içeriklerinin tespiti ve sınıflandırılması için makine öğrenmesi ve derin öğrenmemodellerinin geliştirilmesi gerekmektedir. Günümüzde hem hukuki hem de teknolojik alanlarda hakaret ve tehdit içeriklerinin daha geniş çaplı ve etkin bir şekilde tespitine ve önlenmesine yönelik bir boşluk olduğu görülmektedir.

Bu çalışmada, söz konusu boşluğu doldurmak amacıyla Türkçe dilinde hakaret ve tehdit suçlarının daha kapsamlı tespiti için makine öğrenmesi ve derin öğrenme tabanlı bir model önerilmektedir. Çalışmanın temel amacı tehdit ve hakaret içeren paylaşımların tespit tehdit ve hakaret içeriklerinin tespitine yönelik doğal dil işleme tekniklerinin etkinliğini incelemektedir. Bunun için Türkçe sosyal medya paylaşımlarından suç içeriklerini tespit etmek amacıyla derin öğrenme ve makine öğrenmesi yöntemleri geliştirilmiştir ve önerilen modeller üzerinden performans karşılaştırmaları yapılmıştır.

Çalışmanın ikinci bölümünde geliştirilen sınıflandırma modelinde kullanılan makine öğrenmesi ve derin öğrenme metotlarına yer verilmiştir. Veri kümesi ve veri önişleme aşamalarından bahsedilerek karşılaştırmalı performens sonuçları raporlanmıştır. Son olarak sonuçlar ve gelecek çalışmalar önerileri ile çalışma sonlandırılmıştır.

2 Materyal ve metot

Türk Hukukuna dayalı suç tespiti problemi, bir duygu analizi problemi değildir. Bir Türkçe cümlenin tehdit ya da hakaret ifadesi içeriyor olmasını belirleyebilmek için, cümlenin genel anlamını öğelerden anlayabilmek ve Türk kanunlarına göre karar verebilmek gerekir.

Bu çalışmada, Türkçe hakaret ve tehdit suçlarının tespiti bir metin sınıflandırma problemi olarak ele alınmıştır. Metin sınıflandırma, belirli bir metni içeriğine dayanarak hakaret, tehdit veya diğer suç kategorilerine ayırma sürecini ifade eder. Bu işlem, genellikle denetimli makine öğrenmesi teknikleri ile gerçekleştirilir ve modelin doğru sınıfları öğrenebilmesi için etiketlenmiş veriler kullanılır.

Çalışmanın ilk aşamasında hakaret ve tehdit içeren sosyal medya gönderileri karşılık gelen sınıf etiketleri ile etiketlenerek veri seti oluşturulmuştur. Tehdit ve hakaret içerikleri yasal mevzuat doğrultusunda incelenmiş ve suç teşkil edebilecek içeriklerin belirlenmesine özen gösterilmiştir. Daha sonra veri ön işleme yöntemleri ile metin verisi standart bir forma dönüştürülerek makine öğrenmesi ve derin öğrenme algoritmalarının anlayabileceği sayısal özelliklere dönüştürülmüştür. Elde edilen eğitim verileri makine öğrenmesi yöntemlerin ve derin öğrenme yöntemleri ile eğitilerek sınıflandırma performansları elde edilmiştir. Modeller, eğitim süreçleri boyunca değerlendirilmiş ve önerilen modellerin doğruluğu test verileri üzerinde değerlendirilmiştir. Son aşamada ise, eğitim süreci tamamlanan model, yeni gelen metinleri hakaret ve tehdit içeriklerine göre sınıflandırmak amacıyla kullanılmıştır.

2.1 Veri seti ve veri ön işleme

Çalışma boyunca veri kümesi olarak Yıldız Teknik Üniversitesi Doğal Dil İşleme Grubu tarafından oluşturulan ve 20 milyon Türkçe tweet içeren geniş bir veri kümesi kullanılmıştır [14]. Veri kümesi, 2012 yılında çeşitli zaman dilimlerinde ve birçok farklı konuda atılmış rastgele tweetlerden oluşmaktadır. Bu veriler arasında siyaset, spor, eğlence ve gündelik konular gibi çeşitli kategoriler yer

almaktadır. Tweetlerin seçimi, rastgele örnekleme yöntemi ile yapılmıştır. Veri önişleme kapsamında, tweetlerde bulunan kısaltmalar, hashtagler, emojiler ve bağlantılar gibi metin dışı içerikler zemberek-nlp kütüphanesi kullanılarak temizlenmiştir [15]. Daha sonra metinlerde yapılan yazım yanlışları düzeltilmiştir. Böylece aynı kelimelerin yazım hataları veya kısaltmalar nedeniyle farklı biçimlerde yazılmasının önüne geçilmiştir (Bk. Tablo 1). Normalizasyon sürecinde düzeltilemeyen yazım yanlışları elle düzeltilmiştir.

Tablo 1. Veri ön işleme öncesi ve sonrası

Tweet	Veri Ön işleme Sonrası
Yrn okua gidicem	Yarın okula gideceğim.
Tmm yarin havuza giricem ve aksama kadar yaticam	Tamam yarın havuza gireceğim ve akşama kadar yatacağım.
@zekikayahan alarm çaldıktan sonra kapatıp tekrar uyumanın #hastasiyim	Alarm çaldıktan sonra kapatıp tekrar uyumanın hastasiyim.

Daha sonra tweetlerde bulunan cümleler Tablo 2'de görüldüğü gibi kök ve eklerine ayrılarak anahtar kelime listesinde bulunan herhangi bir kelimeyi içeren cümleler kontrol listesine alınmıştır. Veriler tehdit, hakaret ve nötr olmak üzere üç ayrı sınıfa göre etiketlenmiştir. Veri etiketleme öncesinde, emsal hukuk davalarında geçen ifadeler dikkate alınmış ve tehdit ile hakaret içeriklerine yönelik anahtar kelimeler oluşturulmuştur. Bu kapsamda Türk Hukuku'nda yargıtay kararları kapsamında yapılan araştırmalar sonucunda elde edilen hukuki dosyalardan ve örnek davalardan hakaret kelimelerinin ve tehdit kelimelerini içeren bir anahtar liste oluşturularak bu listeye göre etiketleme işlemi yapılmıştır. Tehdit ve hakaret sınıfları her ne kadar benzer olsalar da bazı anlam farklılıkları içermektedir. Tehdit sınıfı için cümlenin bir kişiye zarar verme, kötü bir durum yaşatma niyetini içermesi gerekirken hakaret sınıflandırması için cümlenin kişinin itibarına saldırma, aşağılama veya küçümseme gibi unsurlar bulundurması gerekmektedir. Tehdit içeren cümleler genellikle gelecekteki olası zararları öne çıkarırken, hakaret içeren cümleler kişinin karakterine veya yeteneklerine yönelik aşağılayıcı ifadeler içerir. Her iki durumda da, cümlenin bağlamını dikkate almak ve dilin tonunu anlamak önemlidir. Bu farkı gözetmek üzere önceden hazırlanmış olan ve tehdit unsuru oluşturabilecek kelimelerin olduğu liste kullanılarak kelime kökü 1. tekil/çoğul şahıs ekleriyle veya kelime kökü gelecek/geniş zaman ekleriyle kullanılan cümlelerin olduğu tehdit anlamı içerebilecek tweetler etiketlenmek üzere ayıklanmıştır. Ayrıca incelenen sosyal medya verilerinde geçen argo ifadeler de, TCK Madde 125 kapsamında hakaret suçu olarak değerlendirilmiştir. Etiketleme sonucunda her bir sınıfa ait 1000 adet tweet içeren veri kümeleri oluşturulmuştur. Nötr sınıfa ait verilerden 500 adet veri hakaret veya tehdit sınıfından kelimeleri içermesine rağmen, tehdit-hakaret anlamı içermeyecek şekilde olması sağlanmıştır. Ayrıca Levenshtein uzaklığı algoritması kullanılarak birbirine çok fazla benzeyen veriler ile yeni veriler değiştirilmiştir.

Tablo 2. Kök ve eklere ayırma öncesi ve sonrası

Ön işlemden geçmiş tweet	Kök ve eklere ayırma sonrası
Ahmaklar anca böyle düşünür zaten	ahmak Adj Zero Noun A3pl an Noun A3sg Equ böyle Adj düşün Verb Aor A3sg zaten Adv
donarak öl aptal	don Verb ByDoingSo Adv ol Verb Imp A2sg aptal Adj
seni öldüreceğim	sen Noun A3sg P1sg Acc öldür Verb Fut A1sg.

Ön işleme sürecinde veri setinde bulunan etkisiz kelimeler (stopwords) listesi çıkarılmış ve minimum belge sıklığı kriteri uygulanmıştır. Python'un scikit-learn kütüphanesindeki CountVectorizer fonksiyonu kullanılarak kelime çantası (Bag of Words) modeli ile veri kümesi vektörize edilmiştir. Vektörizasyon sonrası elde edilen kelime frekansları analiz edilerek en sık geçen kelimeler Tablo 3'te sunulmuştur. Özellikle, veride sık geçen küfür, tehdit ve hakaret içeren kelimelerin yüksek frekanslarda olduğu gözlemlenmiştir. Bu sonuçlar, çalışma kapsamında incelenen veri kümesinin doğasına uygun olup, söz konusu içeriklerin etkili bir şekilde sınıflandırılması için önemli bir temel oluşturmaktadır.

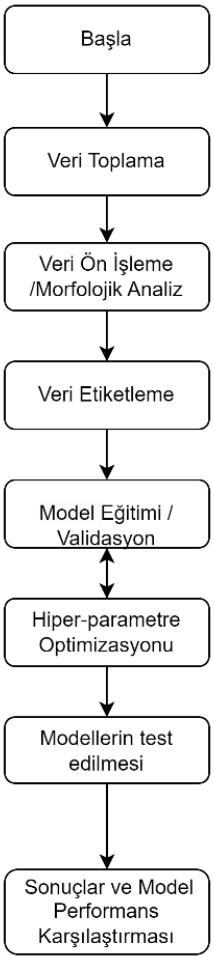
Tablo 3. Veri setinde en çok sıklığa sahip kelimeler

Kelime	Sıklığı / Frekansı
*mına	167
lan	133
or*spu	128
s*keceğim	122
ö*düreceğim	116

2.2 Makine Öğrenmesi ve Derin Öğrenme Yöntemleri

Çalışma boyunca makine öğrenmesi yöntemlerinden Destek Vektör Sınıflandırıcısı (DVS), Rastgele Orman Sınıflandırıcı (ROS), Gradyan Artırma Sınıflandırıcısı (GAS) kullanılmıştır. DVS, makine öğrenmesi alanında sınıflandırma problemlerini çözmek için yaygın olarak kullanılan güçlü bir algoritmadır. Bu algoritma, bir veri noktasının hangi sınıfa ait olduğunu belirlerken, sınıflar arasındaki en uygun karar sınırlarını bulmaya odaklanır [16]. DVS, sınıflar arasındaki boşluğu maksimize ederek karar sınırlarını belirler. Bu karar sınırları, "destek vektörleri" olarak adlandırılan ve sınıflar arasında en kritik olan veri noktalarına dayanarak çizilir. Algoritma, veri noktalarını bir vektör uzayında temsil eder ve bu uzayda en iyi ayırma çizgisini bulur [17]. Şekil 1, önerilen makine öğrenemsi ve derin öğrenme modellerinin iş akış şemasını göstermektedir.

ROS, bir dizi karar ağacından oluşan güçlü bir sınıflandırma ve regresyon modelidir. Bu model, her bir karar ağacının farklı özellik ve veri alt kümeleri üzerinde eğitilmesi ile oluşturulur. Karar ağaçları, veri kümesindeki özelliklere dayalı olarak basit karar kuralları oluşturur ve veri noktalarını sınıflandırır. Rastgele Orman algoritması, her bir ağacın bağımsız olarak tahmin yapmasını sağlar ve bu tahminlerin ortalaması alınarak veya çoğunluk kararıyla nihai sonuç elde edilir [18].



Şekil 1. Makine öğrenmesi ve derin öğrenme modellerinin iş akış şeması.

Bu algoritmanın önemli bir avantajı, yüksek boyutlu ve karmaşık veri kümeleri üzerinde etkili bir şekilde çalışabilmesidir. Ayrıca, aşırı uyum (overfitting) eğilimini azaltarak modelin genelleme kapasitesini artırır. Bu, her bir ağacın rastgele alt kümeler üzerinde eğitilmesi sayesinde sağlanır ve modelin farklı veri alt kümeleri ve özellikler üzerinde öğrenme yapabilmesi mümkün olur. GAS, zayıf öğrenicilerin ardışık olarak bir araya getirilmesiyle güçlü bir sınıflandırıcı oluşturan bir makine öğrenimi yöntemidir. Bu algoritmada, her bir öğrenici, önceki tahmincilerin yaptığı hataları düzeltmeye çalışarak performanslarını artırır. Bu sayede genel hata azalır ve daha güçlü bir model elde edilir [19]. Öğrenme sırasında her iterasyonda kayıp fonksiyonu optimize ederek tahminlerin gerçek değerlerden ne kadar uzak olduğunu ölçülür ve bu farkın minimize edilmesi için gradyan inişi optimizasyonu kullanılır. GAS, genellikle yüksek doğruluk sağlar, aykırı verilere karşı dirençlidir ve çoğu sınıflandırma görevinde diğer yöntemlerden daha iyi performans gösterir. Bununla birlikte, aşırı uyum riskini azaltmak için modelin parametrelerinin dikkatli bir şekilde ayarlanması gereklidir [20].

Derin öğrenme, makine öğrenmesinin bir alt dalı olup, çok katmanlı yapay sinir ağları kullanarak karmaşık veri örüntülerini modellemeye odaklanmaktadır. Böylece düşük seviyeli özelliklerden yüksek seviyeli özelliklerin öğrenildiği hiyerarşik bir özellik öğrenme süreci sağlanır. Bu çalışma kapsamında derin öğrenme metotlarından Uzun Kısa Süreli Bellek (LSTM), ve BERT (Bidirectional Encoder Representations from Transformers) modelleri kullanılmıştır.

LSTM modeli özellikle zaman serisi verileri ve sıralı veri analizlerinde yaygın olarak kullanılan bir tekrarlayan sinir ağı (Recurrent Neural Network, RNN) mimarisidir. LSTM'ler, geleneksel RNN'lerin uzun vadeli bağımlılıkları öğrenme konusundaki zayıflıklarını gidermek amacıyla geliştirilmiştir. Sepp Hochreiter ve Jürgen Schmidhuber tarafından 1997 yılında sunulan makalelerinde tanıtılan LSTM'lerin en büyük yeniliği, uzun süreli bağımlılıkları öğrenme kapasitesine sahip bir bellek hücresine ve bilgi akışını yöneten unutma kapısı (forget gate), giriş kapısı (input gate) ve çıkış kapısı (output gate) olmak üzere üç kapı sistemine sahip olmalarıdır [21]. Bu kapılar, hangi bilgilerin bellekte tutulacağı, hangi bilgilerin güncelleneceği ve hangi bilgilerin çıktı olarak verileceği konusunda karar vererek modelin uzun süreli bağımlılıkları daha etkin bir şekilde öğrenmesini sağlamaktadır.

BERT, doğal dil işleme alanında kullanılan bir model olup, çift yönlü dönüştürücülerden elde edilen temsilciliği ifade eder. Google tarafından 2018 yılında tanıtılan BERT, çeşitli NLP görevlerinde önemli ilerlemeler kaydetmiştir. BERT'in temel yeniliği, çift yönlü eğitim süreci sayesinde bağlamı hem sol hem de sağ yönden anlayabilmesidir. Bu özellik, modelin cümle içindeki kelimelerin anlamını daha iyi kavramasına olanak tanır. BERT'in mimarisi, Transformer modeline dayanmakta olup, metin temsilciliğini öğrenmek için dikkat (attention) mekanizmalarını kullanır [22].

3 Bulgular ve tartışma

Bu bölümde oluşturulan veri kümesi üzerinde uygulanan sınıflandırma modelleri sonucu elde edilen bulgular verilmiştir.

Makine öğrenmesi modellerinin eğitimi için, scikit-learn kütüphanesi kullanılarak çeşitli algoritmalar uygulanmış ve hiperparametre optimizasyonu Grid Search ve Cross Validation yöntemleri ile gerçekleştirilmiştir. Eğitme ve test isabet oranları karşılaştırılarak aşırı uyum (overfitting) riski minimize edilmiş ve sonuçlar değerlendirilmiştir. Modellerin performans değerlendirmeleri sonucunda, elde edilen sınıflandırma doğruluğu, makine öğrenmesi algoritmalarının etkinliğini göstermiştir.

Derin öğrenme aşamasında, Google Colab ortamında LSTM ve BERT modelleri geliştirlmiştir. LSTM modeli en fazla 5 devir boyunca eğitilmiştir. Devir sayısı modellerin aşırı öğrenmesini (overfitting) önleyerek aşırı uyuma gitmeden en iyi performansa ulaşmasını sağlamak amacıyla küçük tutulmuştur. Öğrenme oranı için learning rate ayarı yapılmış ve modellerin performansını optimize etmek amacıyla erken durdurma (early stopping) stratejisi kullanılmıştır.

3.1 Deneysel sonuçlar

Sınıflandırma modellerinin performansı çeşitli metriklerle değerlendirilmiştir. Performans ölçütleri

arasında aşağıdaki denklemlerle ifade edilen doğruluk (accuracy), kesinlik (precision), duyarlılık (recall) ve F1 skoru (F1 score) yer almaktadır.

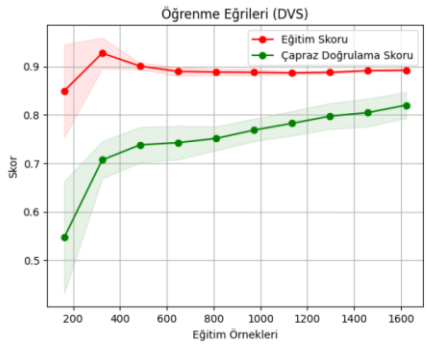
Doğruluk, modelin tüm veri kümesinde ne kadar doğru tahmin yaptığını gösterir ve doğru pozitifler (TP, True Positives) ile doğru negatiflerin (TN, True Negatives) toplam tahminler içindeki oranı olarak hesaplanır. Kesinlik, modelin pozitif olarak tahmin ettiği örnekler arasında gerçekten doğru olanların oranını ifade eder. Bu, doğru pozitiflerin (TP) yanlış pozitifler (FP, False Positives) ile birlikte toplam pozitif tahminler içindeki oranıdır. Duyarlılık (recall), modelin aslında pozitif olan verileri ne kadar doğru tespit ettiğini ölçer; bu, doğru pozitiflerin (TP) yanlış negatifler (FN, False Negatives) ile birlikte toplam pozitif örnekler içindeki oranı olarak hesaplanır. F1 skoru (F1 score) ise kesinlik ve duyarlılığın harmonik ortalamasıdır ve bu iki metriğin dengesini sağlayarak modelin genel performansını özetlemektedir.

3.1.1 Makine öğrenmesi modellerinin deneysel sonuçları

Karşılaştırmalı performans değerlendirmeleri sırasıyla makine öğrenmesi ve derin öğrenme modellerinin performansları üzerinden verilmiştir. Tablo 4'de DVS modeline ait 5-Katlı Çapraz Doğrulama (5-Fold Cross Validation) sonuçları verilmektedir. Buna göre DVS modelinin 5-katlı çapraz doğruluk oranı ortalaması 0.82 olarak hesaplanmıştır. Şekil 2'te görüldüğü üzere, DVS modeli için eğitim ve çapraz doğrulama skorları karşılaştırılmıştır.

Tablo 4. DVS 5-Katlı çapraz doğrulama sonuçları

K-Kat	Doğruluk
1	0.869458
2	0.800492
3	0.827586
4	0.804938
5	0.797530



Şekil 2. DVS'ye ait öğrenme eğrileri

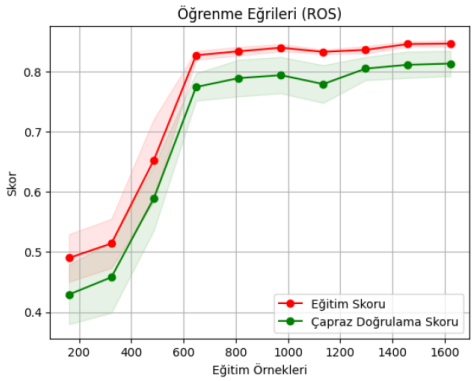
DVS modelinde eğitim skoru başlangıçta oldukça yüksek (0.9'un üzerinde) başlamış ve eğitim örnekleri arttıkça bu yüksek seviyede kalmıştır. Bu durum, modelin eğitim verisi üzerinde aşırı öğrenme gösterdiğini ve modele verilen yeni veriyle eğitim performansının dengelenmediğini işaret eder. Çapraz doğrulama skoru ise eğitim skoru ile karşılaştırıldığında başlangıçta düşük kalmış ve eğitim örneklerinin sayısı arttıkça yavaş bir şekilde yükselmiştir. Ancak eğitim skoru ile arasında belirgin bir fark kalmış, bu da modelin eğitim verisine fazla uyum sağladığını ve test verisinde daha düşük performans gösterdiğini ortaya koymuştur.

Tablo 5. DVS mimarisine ait hakaret, tehdit ve nötr anlam içeren sınıfların performans karşılaştırmaları

Sınıflar/Metrikler	Kesinlik	Duyarlılık	F1 Skoru
Hakaret	0.93	0.66	0.78
Nötr	0.68	0.89	0.77
Tehdit	0.89	0.87	0.88

Tablo 5'te, DVS modeli ile hakaret, tehdit ve nötr anlam içeren sınıfların performans karşılaştırmaları yapılmıştır. **Kesinlik** (precision), **duyarlılık** (recall) ve **F1 skoru** metrikleri her sınıf için değerlendirilmiştir. Hakaret sınıfında kesinlik değeri oldukça yüksek (0.93) olmasına rağmen, duyarlılık değeri (0.66) görece daha düşüktür, bu da modelin hakaret içeren örnekleri doğru tanımlama oranının nispeten sınırlı olduğunu göstermektedir. Buna karşın, nötr sınıfta duyarlılık değeri yüksek (0.89) olmasına rağmen kesinlik (0.68) biraz daha düşüktür. Tehdit sınıfında ise hem kesinlik (0.89) hem de duyarlılık (0.87) dengeli olup, F1 skoru da 0.81 ile oldukça iyi bir performans göstermektedir. Genel olarak, DVS modelinin tehdit sınıfında daha dengeli ve tutarlı bir performans sergilediği görülmüştür.

Şekil 3'te, ROS modeline ait öğrenme eğrileri sunulmuştur. ROS modelinde eğitim skoru, eğitim örnekleri sayısı arttıkça hızla artmış ve yaklaşık 600 örnekten sonra 0.85 civarına ulaşarak sabitlenmiştir. Bu durum, modelin eğitim verisine aşırı uyum göstermeden yüksek bir performans seviyesine ulaştığını gösterir. Çapraz doğrulama skoru ise başlangıçta düşük başlamış olmasına rağmen, eğitim örnekleri arttıkça hızlı bir şekilde yükselmiş ve yaklaşık 0.8 seviyesine ulaşmıştır. Eğitim skoru ile çapraz doğrulama skoru arasındaki fark, DVS modeline kıyasla çok daha küçüktür. Bu da ROS modelinin genelleme kapasitesinin daha iyi olduğunu ve aşırı uyum göstermeden test verisi üzerinde başarılı bir performans sergilediğini ortaya koymaktadır. Tablo 6, ROS modeline ait hakaret, tehdit ve nötr anlam içeren sınıfların performanslarını raporlamaktadır.

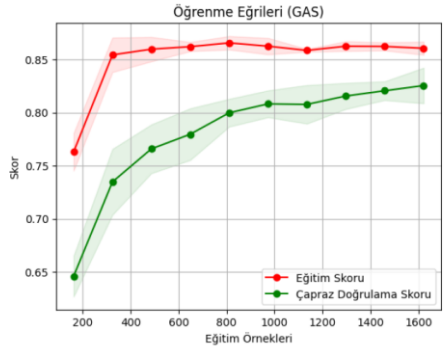


Şekil 3. ROS'a ait öğrenme eğrileri

Tablo 6. ROS mimarisine ait hakaret, tehdit ve nötr anlam içeren sınıfların performans karşılaştırmaları

Sınıflar/Metrikler	Kesinlik	Duyarlılık	F1 Skoru
Hakaret	0.87	0.74	0.80
Nötr	0.71	0.84	0.77
Tehdit	0.86	0.84	0.85

Şekil 4'te, GAS modelinin öğrenme eğrileri görülmektedir. Eğitim skoru, diğer modellerde olduğu gibi hızlı bir yükseliş göstermiş ve yaklaşık 400 örnekten sonra 0.85 seviyesine ulaşarak sabit kalmıştır. Eğitim örneklerinin artmasıyla birlikte GAS modelinin performansında büyük bir değişiklik olmamıştır. Bu GAS modelinin eğitim verisi üzerindeki performansının dengelendiğini göstermektedir. Çapraz doğrulama skoru, başlangıçta düşük olmasına rağmen eğitim örnekleri arttıkça düzenli bir şekilde yükselmiş ve yaklaşık 0.8 seviyesine ulaşmıştır. GAS modelinin çapraz doğrulama skoru ile eğitim skoru arasındaki farkın küçük olması, modelin genelleme yeteneğinin güçlü olduğunu ve aşırı uyum göstermediğini ortaya koymaktadır. Tablo 7 GAS modeline ait hakaret, tehdit ve nötr anlam içeren sınıfların performanslarını raporlamaktadır.



Şekil 4. GAS'a ait öğrenme eğrileri

Tablo 7. GAS mimarisine ait hakaret, tehdit ve nötr anlam içeren sınıfların performans karşılaştırmaları

Sınıflar/Metrikler	Kesinlik	Duyarlılık	F1 Skoru
Hakaret	0.92	0.72	0.81
Nötr	0.71	0.86	0.78
Tehdit	0.88	0.87	0.87

Tablo 8'de sunulan performans karşılaştırmalarına göre, makine öğrenmesi modelleri doğruluk metriği üzerinden değerlendirildiğinde farklı sonuçlar ortaya çıkmaktadır. DVS eğitim seti üzerinde en yüksek doğruluk oranını (0.8920) elde etmesine karşın, test setindeki doğruluğu 0.8087'ye düşerek modelin aşırı uyum (overfitting) eğilimi gösterdiğini ortaya koymaktadır. ROS ise daha dengeli bir performans sergileyerek eğitim setinde 0.8402, test setinde ise 0.8028 doğruluk oranına ulaşmış, 5-katlı çapraz doğrulama ortalaması ise 0.8111 olmuştur. Bu da ROS modelinin genelleme yeteneğinin kabul edilebilir seviyelerde olduğunu göstermektedir. GAS modeli ise hem eğitim setinde (0.8555), hem test setinde (0.8185) hem de 5-katlı çapraz doğrulama ortalamasında (0.8249) en yüksek sonuçları elde ederek, genelleme kapasitesi açısından en iyi performansı sergilemiştir. Bu sonuçlar, GAS modelinin genel performans açısından diğer modellere kıyasla daha güçlü ve tutarlı olduğunu ortaya koyarken, ROS modeli dengeli bir performans sunmakta, DVS modeli ise aşırı uyum göstererek genelleme yeteneğinde sınırlamalara sahip olduğunu göstermektedir.

Tablo 8. Makine öğrenmesi yöntemlerinin doğruluk metriği üzerinden performans karşılaştırmaları

	Destek Vektör	Rastgele Orman	Gradyan Artırma
Eğitim Seti	**0.8920**	0.8402	0.8555
Test Seti	0.8087	0.8028	**0.8185**
5-Katlı Çapraz Doğrulama Ort.	0.8200	0.8111	**0.8249**

3.1.2 Derin öğrenme modellerinin performans sonuçları

LSTM ve BERT modellerinin genelleme yeteneğini artımak ve aşırı uyum (overfitting) riskini önlemek amacıyla devir sayısı (epoch) 4 ile sınırlandırılmıştır. Tablo 9 ve Tablo 10'da sunulan LSTM modeline ait sonuçlar, modelin her devirde (epoch) gösterdiği performansı doğruluk, F1 skoru, kesinlik ve duyarlılık metrikleri üzerinden değerlendirmektedir. Tablo 9'da, LSTM modelinin eğitim ve doğrulama kayıpları her devirde düzenli olarak azalmıştır; ilk devirde 1.0931 olan eğitim kaybı, dördüncü devirde 0.4020'ye düşerken, doğrulama kaybı da benzer şekilde 1.0800'den 0.4596'ya gerilemiştir. Buna paralel olarak, doğruluk oranı ilk devirde 0.3915 gibi düşük bir seviyede başlarken, dördüncü devirde 0.8649 gibi oldukça yüksek bir seviyeye ulaşmıştır. Tablo 10'da ise LSTM modelinin F1 skoru, kesinlik ve duyarlılık metrikleri incelendiğinde, modelin her üç metrikte de gelişim gösterdiği görülmektedir. İlk devirde F1 skoru 0.5449 iken, dördüncü devirde 0.8298'e yükselmiştir. Kesinlik ve duyarlılık değerleri de sırasıyla 0.6786 ve 0.5542 seviyelerinden başlayarak, dördüncü devirde 0.8414 ve 0.8284 seviyelerine ulaşmıştır. Bu sonuçlar, LSTM modelinin her devirde eğitim verisinden daha fazla öğrenme sağladığını, doğruluk, kesinlik ve duyarlılık açısından performansını önemli ölçüde artırdığını ve modelin genel genelleme yeteneğinin her aşamada iyileştiğini göstermektedir.

Tablo 9. LSTM modelinin doğruluk (isabet) metriği üzerinden performans karşılaştırmaları

Devir	Eğitim Kaybı	Doğrulama Kaybı	Doğruluk (İsabet) Oranı
1	1.093174	1.080025	0.391519
2	1.019725	0.914952	0.564103
3	0.707954	0.647656	0.731262
4	0.402070	0.459576	0.864892

Tablo 10. LSTM modelinin F1 skoru, kesinlik ve duyarlılık metriği üzerinden performans karşılaştırmaları

Devir	F1 Skoru	Kesinlik	Duyarlılık
1	0.544923	0.678634	0.554241
2	0.644545	0.696252	0.637081
3	0.739120	0.792398	0.743590
4	0.829840	0.841367	0.828402

Tablo 11 ve Tablo 12'de sunulan BERT modeli sonuçlarına göre, modelin her devirde eğitim ve doğrulama kayıpları düzenli olarak azalmış ve doğruluk oranı ilk devirde 0.8225 iken üçüncü devirde 0.9211'e yükselmiştir. Benzer şekilde, F1 skoru, kesinlik ve duyarlılık metrikleri de her devirde iyileşme göstermiştir. İlk devirde 0.8221 olan F1 skoru, üçüncü devirde 0.9209'a ulaşmıştır. Kesinlik ve duyarlılık da sırasıyla 0.8385 ve 0.8225'ten üçüncü devirde 0.9245 ve 0.9211'e yükselmiştir. Bu sonuçlar, BERT modelinin her aşamada daha yüksek doğruluk ve genelleme kapasitesine ulaştığını göstermektedir.

Tablo 11. BERT modelinin doğruluk (isabet) metriği üzerinden performans karşılaştırmaları

Devir	Eğitim Kaybı	Doğrulama Kaybı	Doğruluk (İsabet) Oranı
1	1.093174	0.542680	0.822485
2	0.463300	0.287721	0.887574
3	0.256500	0.277191	0.921105

Tablo 12. BERT modelinin F1 skoru, kesinlik ve duyarlılık metriği üzerinden performans karşılaştırmaları

Devir	F1 Skoru	Kesinlik	Duyarlılık
1	0.822120	0.838480	0.822485
2	0.887493	0.888696	0.887574
3	0.920852	0.924501	0.921105

Şekil 5'te sunulan LSTM (a) ve BERT (b) modellerine ait eğitim kaybı, doğrulama kaybı ve doğruluk (isabet) oranı grafiklerine göre iki modelin performansı karşılaştırılmaktadır. Buna göre, her iki model de eğitim ve doğrulama kayıplarını düzenli olarak azaltarak doğruluk oranını artırmıştır. Ancak, BERT modeli daha hızlı öğrenme sağlamış ve daha düşük eğitim ve doğrulama kaybı ile üçüncü devirde LSTM modelinin dördüncü devrine kıyasla daha yüksek bir doğruluk oranına (0.9211) ulaşmıştır. Bu sonuçlar, BERT modelinin daha hızlı yakınsama sağladığını ve LSTM modeline kıyasla daha kısa sürede daha yüksek doğruluk elde ettiğini göstermektedir.

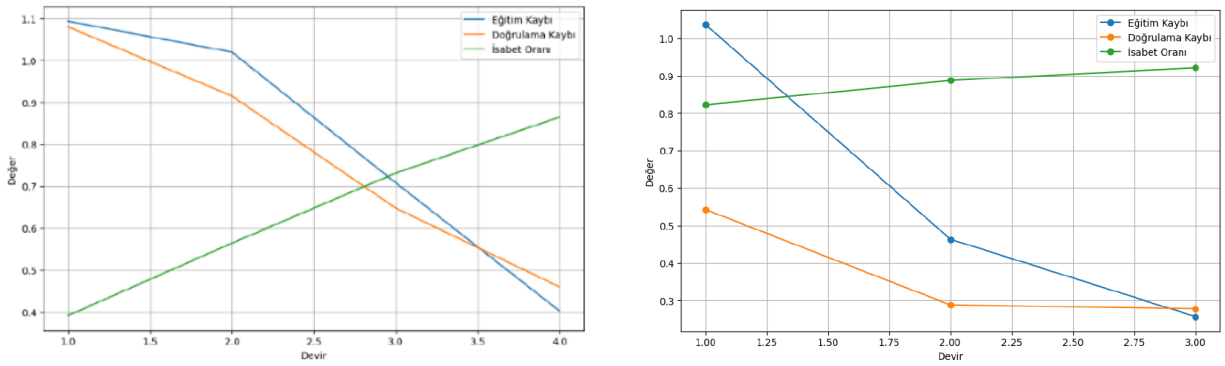
4 Sonuçlar

Bu çalışmada, Türkçe sosyal medya verileri üzerinde tehdit ve hakaret içeriklerinin tespiti amacıyla derin öğrenme ve makine öğrenmesi tabanlı yaklaşımlar karşılaştırılmıştır. LSTM ve BERT gibi derin öğrenme modelleri ile DVS, RO ve GAS gibi makine öğrenmesi modelleri değerlendirilmiştir. Elde edilen bulgular, BERT modelinin diğer modellere kıyasla daha yüksek doğruluk ve genelleme kapasitesine sahip olduğunu göstermektedir. Özellikle BERT modeli, karmaşık dil yapılarının anlaşılmasında ve tehdit içeriklerinin başarılı bir şekilde tespit edilmesinde üstün performans sergilemiştir.

Şimdiye kadar, Türkçe için yapılan çalışmalar büyük ölçüde nefret söylemi ve taciz içeriklerinin tespitine odaklanmıştır. Ancak, tehdit ve hakaret içeriklerinin tespiti bu alanda yeterince ele alınmamıştır. Çalışmamız, bu eksikliği gidermek amacıyla, tehdit ve hakaret içerikli metinlerin otomatik sınıflandırılması için bir altyapı sunmayı hedeflemiştir. Bu tür içeriklerin doğru bir şekilde sınıflandırılması, bireylerin ve toplulukların haklarının korunması için büyük önem taşımaktadır. Ayrıca, sosyal medya platformlarında suç teşkil eden içeriklerin hızlı tespiti, hukuki süreçlere destek sağlayacak ve hukuk sisteminin işleyişini kolaylaştıracaktır.

Gelecek çalışmalar için, tehdit ve hakaret içerikli mesajların tüzel kişilere yönelik olup olmadığını belirleyecek mekanizmaların geliştirilmesi önerilmektedir. Bunun yanı sıra, daha büyük ve dengeli veri setleriyle yapılacak deneyler, model performansını daha da artırabilir. Varlık İsmi Tanıma (NER) gibi yöntemlerin modellerle entegre edilmesi, hukuki metinlerin daha güvenilir bir şekilde sınıflandırılmasına katkı sunacaktır. Ayrıca, veri setinin genişletilmesi ve çeşitlendirilmesiyle birlikte, modelin dil işleme kapasitesi güçlendirilecektir.

Gelecekte, bu alandaki çalışmaların daha geniş alanlara yayılması ve farklı dil modelleriyle geliştirilmesi, yapay zekanın hukuk uygulamalarında yaygın kullanımına katkı sunacaktır. Yapay zeka destekli sistemlerin, dijital ortamdaki suç unsurlarını hızlı ve doğru şekilde tespit etmesi, hukuk sisteminin işleyişine önemli faydalar sağlayacak ve toplum düzeninin korunmasında etkili olacaktır.



(a) (b)

Şekil 5. LSTM (a) ve BERT (b) modellerinin eğitim kaybı, doğrulama kaybı ve isabet (doğruluk) oranı.

Çıkar çatışması

Yazarlar, araştırmanın yürütülmesi, sonuçların değerlendirilmesi ve yayınlanma sürecinde herhangi bir çıkar çatışması olmadığını ve bu çalışmadan maddi veya kişisel bir kazanç elde edilmediğini beyan eder.

Benzerlik oranı (iThenticate): %7

Kaynaklar

[1] İ. Mayda, B. Diri ve T. Yıldız. Türkçe tweetler üzerinde makine öğrenmesi ile nefret söylemi tespiti. Avrupa Bilim Ve Teknoloji Dergisi (24), 328-334. https://doi.org/10.31590/ejosat.903854

[2] Türkiye Cumhuriyeti, 5237 sayılı Türk Ceza Kanunu. Resmî Gazete, 12 Ekim 2004, sayı 25611.

[3] A. Bozyiğit, S. Utku, and E. Nasiboğlu. Cyberbullying detection by using artificial neural network models. 4th International Conference on Computer Science and Engineering (UBMK), 2019. https://doi.org/10.1109/UBMK.2019.8907118

[4] H. Karayigit, C. Aci, and A. Akdagli. Detecting abusive instagram comments in Turkish using convolutional neural network and machine learning methods. Expert Systems with Applications, 114802, 2021. https://doi.org/10.1016/J.ESWA.2021.114802

[5] S. Küçük, and İ. Şahin. Facebook zorbalığı ve mağduriyeti ölçeklerinin Türkçeye uyarlanması, Eğitim ve Bilim Dergisi, vol. 40, no. 178, pp. 1-15, 2015.

[6] O. Kaynar, B. A. Erdoğan, and M. Kaya. Makine öğrenmesi ve öznitelik seçim yöntemleriyle saldırı tespiti, Bilgisayar Mühendisliği ve Bilimleri Dergisi, 11 (2), 56-68, 2018.

[7] F. Beyhan, B. Çarık, İ. Arın, A. Terzioğlu, B. Yanikoglu ve R. Yeniterzi. A Turkish hate speech dataset and detection system. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 4177–4185, Marseille, France. European Language Resources Association, 2022.

[8] A. Najafi ve O. Varol. Turkish hate speech detection online with TurkishBERTweet. In Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024), pages 185–189, St. Julians, Malta. ACL, 2024.

[9] S. Dehghan ve B. Yanikoglu. Evaluating ChatGPT's ability to detect hate speech in Turkish tweets. In Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024), pages 54–59, St. Julians, Malta, ACL, 2024.

[10] Ş. Bayrak, A. Karaca, F. Toson, A. Kocabey ve F. B. Arslanoglu. Detection of hate speech in Turkish social media posts with BERT-Base model. 31st Signal Processing and Communications Applications Conference (SIU), pp. 1-4, Istanbul, Turkiye, 2023.

[11] Z. M. Husunbeyi, D. Akar ve A. Ozgur. Identifying hate speech using neural networks and discourse analysis techniques. In Proceedings of the First Workshop on Language Technologies and Resources for a Fair, Inclusive, and Safe Society within the 13th Language Resources and Evaluation Conference, pp. 32–41, Marseille, France. European Language Resources Association, 2022.

[12] S. Özar. Türk hukukunda nefret suçlarına Avrupa güvenlik ve iş birliği teşkilatı taahhütleri çerçevesinde-genel bir bakış. TAAD, (1009196), 2021. https://doi.org/10.54049/taad.1009196

[13] Ç. Çöltekin, Ç. A Corpus of Turkish offensive language on social media. Proceedings of the 12th Language Resources and Evaluation Conference (LREC), 2020.

[14] Kemik Doğal Dil İşleme Grubu Veri Kümeleri. http://www.kemik.yildiz.edu.tr/veri_kumelerimiz.html , Erişildi 11 Eylül 2024.

[15] A. Akın, zemberek-nlp. https://github.com/ahmetaa/zemberek-nlp, Erişildi 11 Eylül 2024.

[16] C. Cortes, V. Vapnik. Support-vector networks. Machine Learning, pp. 273–297, 1995 https://doi.org/10.1007/BF00994018

[17] J. Yang, A.J. Awan, & G. Vall-llosera. Support vector machines on noisy intermediate-scale quantum computers. 2019, *ArXiv, abs/1909.11988*.

[18] L. Breiman. Random Forests. Machine Learning. 45. 5-32, 2001.

[19] T. Zhang. Improving convection trigger functions in deep convective parameterisation schemes using machine learning, Journal of Advances in Modeling Earth Systems, vol. 13, no. 5, 2021.

[20] F. Jerome. Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics. 29, 2000. 10.1214/aos/1013203451.

[21] S. Hochreiter and J. Schmidhuber. (1997). Long short-term memory. Neural Computation, 9(8), pp. 1735-1780, 1997.

[22] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT, 2019.