T.C.
İSTANBUL GEDİK ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

SOSYAL MEDYADAKİ NEFRET SÖYLEMİ İÇEREN YAYINLARIN TESPİTİNDE YAPAY ZEKA TEMELLİ MAKİNE ÖĞRENMESİ ALGORİTMALARININ PERFORMANS DEĞERLENDİRMESİ

YÜKSEK LİSANS TEZİ

Kadir TURGUT

Yapay Zeka Mühendisliği Anabilim Dalı

Yapay Zeka Mühendisliği Tezli Yüksek Lisans Programı

ARALIK 2023
İSTANBUL

T.C.
İSTANBUL GEDİK ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

SOSYAL MEDYADAKİ NEFRET SÖYLEMİ İÇEREN YAYINLARIN TESPİTİNDE YAPAY ZEKA TEMELLİ MAKİNE ÖĞRENMESİ ALGORİTMALARININ PERFORMANS DEĞERLENDİRMESİ

YÜKSEK LİSANS TEZİ

Kadir TURGUT
210039016
0000-0002-8577-0500

Yapay Zeka Mühendisliği Anabilim Dalı

Yapay Zeka Mühendisliği Tezli Yüksek Lisans Programı

Tez Danışmanı: Dr. Öğr. Üyesi Aytaç Uğur YERDEN

İstanbul 2023

T.C.
İSTANBUL GEDİK ÜNİVERSİTESİ
Lisansüstü Eğitim Enstitüsü Müdürlüğü

Jüri Tez Onay Formu

11.12.2023

LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ MÜDÜRLÜĞÜ

Bu çalışma 11.12 2023 tarihinde aşağıdaki jüri tarafından Yapay Zeka Mühendisliği Anabilim Dalı, Yapay Zeka Mühendisliği (Tezli Yüksek Lisans) Programı Yüksek Lisans Tezi olarak kabul edilmiştir.

TEZ JÜRİSİ

Dr. Öğr. Üyesi Aytaç Uğur YERDEN
Danışman
İstanbul Gedik Üniversitesi

Dr. Öğr. Üyesi Özgür YURTSEVER
Üye (İmza)
İstanbul Gedik Üniversitesi

Dr. Öğr. Üyesi Onur AKAR
Üye (İmza)
Marmara Üniversitesi

YEMİN METNİ

Yüksek Lisans Tezi olarak sunduğum "Sosyal Medyadaki Nefret Söylemi İçeren Yayınların Tespitinde Yapay Zeka Temelli Makine Öğrenmesi Algoritmalarının Performans Değerlendirmesi" başlıklı bu çalışmanın, bilimsel ahlak ve geleneklere uygun şekilde tarafımdan yazıldığını, bu tezdeki bütün bilgileri akademik ve etik kurallar içinde elde ettiğimi, yararlandığım eserlerin tamamının kaynaklarda gösterildiğini ve çalışmamın içinde kullanıldıkları her yerde bunlara atıf yapıldığını, patent ve telif haklarını ihlal edici bir davranışımın olmadığını belirtir ve bunu onurumla doğrularım (11/12/2023).

Kadir TURGUT

ÖNSÖZ

Bu çalışmanın tüm süreçlerinde yol gösteren değerli tez danışmanım Dr. Öğretim Üyesi Aytaç Uğur YERDEN'e ve desteğini her zaman hissettiğim sevgili aileme teşekkür ederim.

Aralık 2023 Kadir TURGUT

İÇİNDEKİLER

Sayfa

KISALTMALAR

BERT : Transformatörlerden Çift Yönlü Kodlayıcı Temsilleri (Bidirectional Encoder Representations from Transformers)

CNN : Evrişimsel Sinir Ağları (Convolutional Neural Networks)

DT : Karar Ağaçları (Decision Trees)

KNN : K-En Yakın Komşu (K-Nearest Neighbours)

LR : Lojistik Regresyon (Logistic Regression)

LSTM : Uzun Kısa Dönemli Bellek (Long Short-Term Memory)

ML : Makine Öğrenmesi (Machine Learning)

MLP : Çok Katmanlı Algılayıcılar (Multilayer Perceptron)

MNB : Çok Terimli Naif Bayes (Multinomial Naive Bayes)

NLP : Doğal Dil İşleme (Natural Language Processing)

RF : Rastgele Orman (Random Forest)

SVM : Destek Vektör Makineleri (Support Vector Machines)

TF-IDF : Terim Frekansı - Tersine Doküman Frekansı (Term Frequency – Inverse Document Frequency)

ÇİZELGE LİSTESİ

Sayfa

ŞEKİL LİSTESİ

Sayfa

SOSYAL MEDYADAKİ NEFRET SÖYLEMİ İÇEREN YAYINLARIN TESPİTİNDE YAPAY ZEKA TEMELLİ MAKİNE ÖĞRENMESİ ALGORİTMALARININ PERFORMANS DEĞERLENDİRMESİ

ÖZET

Sosyal medya üzerindeki nefret söylemi tespiti, insanlar ve topluluklar üzerinde olumsuz etkileri önlemek ve bu tür içerikleri kaldırmak için büyük öneme sahiptir. Ancak nefret söylemi tespiti, dilbilimsel ve kültürel çeşitlilik nedeniyle karmaşık ve zorlu bir süreçtir. Bu nedenle, güçlü ve etkili makine öğrenmesi algoritmaları geliştirmek önemlidir. Geleneksel yöntemlerle bu tür içeriklerin tespiti zaman alıcı ve maliyetli olabileceğinden, yapay zeka temelli makine öğrenmesi algoritmalarının bu konuda büyük bir potansiyel taşıdığı belirtilmektedir.

Bu çalışmanın amacı, sosyal medyadaki nefret söylemi içeren yayınların tespitinde kullanılan yapay zeka temelli makine öğrenmesi algoritmalarının performansını değerlendirmektir. Çalışma, sosyal medya platformlarında nefret söylemini tespit etme ve yönetme sorununa odaklanmaktadır. Bu çalışmada, farklı algoritmaların performanslarını karşılaştıracak ve en uygun yöntemleri belirleyecektir. Ayrıca, veri kümesi ve özellik çıkarımı yöntemlerinin algoritma performansı üzerindeki etkileri analiz edilecektir. Algoritmalar genellikle doğal dil işleme tekniklerine dayanır ve metinlerdeki özellikleri öğrenerek nefret söylemini tespit etmeye çalışır. Bu algoritmaların performansı, dil, kültür, kullandıkları öznitelikler ve eğitim veri kümesi gibi faktörlere bağlı olarak değişebilir, bu nedenle kapsamlı bir analiz gereklidir. Araştırmada, nefret söylemi tespitinde kullanılan algoritmaların performansı, veri kümesi ve özellik çıkarımı yöntemleriyle karşılaştırılmıştır. Bu süreçte algoritmaların dil ve kültürlerarası etkinliği, özellik seçimi ve temsilini, yanlış pozitif ve yanlış negatif oranlarını ve genel doğruluklarını analiz edilecektir.

Bu çalışma, algoritmaların yanlış pozitif ve yanlış negatif oranlarını düşürmek için sürekli iyileştirme ve optimizasyon çalışmalarının önemini vurgulamaktadır. Çalışmada, algoritmaların dil ve kültürlerarası etkinliği, özellik seçimi ve temsilini, yanlış pozitif ve yanlış negatif oranlarını ve genel doğruluklarını analiz edilecek ve bu analiz, nefret söylemi tespiti için en uygun ve etkili yöntemleri belirlemeye yardımcı olacaktır.

Sonuç olarak bu çalışma nefret söyleminin yayılmasını önlemek ve sosyal medya ortamlarını daha güvenli hale getirmek için önemli bir adım olarak görülebilir.

Anahtar Kelimeler: *Sosyal Medya, Nefret Söylemi, Yapay Zeka*

PERFORMANCE EVALUATION OF ARTIFICIAL INTELLIGENCE BASED MACHINE LEARNING ALGORITHMS IN DETECTING PUBLICATIONS CONTAINING HATE SPEECH ON SOCIAL MEDIA

ABSTRACT

The detection of hate speech on social media holds significant importance in preventing adverse effects on individuals and communities, as well as in removing such content. However, hate speech detection is a complex and challenging process due to linguistic and cultural diversity. Therefore, it is crucial to develop robust and effective machine learning algorithms. Given that the detection of such content through traditional methods can be time-consuming and costly, there is an indication that artificial intelligence-based machine learning algorithms carry substantial potential in this regard.

The objective of this study is to evaluate the performance of artificial intelligence-based machine learning algorithms employed in the detection of publications containing hate speech on social media. The research focuses on addressing the issue of detecting and managing hate speech on social media platforms. In this study, different algorithms will be compared in terms of their performances, aiming to identify the most suitable methods. Additionally, the impact of dataset and feature extraction methods on algorithm performance will be analyzed. These algorithms typically rely on natural language processing techniques, learning features in texts to identify hate speech. The performance of these algorithms may vary based on factors such as language, culture, features employed, and training dataset; hence, a comprehensive analysis is necessary.

In the research, the performance of algorithms used for hate speech detection is compared with dataset and feature extraction methods. The cross-cultural and cross-linguistic effectiveness of algorithms, feature selection and representation, false positive and false negative rates, and overall accuracy will be analyzed. This study underscores the importance of continuous improvement and optimization efforts to reduce false positive and false negative rates of algorithms. The analysis of cross-cultural and cross-linguistic effectiveness, feature selection and representation, false positive and false negative rates, and overall accuracy will help determine the most suitable and effective methods for hate speech detection.

In conclusion, this study can be perceived as a crucial step in preventing the spread of hate speech and enhancing the safety of social media environments.

Keywords: *Social Media, Hate Speech, Artificial Intelligence*

1. GİRİŞ

Sosyal medya, insanların düşüncelerini, fikirlerini ve duygularını paylaşma konusunda büyük bir özgürlük sunar. Bu platformlar, insanlar arasında iletişimi ve bilgi alışverişini kolaylaştırarak toplumsal fayda sağlamaktadır (Kaplan ve Haenlein, 2010). Ancak, sosyal medyanın yaygın kullanımıyla birlikte, nefret söylemi gibi olumsuz etkiler de ortaya çıkmıştır. Nefret söylemi, belirli bir grup, topluluk veya bireylere yönelik hoşgörüsüzlük ve düşmanlık içeren ifadelerdir (Allport, 1954). Bu tür ifadeler, çeşitli toplumlar ve ülkelerde ayrımcılığa, şiddete ve sosyal gerilime yol açabilir (Perry, 2001).

Nefret söylemi içeren yayınların tespiti ve yönetimi, sosyal medya platformları için önemli bir sorun haline gelmiştir (Chetty ve Alathur, 2018). Geleneksel yöntemlerle (ör. moderatörler tarafından manuel olarak inceleme), bu tür yayınların tespiti ve yönetimi zaman alıcı ve maliyetli olabilir (Davidson ve diğ., 2017). Bu nedenle, yapay zeka temelli makine öğrenmesi algoritmaları, sosyal medyadaki nefret söylemini otomatik olarak tespit etmek ve yönetmek için büyük potansiyele sahiptir (Schmidt ve Wiegand, 2017).

Bu çalışmada, sosyal medyadaki nefret söylemi içeren yayınların tespitinde kullanılan yapay zeka temelli makine öğrenmesi algoritmalarının performans değerlendirmesi ele alınmaktadır. Çeşitli algoritmaların performansı, nefret söylemi tespiti için kullanılan veri kümesi ve özellik çıkarımı yöntemleri ile karşılaştırılacaktır (Fortuna ve Nunes, 2018).

Birçok çalışma, nefret söyleminin tespitinde sınıflandırma algoritmaları kullanarak başarılı sonuçlar elde etmiştir (Djuric ve diğ., 2015; Founta ve diğ., 2018). Özellikle, derin öğrenme yöntemlerinin (ör. evrişimli sinir ağları ve tekrarlayan sinir ağları) doğal dil işleme alanında önemli gelişmelere yol açtığı bilinmektedir (Young ve diğ., 2018). Bu yöntemlerin nefret söylemi tespitindeki etkinliği, literatürdeki önemli bir araştırma konusudur (Sap ve diğ., 2019).

Bu tezde, makine öğrenmesi algoritmalarının sosyal medyadaki nefret söylemi tespiti üzerindeki performanslarını değerlendirmek için öncelikle, algoritmaların çalışma prensipleri ve uygulanabilirliği incelenmiştir. Yapılan literatür taramasında, en çok kullanılan algoritmaların Destek Vektör Makineleri (SVM), Naive Bayes, Karar Ağaçları ve Rastgele Ormanlar olduğu görülmüştür (Gao ve diğ., 2017; Zhang ve diğ., 2018). Ayrıca, derin öğrenme yöntemleri, özellikle evrişimli sinir ağları (CNN) ve tekrarlayan sinir ağları (RNN), nefret söylemi tespitinde önemli başarılar göstermiştir (Gambäck ve Sikdar, 2017; Saha ve diğ., 2018).

Veri kümesi ve özellik çıkarımı yöntemlerinin algoritma performansı üzerindeki etkisini incelemek adına, çeşitli çalışmalarda kullanılan veri kümeleri ve özellik çıkarım yöntemleri değerlendirilmiştir. Özellik çıkarımı yöntemlerine örnek olarak, kelime frekansları, terim frekansı–ters belge frekansı (TF-IDF) ve Word2Vec gibi temsiller verilebilir (Mikolov ve diğ., 2013; Salminen ve diğ., 2018).

Bu tezin amacı, sosyal medyadaki nefret söylemi içeren yayınların tespitinde kullanılan yapay zeka temelli makine öğrenmesi algoritmalarının performans değerlendirmesini gerçekleştirmektir. Bu bağlamda, farklı algoritmaların performansları karşılaştırılacak ve en uygun yöntemler belirlenecektir. Ayrıca, veri kümesi ve özellik çıkarımı yöntemlerinin algoritma performansı üzerindeki etkisi analiz edilecektir.

Sosyal medya, günümüzde bireylerin düşüncelerini ve duygularını ifade etmeleri için önemli bir platform haline gelmiştir. Bununla birlikte, nefret söylemi ve ayrımcılığın yayılması da sosyal medya üzerinden hızla gerçekleşmektedir. Nefret söylemi, belirli bir kişi veya gruplara karşı nefret, aşağılama veya şiddeti teşvik eden sözler, görseller veya diğer iletişim şekilleri olarak tanımlanabilir (Brown, 2018). Bu çalışmada, sosyal medyadaki nefret söylemi içeren yayınların tespitinde yapay zeka temelli makine öğrenmesi algoritmalarının performans değerlendirmesi üzerine odaklanacağız.

Nefret söyleminin sosyal medya üzerinde tespiti, insanlar ve topluluklar üzerinde olumsuz etkileri önlemek ve bu tür içerikleri kaldırmak için büyük öneme sahiptir (ElSherief ve diğ., 2018). Ancak, nefret söylemi tespiti, dilbilimsel ve kültürel çeşitlilik nedeniyle karmaşık ve zorlu bir süreçtir (Waseem ve Hovy, 2016). Bu nedenle, otomatik nefret söylemi tespiti için güçlü ve etkili makine öğrenmesi algoritmaları geliştirmek önemlidir.

Yapay zeka temelli makine öğrenmesi algoritmaları, nefret söylemini otomatik olarak tespit etmek için kullanılabilir (Schmidt ve Wiegand, 2017). Bu algoritmalar, genellikle doğal dil işleme (NLP) tekniklerine dayanır ve metinlerdeki özellikleri öğrenerek nefret söylemini tespit etmeye çalışır (Zhang ve Luo, 2018). Bu yöntemler arasında, destek vektör makineleri (SVM), karar ağaçları, Bayes sınıflandırıcıları ve derin öğrenme gibi çeşitli makine öğrenmesi teknikleri bulunmaktadır (Badjatiya ve diğ., 2017).

Ancak, bu algoritmaların performansı, dil, kültür, kullandıkları öznitelikler ve eğitim veri kümesi gibi faktörlere bağlı olarak değişir (Davidson ve diğ., 2017). Bu nedenle, bu algoritmaların performansını değerlendirmek ve en uygun yöntemi seçmek için kapsamlı bir analiz gereklidir. Ayrıca, algoritmaların yanlış pozitif ve yanlış negatif oranlarını düşürmek için sürekli iyileştirme ve optimizasyon çalışmaları yapılması önemlidir (Fortuna ve Nunes, 2018).

Bu tezde, sosyal medyadaki nefret söylemi içeren yayınların tespitinde kullanılan yapay zeka temelli makine öğrenmesi algoritmalarının performans değerlendirmesini gerçekleştireceğiz. Bu süreçte, algoritmaların dil ve kültürlerarası etkinliğini, özellik seçimi ve temsilini, yanlış pozitif ve yanlış negatif oranlarını ve genel doğruluklarını analiz edeceğiz. Bu analiz, nefret söylemi tespiti için en uygun ve etkili yöntemleri belirlememize yardımcı olacaktır.

Sonuç olarak, sosyal medyadaki nefret söylemi içeren yayınların tespitinde yapay zeka temelli makine öğrenmesi algoritmalarının performans değerlendirmesine odaklanan bu tez, nefret söyleminin yayılmasını önlemek ve sosyal medya ortamlarını daha güvenli hale getirmek için önemli bir adım olarak görülebilir.

2. KAVRAMSAL ÇERÇEVE

2.1 Sosyal Medya

Sosyal medya, son yıllarda hızla gelişen ve insanlar arasındaki iletişim biçimini büyük ölçüde değiştiren bir fenomendir. Bu yazıda, sosyal medyanın tanımı, tarihsel gelişimi, kullanım alanları ve etkilerine dair akademik kaynaklardan yola çıkarak detaylı ve uzun bir açıklama yapılacaktır.

Sosyal medya, temel olarak internet üzerinden paylaşılan bilgi, fikir, düşünce, görsel ve diğer içeriklerle insanlar arasında etkileşim ve iletişim sağlayan platformlar ve uygulamalar olarak tanımlanabilir (Kaplan ve Haenlein, 2010). Sosyal medyanın tarihsel gelişimine bakacak olursak, ilk sosyal medya platformlarından biri olarak 1997 yılında kurulan Six Degrees'i sayabiliriz (Boyd ve Ellison, 2007). 2000'li yılların başında ise Friendster, MySpace ve LinkedIn gibi platformlar ortaya çıkmış ve sosyal medya kullanımı hızla yaygınlaşmıştır. 2006 yılında kurulan Twitter ve 2004 yılında kurulan Facebook, sosyal medyanın gelişiminde önemli dönüm noktalarıdır (Boyd ve Ellison, 2007).

Sosyal medya platformlarının kullanım alanları oldukça çeşitlidir ve sadece kişisel iletişimle sınırlı değildir. İşletmeler ve markalar, sosyal medya platformlarını pazarlama ve müşteri ilişkileri yönetimi için kullanmaktadırlar (Kaplan ve Haenlein, 2010). Ayrıca, sosyal medya politikacılar, aktivistler ve sivil toplum örgütleri tarafından da toplumsal ve politik kampanyaları yaymak ve destekçi toplamak için kullanılmaktadır (Castells, 2012).

Sosyal medyanın etkileri üzerine yapılan akademik çalışmalar da oldukça geniştir. Sosyal medyanın olumlu etkilerine örnek olarak, insanlar arasında sınırsız ve hızlı iletişim imkanı sunması, bilgi paylaşımını kolaylaştırması ve düşüncelerin çeşitlenmesine katkıda bulunması gösterilebilir (Ellison ve diğ., 2007). Ancak sosyal medyanın olumsuz etkileri de bulunmaktadır. Özellikle gençler üzerinde yapılan çalışmalar, sosyal medya kullanımının bağımlılık, sosyal izolasyon ve özgüven düşüklüğü gibi psikolojik sorunlara yol açabileceğini göstermektedir (Kuss ve

Griffiths, 2011). Ayrıca, sosyal medya platformlarında yayılan yanlış bilgiler ve kışkırtıcı içeriklerin toplumsal kutuplaşma ve ayrışmaya yol açabileceği üzerine de çalışmalar mevcuttur (Allcott ve Gentzkow, 2017).

Sosyal medya ve eğitim alanındaki ilişki de önemli bir konudur. Öğretmenler ve öğrenciler arasında etkili bir iletişim sağlayarak, öğrenme süreçlerini zenginleştirebilir ve sınıf dışında da öğrenme fırsatları sunabilir (Greenhow ve diğ., 2009). Bununla birlikte, sosyal medya kullanımının dikkat dağıtıcı etkisi ve akademik başarı üzerinde olumsuz bir etkiye sahip olabileceği de belirtilmiştir (Junco, 2012).

Sosyal medyanın gizlilik ve güvenlik ile ilgili konuları da akademik çevrelerde tartışılmaktadır. Kişisel verilerin korunması ve özel bilgilerin kötüye kullanılması riskleri, sosyal medya platformlarının gizlilik politikaları ve güvenlik önlemleriyle ilgili endişeleri gündeme getirmektedir (Fogel ve Nehmad, 2009). Bu bağlamda, sosyal medya kullanıcılarının gizlilik ve güvenlik konularında bilinçli olmaları ve platformlarda paylaştıkları bilgiler konusunda dikkatli davranmaları önem taşımaktadır (Debatin ve diğ., 2009).

Sonuç olarak, sosyal medya günümüzde iletişim, bilgi paylaşımı, eğitim, politika ve pazarlama gibi pek çok alanda etkili bir araç haline gelmiştir. Bununla birlikte, olumlu etkilerinin yanı sıra, sosyal medyanın olumsuz etkileri ve riskleri de göz ardı edilmemelidir. Bu nedenle, sosyal medya kullanımının dengeli ve bilinçli bir şekilde yapılması önemlidir.

2.2 Nefret Söylemi

Nefret söylemi, bireylerin veya grupların dil yoluyla aşağılanması, dışlanması ve tehdit edilmesi olarak tanımlanabilir (Brown, 2017). Bu tür söylemler, toplumların sosyal dokusunu bozarak, ayrımcılığı ve hoşgörüsüzlüğü körükler ve insan haklarını ihlal eder (Gelber ve McNamara, 2016).

Nefret söyleminin tarihsel kökenleri, toplumların etnik, dini ve kültürel çeşitlilik gösterdiği dönemlere kadar uzanır (Perry, 2001). Bu tür söylemler, baskın grupların güçlerini sürdürmek ve azınlıkları sindirmek için kullandığı bir araç olarak işlev görmüştür (Richardson, 2010).

Nefret söylemi kavramı, sosyal bilimlerde farklı teorik çerçevelerle ele alınmıştır. Örneğin, sosyal kimlik teorisi (Tajfel ve Turner, 1986) ve gerçekçi grup çatışması teorisi (Sherif, 1966), insanların diğer grupları aşağılamaya ve düşmanca tutumlar sergilemeye yönlendiren süreçleri açıklar. Ayrıca, dil ve güç ilişkisine odaklanan kritik dilbilim çalışmaları (Fairclough, 1989; van Dijk, 1993), nefret söyleminin nasıl yapılandığı ve toplumlar üzerindeki etkilerini ortaya koymaktadır.

Nefret söyleminin olumsuz etkileri, bireysel ve toplumsal düzeyde kendini gösterir (Bleich, 2011). Bireysel düzeyde, nefret söylemine maruz kalan bireylerde travma, özsaygı kaybı ve güvende hissetmeme gibi psikolojik sorunlar ortaya çıkabilir (Möller ve Krahé, 2009). Toplumsal düzeyde ise, nefret söylemi ayrımcılık, hoşgörüsüzlük ve şiddet olaylarının artmasına yol açarak sosyal uyumu tehdit eder (Green, McFalls ve Smith, 2001).

2.2.1 Nefret söylemine karşı mücadele yöntemleri

Nefret söylemine karşı mücadelede, yasal düzenlemeler, eğitim ve farkındalık çalışmaları ve medya okuryazarlığı önemli rol oynar (Waldron, 2012; Branković, 2018).

Nefret söylemini önlemeye yönelik yasal düzenlemeler, ifade özgürlüğü ile nefret söyleminin yarattığı zararlar arasında denge kurmaya çalışır (Gelber, 2013). Birçok ülke, nefret söylemini suç olarak kabul eden ve bu tür eylemlere cezaî yaptırım uygulayan yasalar çıkarmıştır (Bleich, 2011). Ancak, bu tür yasaların etkinliği ve uygunluğu, ifade özgürlüğü ve insan hakları açısından tartışmalıdır (Heinze, 2016).

Nefret söylemine karşı eğitim ve farkındalık çalışmaları, insanların hoşgörü ve empati becerilerini geliştirmeyi amaçlar (Banks, 2006). Okullarda yapılan çeşitlilik ve ayrımcılık karşıtı eğitim programları, öğrencilerin diğer kültürler ve yaşamlar hakkında bilgi sahibi olmalarına ve hoşgörüyü benimsemelerine yardımcı olabilir (Pettigrew ve Tropp, 2006).

Medya okuryazarlığı, nefret söyleminin yayılmasını önlemek ve insanların bu tür söylemlere karşı dirençli hale gelmelerine katkı sağlar (Hobbs, 2010). Medya okuryazarlığı eğitimi, bireylerin haber ve bilgi kaynaklarını sorgulama, analiz etme ve değerlendirme becerilerini geliştirerek, nefret söylemine karşı bilinçli ve eleştirel bir tutum sergilemelerine yardımcı olur (Livingstone, 2004).

2.2.2 İnternet ve sosyal medya üzerinde nefret söylemi

İnternet ve sosyal medya platformları, nefret söyleminin yayılmasında önemli bir araç haline gelmiştir (Bartlett ve Krasodomski-Jones, 2015). Bu platformlar, anonimlik ve küresel erişim sağlayarak, nefret söylemlerinin daha hızlı ve geniş kitlelere ulaşmasına olanak tanır (Keats Citron, 2014). Ayrıca, sosyal medya algoritmalarının kullanıcıların ilgi alanlarına göre içerik sunma eğilimi, nefret söylemlerinin hedef kitlelere daha kolay ulaşmasına ve bu tür içeriklerin daha fazla paylaşılmasına neden olabilir (Tufekci, 2018).

İnternet ve sosyal medya üzerinde nefret söylemine karşı mücadelede, platformların sorumluluğu ve kullanıcıların bilinçlendirilmesi önemlidir (Gagliardone ve diğ., 2015). Örneğin, platformlar, kullanıcıların nefret söylemi içeren içerikleri bildirebilmesi ve bu tür içeriklerin kaldırılması için politikalar ve mekanizmalar geliştirebilir (Suzor ve diğ., 2019). Ayrıca, kullanıcılar, nefret söylemi içeren içeriklerle karşılaştıklarında nasıl tepki verecekleri ve bu tür içerikleri nasıl rapor edecekleri konusunda eğitilebilir (Matamoros-Fernández, 2017).

2.2.3 Yeni teknolojilerin rolü

Yapay zeka, büyük veri analizi ve otomatik dil işleme gibi yeni teknolojilerin, nefret söylemi tespiti ve analizi için kullanılması üzerine araştırmalar yapılabilir (Schmidt ve Wiegand, 2017). Bu tür teknolojiler, nefret söyleminin izlenmesi ve kaldırılması süreçlerini hızlandırarak, bu tür içeriklerin yayılmasını ve etkilerini sınırlamada etkili olabilir.

2.2.4 Uygulamalar ve politika önerileri

Nefret söylemiyle mücadelede etkili uygulamalar ve politika önerileri geliştirmek için, araştırmaların bulguları ve teorik çerçeveleri dikkate alınmalıdır. İşte bu bağlamda bazı öneriler:

Nefret söylemiyle mücadelede yasal düzenlemelerin, ifade özgürlüğü ve insan haklarıyla uyumlu olmasına dikkat edilmelidir. Mevcut yasaların etkinliği ve uygulanabilirliği düzenli olarak değerlendirilmeli ve gerektiğinde güncellenmelidir (Heinze, 2016).

Okullarda ve toplum merkezlerinde nefret söylemi ve ayrımcılığa karşı eğitim programları düzenlenmeli ve desteklenmelidir. Bu tür programlar, bireylerin hoşgörü

ve empati becerilerini geliştirerek, nefret söyleminin toplumsal etkilerini azaltmaya katkı sağlayabilir (Banks, 2006).

Bireylerin medya okuryazarlığı becerilerinin geliştirilmesi ve doğru bilgiye erişimin kolaylaştırılması, nefret söylemine karşı dirençli hale gelmelerine yardımcı olabilir (Livingstone, 2004).

İnternet ve sosyal medya platformlarının, nefret söylemiyle mücadelede daha etkin ve sorumlu olmaları teşvik edilmelidir. Bu amaçla, platformlar kullanıcıları için şeffaf ve hızlı içerik bildirme ve kaldırma süreçleri oluşturmalı ve bu süreçleri sürekli iyileştirmelidir (Suzor ve diğ., 2019).

2.3 Yapay Zeka

Yapay Zeka (YZ), genel olarak, insan zekâsının özelliklerini ve işlevlerini modelleyen ve simüle eden bilgisayar sistemlerinin çalışmasıdır (Russell ve Norvig, 2016). Bu kapsamlı bir alan olup, öğrenme, problem çözme, algılama, dil anlama ve daha birçok alanı içerir (Nilsson, 1998). Yapay zekâ çalışmalarının temel amacı, insan zekâsının bilişsel işlemlerini anlamak ve bilgisayar sistemlerinde bu işlemleri yeniden yaratmaktır (Poole ve Mackworth, 2017).

Yapay zekâ, tarih boyunca birçok evre geçirerek gelişmiştir. İlk aşamada, sembolik mantık temelli yapay zekâ çalışmaları yapılmıştır (Newell ve Simon, 1972). Bu dönemde, bilgisayarlar mantık kurallarını kullanarak problemleri çözmeye çalışmıştır (McCarthy, 1959). Bu süre zarfında, oyun oynayan ve geometri problemleri çözen programlar geliştirilmiştir (Newell ve Simon, 1963).

Daha sonra, bilgisayarların insan gibi öğrenmesini sağlamak için makine öğrenimi (ML) alanı ortaya çıkmıştır (Mitchell, 1997). Bu alanda, yapay sinir ağları ve genetik algoritmalar gibi teknikler kullanılarak bilgisayarlar, verilerden öğrenme yeteneği kazanmıştır (Rumelhart ve McClelland, 1986; Goldberg, 1989). Bu dönemde, tanıma, tahmin ve sınıflandırma gibi görevlerde başarılı sonuçlar elde edilmiştir (LeCun ve diğ., 1989).

Son yıllarda, derin öğrenme adı verilen bir alt dal daha büyük bir öneme sahip olmuştur (Goodfellow ve diğ., 2016). Bu alan, büyük veri kümeleri ve daha güçlü bilgisayar sistemleri sayesinde, yapay sinir ağlarının daha derin katmanlara sahip olmasını ve böylece daha karmaşık görevleri başarıyla tamamlamasını sağlamıştır

(Krizhevsky ve diğ., 2012). Özellikle, doğal dil işleme ve görüntü tanıma gibi alanlarda büyük başarılar elde edilmiştir (Devlin ve diğ., 2018; He ve diğ., 2016).

Yapay zeka alanında yapılan çalışmalar, gittikçe artan başarıları ve sürekli gelişen teknolojiyle, gelecekte daha da büyük etkilere ve uygulama alanlarına sahip olması beklenmektedir. Özellikle, otomasyon, sağlık, eğitim ve sürdürülebilirlik gibi alanlarda yapay zeka teknolojilerinin insan yaşamını daha da iyileştireceği düşünülmektedir (Daugherty ve Wilson, 2018; Ng, 2017).

Yapay zeka ve etik konuları da giderek daha fazla önem kazanmaktadır. Teknolojinin hızlı ilerlemesi, insanlar ve toplum için bazı etik ve sosyal sorunları beraberinde getirebilir. Örneğin, veri gizliliği, algoritmik önyargı ve insanlarla makineler arasındaki iş güvenliği gibi konular üzerinde durulması gereken önemli alanlardır (Cath ve diğ., 2018; Bostrom ve Yudkowsky, 2014).

Ayrıca, yapay zeka ve robotik alanındaki gelişmeler, insansı ve otonom robotların sosyal ve hukuki konularını da gündeme getiriyor. Robotların insanlarla etkileşimi ve onların hakları üzerinde düşünülmesi gereken konular mevcuttur (Darling, 2016; Bryson, 2018).

Yapay zeka ve beyin-bilgisayar arayüzleri (BCI) gibi nöroteknolojilerin gelişimi de, insan zekâsını ve bilincini daha iyi anlamamıza ve geliştirmemize yardımcı olabilir. Bu alandaki çalışmalar, insan beyninin doğrudan bilgisayar sistemlerine bağlanmasını ve böylece zihinsel yeteneklerin iyileştirilmesini veya engellerin aşılmasını sağlar (Nicolelis ve Lebedev, 2009; Wolpaw ve diğ., 2002).

Yapay zeka, yukarıda belirtilen başarıları, uygulama alanları ve gelecekteki potansiyeliyle birlikte, sürekli gelişen ve genişleyen bir alandır. Akademik kaynaklar ve araştırmalar, bu alandaki temel bilgi ve becerilerin geliştirilmesine katkıda bulunarak, yapay zeka teknolojilerinin daha etkili ve sorumlu bir şekilde kullanılmasına yönlendirmektedir. Bu doğrultuda, yapay zeka ve onunla ilişkili alanlarda çalışan araştırmacılar ve mühendisler, bu teknolojilerin insanlık için daha iyi bir gelecek inşa etme potansiyelini en üst düzeye çıkaracak şekilde sürekli ilerleme kaydetmeye çalışmaktadır.

Yapay zeka araştırmalarının önemli bir bileşeni, açık kaynak topluluklarının ve veri paylaşımının rolüdür. Bu tür girişimler, araştırmacıların ve geliştiricilerin yeni teknikler ve yöntemler geliştirmelerine yardımcı olurken, aynı zamanda dünya

çapında işbirliğini ve yenilikçiliği teşvik eder (Bengio ve diğ., 2016; Barret ve diğ., 2020).

Yapay zeka alanında ilerleme kaydedilmesine rağmen, bazı temel zorluklar ve sınırlamalar hâlâ varlığını sürdürmektedir. Örneğin, açıklanabilirlik ve şeffaflık gibi konular, karmaşık yapay zeka sistemlerinin günlük hayatta ve kritik uygulamalarda daha yaygın olarak benimsenmesinin önündeki engellerdir (Guidotti ve diğ., 2018; Arrieta ve diğ., 2020).

Sonuç olarak, yapay zeka alanındaki akademik kaynaklar ve araştırmalar, bu alandaki ilerlemelerin anlaşılmasına ve daha fazla gelişmeye yönlendirmeye katkıda bulunmaktadır. Hem teorik bilgi birikimi hem de uygulama alanlarındaki başarılarla birlikte, yapay zeka teknolojilerinin insan yaşamını ve toplumu dönüştürme potansiyeli giderek daha büyük bir öneme sahip olmaktadır. Bu nedenle, akademik kaynakların ve araştırmaların sürekli güncellenmesi ve paylaşılması, yapay zeka alanındaki gelişmelerin etkili ve sorumlu bir şekilde yönlendirilmesi için önemlidir.

Yapay zeka eğitimi ve öğretimi de büyük önem taşımaktadır. Üniversiteler ve eğitim kurumları, yapay zeka alanında yetenekli öğrenciler yetiştirmeye yönelik dersler ve programlar sunarak bu alandaki büyümeyi ve ilerlemeyi desteklemektedir (Coppola ve diğ., 2020; Minsky ve diğ., 2021).

Ayrıca, yapay zeka alanında yapılan araştırmalar, sürekli olarak yeni paradigmalar ve yaklaşımlar doğurmaktadır. Örneğin, insan benzeri öğrenme ve genel yapay zeka gibi konular, gelecekteki yapay zeka sistemlerinin daha esnek, uyarlanabilir ve özerk olmasını sağlamak için araştırılmaktadır (Lake ve diğ., 2017; Silver ve diğ., 2018).

Yapay zeka alanındaki gelişmelerin ve araştırmaların değerlendirilmesi ve güncellenmesi, bu teknolojilerin insanlık için daha iyi bir gelecek inşa etme potansiyelini en üst düzeye çıkaracak şekilde sürekli ilerleme kaydetmeye yardımcı olacaktır. Bu süreçte, akademik kaynaklar ve çalışmalar, bu alandaki temel bilgi ve becerilerin geliştirilmesine ve uygulanmasına katkıda bulunarak, yapay zeka teknolojilerinin daha etkili ve sorumlu bir şekilde kullanılmasına yönlendirmektedir.

2.4 Makine Öğrenmesi Algoritmaları ve Nefret Söylemi Tespiti

Makine öğrenimi, verilerden otomatik olarak model oluşturarak ve bu modelleri kullanarak öğrenme yeteneğine sahip algoritmaların tasarımı ve geliştirilmesi ile ilgilenen yapay zeka alanının bir alt dalıdır (Mitchell, 1997). Makine öğrenimi algoritmaları temel olarak üç kategoriye ayrılır: denetimli öğrenme (supervised learning), denetimsiz öğrenme (unsupervised learning) ve takviyeli öğrenme (reinforcement learning) (Kelleher, Mac Namee ve D'Arcy, 2015).

Makine öğrenimi alanındaki sürekli gelişmeler ve yenilikler, insan yaşamını ve çalışma şekillerini daha da dönüştürmeye ve farklı endüstrilerde ve uygulama alanlarında önemli başarılar elde etmeye yardımcı olmaktadır. Gelecekteki araştırmaların, mevcut algoritmaların daha da geliştirilmesi ve yeni algoritmaların ortaya çıkması yönünde çalışmalar yaparak makine öğrenimi teknolojisinin daha geniş bir yelpazede kullanılmasına katkıda bulunması beklenmektedir. Bu süreç, teknolojik ilerlemelerin temelini atmaya ve insan yaşamını ve çalışma şekillerini daha da iyileştirmeye devam edecektir.

Nefret söylemi, bireylerin veya grupların ırk, etnik köken, cinsiyet, din veya cinsel yönelim gibi özelliklerine yönelik saldırgan ve ayrımcı dil kullanımıdır. İnternet ve sosyal medya platformları, bu tür dil kullanımının hızla yayılmasına ve artmasına olanak tanımaktadır. Bu nedenle, nefret söyleminin otomatik olarak tespit edilmesi ve engellenmesi, teknolojinin ve yapay zekanın önemli bir uygulama alanı haline gelmiştir. Bu alandaki çalışmalar, farklı makine öğrenmesi algoritmalarının kullanımını içerir.

2.4.1 Denetimli öğrenme algoritmaları

Nefret söylemi tespitinde kullanılan temel yaklaşımlardan biri denetimli makine öğrenmesidir. Bu yöntemde, etiketli veri kümesi kullanılarak bir model eğitilir ve bu model, yeni veriler üzerinde nefret söylemini tespit etmek için kullanılır (Waseem ve Hovy, 2016). Denetimli makine öğrenmesi algoritmaları arasında Naive Bayes, Destek Vektör Makineleri (SVM), Rastgele Ormanlar (RF), Karar Ağaçları (DT) ve Lojistik Regresyon bulunmaktadır (Fortuna ve Nunes, 2018).

2.4.2 Denetimsiz öğrenme algoritmaları

Denetimsiz öğrenme, etiketlenmemiş verilerden yapısal özellikleri ve örüntüleri öğrenmeyi amaçlayan algoritmaları içerir (Hastie, Tibshirani ve Friedman, 2009).

Denetimsiz makine öğrenmesi, etiketli veri eksikliği durumunda kullanılan bir yöntemdir. Bu yaklaşımda, metinlerin yapısal ve anlamsal benzerliklerine dayalı olarak kümeleme algoritmaları kullanılır. K-means, DBSCAN ve hiyerarşik kümeleme, nefret söylemi tespiti için kullanılan denetimsiz makine öğrenmesi algoritmaları arasındadır (Zhao ve Mao, 2020). Denetimsiz yöntemler aynı zamanda, yarı denetimli makine öğrenmesi yaklaşımlarına da temel teşkil edebilir (Zhang ve diğ., 2017). Başlıca denetimsiz öğrenme algoritmaları Kümeleme Algoritmaları ve Boyut Azaltma Algoritmalarıdır.

2.4.3 Takviyeli öğrenme algoritmaları

Takviyeli öğrenme, öğrenme sürecinde ödül ve ceza mekanizmaları kullanarak kararlar alan bir ajanın eylemlerini optimize etmeye çalışır (Sutton ve Barto, 2018). Başlıca takviyeli öğrenme algoritmaları şunlardır: Q-learning (Watkins ve Dayan, 1992), Sarsa (Rummery ve Niranjan, 1994) ve Deep Q-Networks (DQN) (Mnih ve diğ., 2015).

2.4.5 Derin öğrenme (Deep learning)

Derin öğrenme, nefret söylemi tespiti için kullanılan başka bir yaklaşımdır. Derin öğrenme modelleri, özellikle metin verilerinin karmaşık yapıları ve anlamsal ilişkilerini öğrenebilme yetenekleri nedeniyle tercih edilmektedir. Derin öğrenme, karmaşık ve büyük veri kümelerinden özellikler ve örüntüler öğrenmek için derin yapay sinir ağları kullanır (LeCun, Bengio ve Hinton, 2015). Bu alanda yaygın olarak kullanılan modeller arasında Evrişimli Sinir Ağları (Convolutional Neural Networks-CNN) (Zhang ve Wallace, 2015), Uzun Kısa Vadeli Bellek (LSTM) (Tai ve diğ., 2015), Geri Beslemeli Sinir Ağları (Recurrent Neural Networks - RNN) ve Dikkat Mekanizmaları (Attention Mechanisms) bulunmaktadır.

2.4.6 Aktarım öğrenmesi (Transfer learning)

Aktarım öğrenmesi, önceden eğitilmiş bir modelin bilgilerini, yeni ve ilgili görevlere uyarlayarak öğrenme sürecini hızlandırmayı amaçlayan bir yöntemdir. Bu yaklaşım, nefret söylemi tespitinde özellikle BERT (Devlin ve diğ., 2019) ve RoBERTa (Liu ve diğ., 2019) gibi önceden eğitilmiş dil modelleri kullanarak başarı sağlamıştır.

Bu yaklaşım, yeni görevler için daha az etiketli veriyle çalışarak ve eğitim süresini kısaltarak model performansını artırmayı amaçlar (Pan ve Yang, 2010). Özellikle derin öğrenmede, büyük ölçekli veri kümesi üzerinde eğitilmiş modellerin özellikleri daha küçük veri kümelerindeki benzer görevler için kullanılabilir (Yosinski ve diğ., 2014).

2.4.7 Aktif öğrenme (Active learning)

Aktif öğrenme, etiketli veri eksikliği veya etiketleme maliyetlerinin yüksek olduğu durumlarda kullanılan bir yöntemdir. Bu yaklaşımda, modelin performansını en çok artırabilecek örneklerin etiketlenmesi ve eğitime dahil edilmesi amaçlanır. Nefret söylemi tespitinde aktif öğrenme, etiketleme maliyetlerini düşürerek ve modelin doğruluğunu artırarak etkili sonuçlar elde etmeyi sağlamıştır (Alfina ve diğ., 2020).

2.4.8 Ensemble learning

Ensemble learning, birden fazla modelin veya algoritmanın bir araya getirilerek daha güçlü ve genelleştirilebilir bir model oluşturma yaklaşımıdır (Dietterich, 2000). Ensemble learning teknikleri, bagging, boosting ve stacking gibi yöntemlerle, modelin genel performansını artırarak ve aşırı uyuma karşı direnç sağlayarak daha iyi sonuçlar elde etmeyi hedefler.

3. LİTERATÜR

Sosyal medyadaki nefret söylemi içeren yayınların tespiti konusunda birçok akademik çalışma yapılmıştır ve yapay zeka temelli makine öğrenmesi algoritmalarının performansı değerlendirilmiştir.

Sosyal medyadaki nefret söylemini tespit etmek için yapay zeka temelli makine öğrenmesi algoritmalarının performans değerlendirmesi üzerine yapılan akademik çalışmalar, nefret söyleminin önlenmesi ve kontrolüne katkıda bulunmak amacıyla bu alanda yapılan çalışmalar üzerine araştırmalar yapılmıştır. Bu tezde, sosyal medya platformlarında nefret söylemi içeren yayınların tespitinde kullanılan yapay zeka algoritmalarının performans değerlendirmesine odaklanılmış ve daha ziyade bu yöndeki çalışmalar incelenmiştir.

Waseem ve Hovy (2016) tarafından yapılan bir çalışmada, Twitter üzerindeki nefret söylemi ve ırkçılığı tespit etmek için kullanılan doğal dil işleme (NLP) ve makine öğrenmesi (ML) yöntemleri ele alınmıştır. Bu çalışma, destek vektör makineleri (SVM) ve lojistik regresyon gibi algoritmaların performansını değerlendirerek, doğruluk, kesinlik, duyarlılık ve F1 puanı gibi metriklerle sonuçları analiz etmiştir.

Davidson ve diğ. (2017), Twitter üzerindeki nefret söylemini tespit etmek için derin öğrenme yöntemleri kullanmıştır. Bu çalışma, CNN (Convolutional Neural Networks) ve LSTM (Long Short-Term Memory) gibi sinir ağlarına dayalı yöntemlerin performansını değerlendirmiştir. Çalışma, bu yöntemlerin kullanımının, geleneksel ML yöntemlerinden daha iyi sonuçlar elde ettiğini göstermiştir.

Founta - (2018), sosyal medya platformlarında nefret söylemi ve saldırgan dilin yayılmasını önlemek için bir dizi makine öğrenmesi algoritması kullanarak çalışmalar yapmıştır. Bu çalışma, Rastgele Orman (RF), Naive Bayes, ve Gradyan Artırma (Gradient Boosting) gibi algoritmaların performansını değerlendirmiş ve bu yöntemlerin nefret söylemini tespit etmede başarılı olduğunu göstermiştir.

Singh ve diğ. (2020), Facebook ve Twitter gibi sosyal medya platformlarında nefret söylemini tespit etmek için BERT (Bidirectional Encoder Representations from Transformers) tabanlı derin öğrenme modellerini kullanmıştır. Bu çalışma, BERT tabanlı modellerin performansını değerlendirmiş ve bu yöntemlerin sosyal medyadaki nefret söylemini tespit etmede yüksek doğruluk elde ettiğini göstermiştir.

Sap ve diğ. (2019) çalışmalarında, nefret söylemi tespitinde ırksal önyargı riskini incelemiş ve bu tür önyargıları azaltmak için farklı ML yöntemlerini kullanmıştır.

Park ve Fung (2017) ise, Twitter'da kötü niyetli dil ve nefret söylemini tespit etmek için bir ve iki adımlı sınıflandırma yöntemlerini kullanarak bu yöntemlerin performansını değerlendirmiştir.

Fortuna ve Nunes (2018) tarafından yapılan kapsamlı bir araştırmada, nefret söyleminin otomatik tespitine ilişkin çeşitli algoritmalar ve yaklaşımlar incelenmiştir. Bu çalışma, mevcut yöntemlerin avantajlarını ve sınırlamalarını değerlendirerek, nefret söyleminin otomatik tespiti için gelecekteki araştırmaların yönünü belirlemeye çalışmıştır.

Alatawi ve Lee (2020) ise, Twitter üzerinde nefret söylemini tespit etmek için derin öğrenme yöntemlerini kullanan bir dizi algoritmayı karşılaştırmış ve değerlendirmiştir. Bu çalışma, algoritmaların performansını doğruluk, hassasiyet, hatırlama ve F1 skoru gibi metriklerle ölçmüştür ve sonuçlar, derin öğrenme tabanlı modellerin sosyal medyada nefret söylemini tespit etmede etkili olduğunu göstermiştir.

Ribeiro ve diğ. (2018), Twitter'da nefret söylemi yayarak topluluğa zarar veren kullanıcıları karakterize etmek ve tespit etmek için farklı makine öğrenmesi yöntemlerini kullanmıştır. Çalışma, kullanıcıların özelliklerini ve nefret söylemi yayma eğilimlerini analiz etmiş ve bu bilgilerin nefret söylemini tespit etmek için kullanılabileceğini göstermiştir.

Gambäck ve Sikdar (2017) tarafından yapılan bir çalışmada, nefret söylemi sınıflandırması için Evrişimli Sinir Ağları (Convolutional Neural Networks, CNN) kullanmışlardır. Bu çalışma, CNN tabanlı modellerin performansını değerlendirmiş ve geleneksel makine öğrenmesi yöntemlerine kıyasla daha iyi sonuçlar elde edildiğini göstermiştir.

ElSherief ve diğ. (2018), sosyal medyada nefret söylemi içeren dilin hedef temelli bir analizini gerçekleştirmişlerdir. Bu çalışma, nefret söylemi ve kötü niyetli dilin yayılmasını anlamaya yönelik dilbilimsel analizlerle makine öğrenmesi yöntemlerini birleştirmiştir ve bu yaklaşımın, nefret söylemi tespitinde daha kapsamlı sonuçlar elde etmeye yardımcı olduğunu göstermiştir.

Vidgen ve Derczynski (2020), kötü amaçlı dil eğitim veri setlerinin kalitesinin önemini vurgulayan bir çalışma yapmışlardır. Bu çalışma, düşük kaliteli veri setlerinin algoritmaların performansını ve güvenilirliğini olumsuz etkileyebileceğini belirtmekte ve nefret söylemi tespitinde daha iyi sonuçlar elde etmek için veri setlerinin kalitesinin önemini vurgulamaktadır.

Badjatiya ve diğ. (2017) tarafından yapılan bir çalışmada, tweetlerdeki nefret söylemi tespiti için derin öğrenme yöntemlerini kullanarak, farklı derin öğrenme modellerinin performansını karşılaştırmışlardır. Bu çalışma, derin öğrenme tabanlı modellerin, geleneksel makine öğrenmesi yöntemlerine göre daha yüksek doğruluk oranları elde ettiğini göstermiştir.

Chen ve diğ. (2018), sosyal medyada gençlerin çevrimiçi güvenliğini sağlamak amacıyla, sosyal medyada saldırgan dilin tespitine yönelik çalışmalar yapmışlardır. Bu çalışma, makine öğrenmesi algoritmalarının, gençlerin çevrimiçi güvenliğini sağlamaya yönelik olarak saldırgan ve kötü niyetli içeriği tespit etmekte etkili olduğunu göstermiştir.

Burnap ve Williams (2015), Twitter'da siber nefret söylemi konusunda politika ve karar verme süreçlerine katkı sağlamak amacıyla makine sınıflandırma ve istatistiksel modelleme yöntemlerini kullanarak çalışmalar yapmışlardır. Bu çalışma, nefret söylemi tespitinde makine öğrenmesi ve istatistiksel modelleme yöntemlerinin uygulanmasının, politika ve karar verme süreçlerine önemli katkılar sağlayabileceğini göstermiştir.

Makine öğrenmesi ve derin öğrenme yöntemlerinin nefret söylemi tespitindeki etkinliği, bu alandaki çalışmaların artmasına ve algoritmaların sürekli geliştirilmesine katkıda bulunmaktadır. Bununla birlikte, algoritmaların önyargıları ve yanlış pozitif/negatif sonuçları azaltmaya yönelik çalışmalar, nefret söylemi tespitinde daha güvenilir ve adil sonuçlar elde etmeye yöneliktir. Gelecekteki araştırmaların, algoritmaların performansını daha da geliştirmeye ve sosyal medyada

nefret söylemini önlemeye yönelik stratejiler üretmeye odaklanması beklenmekte ve bu alandaki gelişmelerin hızla devam etmesi öngörülmektedir.

Bu çalışmalar gibi birçok akademik araştırma, yapay zeka temelli algoritmaların sosyal medyadaki nefret söylemi içeren yayınların tespiti konusunda başarılı sonuçlar verdiğini göstermektedir. Ancak, bu algoritmaların doğruluğu ve tarafsızlığı konusunda bazı endişeler de bulunmaktadır ve bu nedenle daha fazla çalışmaya ihtiyaç duyulmaktadır.

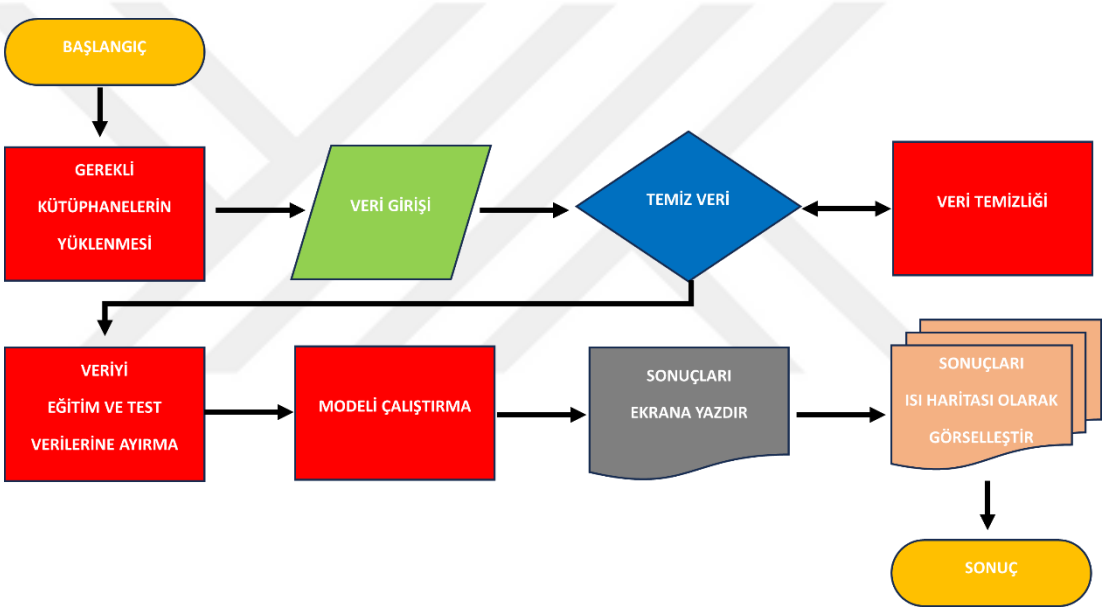
Bu çalışmaların çoğu, doğal dil işleme (NLP) tekniklerini kullanarak, sosyal medyadaki nefret söylemi içeren yayınların tespit edilmesi üzerine odaklanmıştır. Genellikle, algoritmalar birçok özellik kullanarak tweet'leri veya diğer sosyal medya yayınlarını analiz etmektedir. Özellikler arasında kelime dağarcığı, kelime sıklığı, kelime kombinasyonları, n-gramlar, emojiler, kullanıcı hesap bilgileri vb. yer alabilir.

Bununla birlikte, makine öğrenmesi algoritmalarının doğruluğu ve performansı, kullanılan veri kümesinin boyutu ve kalitesi gibi birçok faktöre bağlıdır. Bu nedenle, daha büyük ve çeşitli veri kümeleri kullanmak ve algoritmaların sürekli olarak güncellenmesi gerekebilir.

Sonuç olarak, sosyal medyada nefret söylemi içeren yayınların tespiti için yapay zeka temelli makine öğrenmesi algoritmalarının performans değerlendirmelerine odaklanan akademik çalışmalar, bu alanda önemli gelişmeler kaydetmektedir. Bu çalışmaların sonuçları, nefret söyleminin yayılmasını önlemek ve kontrol etmek için daha etkili ve güvenilir algoritmalar geliştirmeye yönelik çalışmaların devam etmesi gerektiğini göstermektedir. Gelecekte, daha kapsamlı ve güçlü algoritmaların geliştirilmesi ve daha fazla dil ve sosyal medya platformlarında uygulanması, nefret söyleminin tespitinde ve önlenmesinde daha iyi sonuçlar elde etmeye yardımcı olacaktır. Ayrıca, yanlış pozitif/negatif sonuçları ve algoritmaların önyargılarını azaltmaya yönelik çalışmalar, nefret söylemi tespitinde daha adil ve güvenilir sonuçlar sağlamayı amaçlamaktadır.

4. MATERYAL VE METOT

Sosyal medyadaki nefret söylemini içeren yayınların tespitinde yapay zeka temelli makine öğrenmesi algoritmalarının performans değerlendirmesi üzerine bu tez çalışması, nefret söylemini otomatik olarak tespit etmek için çeşitli makine öğrenmesi ve derin öğrenme yöntemlerini kullanmayı amaçlamaktadır. Kullanılan yöntemin akış diyagramı Şekil 4.1 de gösterilmiştir. Bu bölümde, kullanılan yöntem ve tekniklerin detaylı bir açıklamasını sunulmaktadır.



Şekil 4.1: Akış Diyagramı

4.1 Yöntemler

Bu bölümde çalışmada kullanılan Veri Toplama ve Ön İşleme, Özellik Çıkarımı, Sınıflandırma Modelleri, Model Eğitimi ve Performans Değerlendirmesi, Model Karşılaştırması ve Analiz, Sonuçlar ve Gelecekteki Çalışmalar, Uygulama ve Entegrasyon, Etik ve Yasal Hususlar, Özelleştirme ve Transfer Öğrenme, Çalışmanın Sınırlılıkları ve Potansiyel İyileştirmeler başlıklarından ve bu başlıklara ait süreçlerden bahsedilmektedir.

4.1.1 Veri toplama ve ön işleme

Bu çalışmanın temelini oluşturan veri kümesi, nefret söylemi içeren ve içermeyen sosyal medya yayınlarından oluşmaktadır. Veri kümesi, kaggle.com platformundan elde edilen Twitter hate speech isimli verisetidir. Veri kümesi belirlenirken, özellikle nefret söylemiyle ilgili konulara odaklanan hesaplar ve sayfalar incelenmiştir. Ön işleme adımında, gürültüyü azaltmak için metinler temizlenmiş ve özellik çıkarımı için hazırlanmıştır.

4.1.2 Özellik çıkarımı

Makine öğrenmesi ve derin öğrenme yöntemlerini kullanarak nefret söylemini tespit etmek için, metinlerden özellikler çıkarılması gerekmektedir. Çalışmada, hem geleneksel metin madenciliği yöntemleri (ör. TF-IDF, n-gram) (Manning ve diğ., 2008) hem de derin öğrenme tabanlı yöntemler (ör. word2vec, GloVe, ELMo, BERT) (Mikolov ve diğ., 2013; Pennington ve diğ., 2014; Peters ve diğ., 2018; Devlin ve diğ., 2019) kullanılmıştır.

4.1.3 Sınıflandırma modelleri

Nefret söylemi içeren ve içermeyen yayınları sınıflandırmak için çeşitli makine öğrenmesi ve derin öğrenme algoritmaları kullanılmıştır. Bu algoritmalar Karar Ağaçları (DT), Gradyan Artırma (Gradient Boosting), K-En Yakın Komşu (KNN), Lojistik Regresyon (LR), Çok Katmanlı Algılayıcılar (MLP), Çok Terimli Naif Bayes (MNB), Rastgele Orman (RF), Destek Vektör Makineleri (SVM) modelleridir.

4.1.4 Model eğitimi ve performans değerlendirmesi

Veri kümesi, model eğitimi ve performans değerlendirmesi için eğitim, doğrulama ve test alt kümelerine bölünmüştür (Kohavi, 1995). Sınıflandırma modelleri, eğitim veri kümesi üzerinde eğitilmiş ve doğrulama veri kümesi üzerinde hiperparametre ayarlaması yapılmıştır (Bergstra ve Bengio, 2012). Model performansını değerlendirmek için, test veri kümesi üzerinde çeşitli metrikler kullanılmıştır. Bu metrikler arasında doğruluk (accuracy), kesinlik (precision), duyarlılık (recall), F1 skoru (F1-score) ve alan eğrisi altındaki değer AUC-ROC Skoru (AUC-ROC Score) bulunmaktadır (Sokolova ve Lapalme, 2009).

4.1.5 Model karşılaştırması ve analiz

Tüm modellerin performanslarının karşılaştırılması ve analizi, en iyi modelin seçilmesine ve önerilmesine olanak tanımıştır. Ayrıca, hatalı sınıflandırma örneklerinin incelenmesi, modellerin zayıf yönlerinin ve potansiyel iyileştirmelerin belirlenmesine yardımcı olmuştur (Albright ve diğ., 2019).

4.1.6 Sonuçlar ve gelecekteki çalışmalar

Bu tez çalışması, sosyal medyadaki nefret söylemi içeren yayınların tespitinde makine öğrenmesi ve derin öğrenme algoritmalarının etkinliğini değerlendirmiştir. Sonuçlar, bazı yöntemlerin nefret söylemi tespiti konusunda daha başarılı olduğunu göstermiştir. Bununla birlikte, daha fazla araştırma ve geliştirme ile modellerin performanslarının daha da artırılması mümkündür.

4.1.8 Uygulama ve entegrasyon

Bu tez çalışmasında geliştirilen nefret söylemi tespit modelleri, sosyal medya platformlarına entegre edilebilir ve gerçek zamanlı içerik moderasyonu için kullanılabilir. Bu entegrasyon, platformların toksik ve zararlı içerikleri daha hızlı ve etkin bir şekilde belirlemelerine ve önlemelerine olanak tanır (Fortuna ve Nunes, 2018).

4.1.8 Etik ve yasal hususlar

Nefret söylemi tespiti için makine öğrenimi ve derin öğrenme algoritmalarının kullanılması, etik ve yasal düşünceleri de beraberinde getirir. Özellikle, kullanıcıların mahremiyetini koruma ve yanlış pozitif/negatif sınıflandırmaların etkilerini en aza indirgeme gibi konular önemlidir (Schmidt ve Wiegand, 2017).

4.1.9 Özelleştirme ve transfer öğrenme

Geliştirilen nefret söylemi tespit modelleri, farklı diller ve kültürel bağlamlar için özelleştirilebilir ve transfer öğrenme yöntemleri kullanılarak daha geniş kapsamlı ve etkili hale getirilebilir (Ruder, 2019).

4.1.10 Çalışmanın sınırlılıkları ve potansiyel iyileştirmeler

Bu tez çalışması, sosyal medyadaki nefret söylemi içeren yayınları tespit etmek için makine öğrenmesi ve derin öğrenme algoritmalarını kullanırken, bazı sınırlılıklar ve potansiyel iyileştirmeler vardır. Bunlar arasında, veri kümesi dengesizliği, etiketleme tutarlılığı ve algoritmaların önyargıları ve sınırlılıkları gibi konular bulunmaktadır. Gelecekteki çalışmalar, bu sınırlılıkların üstesinden gelmeye ve algoritmaların performansını daha da artırmaya odaklanabilir (Dixon ve diğ., 2018).

4.2 Algoritmalar

Bu bölümde çalışmada kullanılan Karar Ağaçları (DT), Gradyan Artırma (Gradient Boosting), K-En Yakın Komşu (KNN), Lojistik Regresyon (LR), Çok Katmanlı Algılayıcılar (MLP), Çok Terimli Naif Bayes (MNB), Rastgele Orman (RF) ve Destek Vektör Makineleri (SVM) algoritmaları kısaca tanımlanmaktadır.

4.2.1 Karar ağaçları (DT)

Karar Ağaçları (DT), sınıflandırma ve regresyon görevleri için kullanılan popüler bir makine öğrenimi algoritmasıdır. Bu algoritma, bir dizi karar kurallarını kullanarak giriş özelliklerine dayanarak bir çıktı tahmini üretir (Quinlan, 1986). Karar ağaçlarının temel avantajlarından biri, modelin sonuçlarına görsel bir biçimde erişim sağlaması ve bu nedenle yorumlanabilir olmasıdır (Breiman ve diğ., 1986). Ancak, ağaçların derinliği kontrol edilmezse, modelin aşırı uyuma eğilimi vardır (Mitchell, 1997). Aşırı uyum, modelin eğitim verisine mükemmel bir şekilde uymasını, ancak yeni verilere iyi genelleme yapamamasını ifade eder.

4.2.2 Gradyan artırma (Gradient boosting)

Gradyan Artırma, makine öğrenimi algoritmalarından biridir ve özellikle sınıflandırma ve regresyon problemleri için güçlü bir yöntem olarak bilinir. Bu algoritma, modeli adım adım geliştirerek hataları azaltmayı amaçlar. Freund ve Schapire (1997) bu konseptin ilk örneklerinden biri olan Adaboost'u tanıttı. Gradyan Artırma'nın temel mantığı, önceki adımda yapılan hataları düzeltmek için yeni bir tahminci eklemektir. Bu süreç, belirlenen bir hata oranına ya da iterasyon sayısına

ulaşana kadar devam eder (Friedman, 2001). Bu yöntemin başarısı, birçok uygulamada yüksek performans göstermesiyle kanıtlanmıştır.

4.2.3 K-En yakın komşu (KNN)

K-En Yakın Komşu (KNN) sınıflandırma ve regresyon problemlerinin çözümünde kullanılan gözetimli öğrenme algoritmasıdır. KNN, bir örnek için sınıf etiketini tahmin ederken, veri kümesindeki diğer örnekler arasında en yakın k komşusunu bulur ve bu komşuların çoğunluk sınıfını veya ağırlıklı ortalamasını kullanarak tahminde bulunur (Cover ve Hart, 1967). Algoritmanın başarısı, doğru bir uzaklık ölçüsünün seçilmesine, uygun k değerinin belirlenmesine ve veri setinin özelliklerinin ölçeklendirilmesine bağlıdır (Hechenbichler ve Schliep, 2004). KNN algoritması, basitliği, anlaşılabilirliği ve çok çeşitli uygulama alanlarına uygunluğu ile bilinir. Bununla birlikte, büyük veri kümeleri için hesaplama maliyeti yüksek olabilir ve bu durumda verinin indirgenmesi veya özellik seçimi gibi tekniklerle optimize edilmesi gerekebilir.

4.2.4 Lojistik regresyon (LR)

Lojistik Regresyon (LR), ikili ya da çoklu sınıflandırma problemleri için sıkça kullanılan istatistiksel bir analiz yöntemidir. Bu yöntem, bağımlı değişkenin olasılık değerlerini tahmin ederken bağımsız değişkenlerin etkisini belirlemeye yardımcı olur (Hosmer Jr, Lemeshow, ve Sturdivant, 2013). Lojistik regresyon, doğrusal regresyona benzer şekilde, bağımsız değişkenlerin ağırlıklarını öğrenir, ancak sonuçlar logit dönüşümü ile sıkıştırılır, böylece elde edilen değerler 0 ile 1 arasında olur (Agresti, 2007). Bu nedenle, sonuçlar genellikle bir olayın gerçekleşme olasılığı olarak yorumlanır.

4.2.5 Çok katmanlı algılayıcılar (MLP)

Çok Katmanlı Algılayıcılar (MLP), yapay sinir ağlarının en yaygın şekillerinden biridir ve birden fazla katmandan oluşan tam bağlantılı bir sinir ağıdır. Bir girdi katmanı, bir veya daha fazla gizli katman ve bir çıktı katmanından oluşan MLP, türevlenebilir aktivasyon fonksiyonları kullanarak karmaşık özellikleri modelleyebilir (Hornik ve diğ., 1989). MLP, geri yayılım algoritmasıyla birlikte, eğitim sırasında ağırlıkları optimize etmek için yaygın olarak kullanılır (Rumelhart, Hinton, ve Williams, 1986). Bu optimizasyon süreci, gradyan inişi ve onun çeşitli

varyantları ile gerçekleştirilir (LeCun ve diğ., 1998). Son yıllarda, derin öğrenme alanında, derin MLP modelleri de başarılı bir şekilde uygulanmıştır ve bu modeller birçok uygulama alanında yüksek performans göstermektedir (Goodfellow ve diğ., 2016).

4.2.6 Çok terimli naif bayes (MNB)

Çok Terimli Naif Bayes sınıflandırıcısı, temelini Bayes teoreminin prensiplerine dayandıran basit ve etkili bir istatistiksel sınıflandırma yöntemidir. Bayes teoremi, yeni verilere dayanarak önceki bilgiyi güncelleme prensibini tanımlar. Naive Bayes, özellikler arasında bağımsızlık varsayımıyla bilinir ve bu nedenle "naive" (naif/saf) olarak adlandırılır. Bu yöntem, özellikle yüksek boyutlu veri kümelerinde etkili bir şekilde çalışabilir ve sıklıkla metin madenciliği, spam filtreleme gibi uygulamalarda kullanılır (McCallum ve Nigam, 1998). Ancak, gerçek dünyada özellikler arasındaki bağımsızlık varsayımı her zaman gerçekleşmez, bu yüzden uygulama sırasında dikkatli olunmalıdır (Zhang, 2004).

4.2.7 Rastgele orman (RF)

Rastgele Orman (RF), karar ağaçlarının bir araya gelerek daha stabil ve güçlü bir model oluşturduğu bir topluluk öğrenme yöntemidir. Breiman (2001) bu yöntemi tanıtarak, tek bir karar ağacının eğitim veri setine aşırı uyum sağlama eğilimini azaltmak için birden fazla ağacın eğitilmesi ve bu ağaçların oylamalarla veya ortalama alarak bir sonuca varmasını önermiştir. Rastgele Orman'ın avantajlarından biri, özellik seçimi yaparken rastgele özellik alt kümelerini kullanmasıdır, bu da modelin genelleştirme yeteneğini artırır (Liaw ve Wiener, 2002). Ayrıca, Rastgele Orman algoritması, özelliklerin önemini ölçebilme, eksik veriyle başa çıkabilme ve modelin yorumlanabilirliği gibi diğer avantajlara da sahiptir (Cutler ve diğ., 2007).

4.2.8 Destek vektör makineleri (SVM)

Destek Vektör Makineleri (SVM), sınıflandırma ve regresyon analizi için kullanılan popüler bir denetimli öğrenme modelidir. SVM, bir ayrım marjını maksimize ederek veri noktalarını sınıflandırma işlemini gerçekleştirir. Vapnik ve arkadaşları tarafından önerilen bu yöntem, veriyi iki sınıfa ayırmak için bir hiperdüzlem seçer (Vapnik, 1995). Özellikle yüksek boyutlu veri setlerinde etkili bir şekilde çalışabilme kapasitesine sahip olan SVM, çekirdek fonksiyonları sayesinde

doğrusal olmayan sınıflandırmalar da yapabilmektedir (Schölkopf ve Smola, 2002). Aynı zamanda, SVM, büyük veri setleri üzerinde eğitim süresi ve özellik seçiminin zorlukları gibi bazı zorluklara sahip olabilir (Boser, Guyon ve Vapnik, 1992).

5. BULGULAR VE TARTIŞMA

Bu bölümde öncelikle 8 farklı algoritma tek tek çalıştırılmış ardından algoritmaların tamamı aynı anda çalıştırılmak suretiyle ortak ve daha başarılı bir sonuca varılmak hedeflenmiştir.

Tüm algoritmalarda sırasıyla aşağıdaki işlemler uygulanmıştır.

- Veriyi yükleme ve ön işleme.
- Her model için veriyi hazırlama.
- Modeli tanımlama.
- Modeli eğitme.
- Değerlendirme metriklerini hesaplama.
- Sonuçları yazdırma ve görselleştirme.
- Sonuçları yorumlama.

Sonuçları yorumlamak için, ısı haritalarındaki metrik değerlerine odaklanılmalıdır. Yüksek bir Doğruluk (Accuracy), Kesinlik (Precision), Duyarlılık (Recall), F1 Skoru (F1 Score) ve AUC-ROC Skoru (AUC-ROC Score) puanı, modelin iyi performans gösterdiğini gösterir. Düşük Kesinlik (Precision), yanlış pozitif tahminlerin sayısının yüksek olduğunu gösterirken, düşük Duyarlılık (Recall), yanlış negatif tahminlerin sayısının yüksek olduğunu gösterir. İdeal olarak, yüksek Kesinlik (Precision) ve Duyarlılık (Recall) ile yüksek bir F1 Skoru (F1 Score) elde etmek hedeflenir.

Doğruluk (Accuracy): Tüm tahminlerin doğru oranını gösterir.

Kesinlik (Precision): Pozitif olarak tahmin edilen örneklerin gerçekten ne kadarının pozitif olduğunu gösterir.

Duyarlılık (Recall): Gerçek pozitif örneklerin ne kadarının pozitif olarak tahmin edildiğini gösterir.

F1 Skoru (F1 Score): Kesinlik (Precision) ve Duyarlılık (Recall) değerinin harmonik ortalamasını alır ve dengeli bir performans ölçüsü sağlar.

AUC-ROC Skoru (AUC-ROC Score): Sınıflandırıcının rastgele pozitif ve negatif bir örneği doğru bir şekilde sıralama olasılığını gösterir.

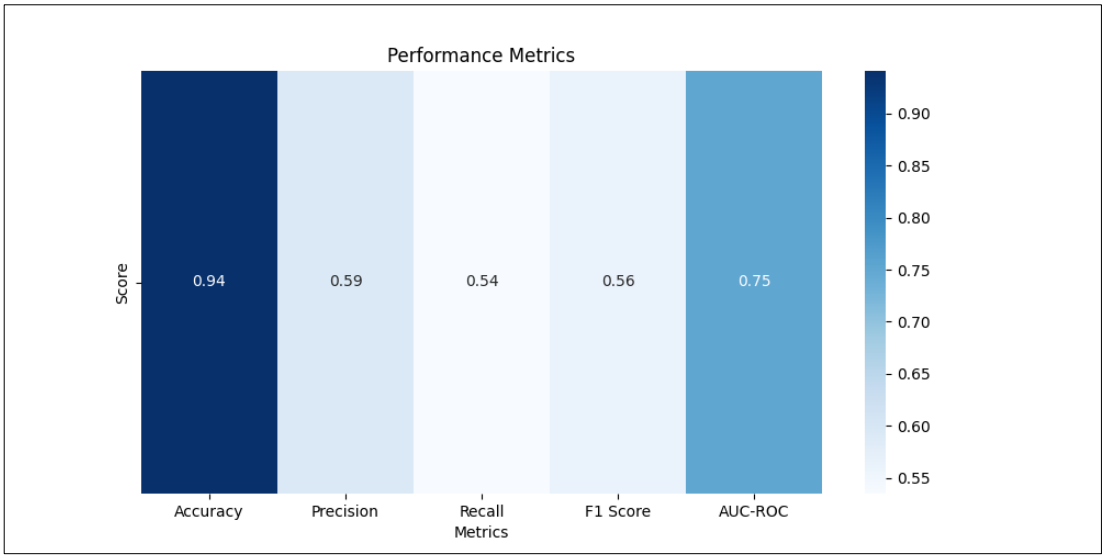
Tüm bu sonuçlara göre, modelin nefret söylemi içeren tweetleri tespit etmedeki performansı değerlendirilebilir. Yüksek doğruluk, doğru sınıflandırılan örneklerin oranını gösterir. Ancak, denge için Kesinlik (Precision), Duyarlılık (Recall) ve F1 Skoru (F1 Score) değerlerine de bakmalısınız. Özellikle, nefret söylemi tespitinde yanlış alarm oranını (yanlış pozitif) azaltmak için, Kesinlik (Precision) değerine dikkat edilmelidir. Duyarlılık (Recall), gerçek pozitiflerin ne kadarının doğru bir şekilde sınıflandırıldığını gösterirken, F1 Skoru (F1 Score) ise Kesinlik (Precision) ve Duyarlılık (Recall) arasında bir denge sağlar. AUC-ROC Skoru (AUC-ROC Score), modelin sınıfları ayırt etme yeteneğini gösterir; 1'e yakın bir değer, mükemmel bir ayırt ediciliğe işaret eder.

5.1 DT Performans Değerlendirme

Karar Ağaçları (DT) algoritması ile twitter-hate-speech veriseti üzerinde yapılan performans denemesi sonucunda ulaşılan metrik değerler Çizelge 5.1 de gösterilmektedir.

Çizelge 5.1: DT Metrik Değerler

Metrik	**Değer**
Doğruluk (Accuracy)	0.9407164085718754
Kesinlik (Precision)	0.5936739659367397
Duyarlılık (Recall)	0.5350877192982456
F1 Skoru (F1 Score)	0.5628604382929643
AUC-ROC Skoru (AUC-ROC Score)	0.753479517388722



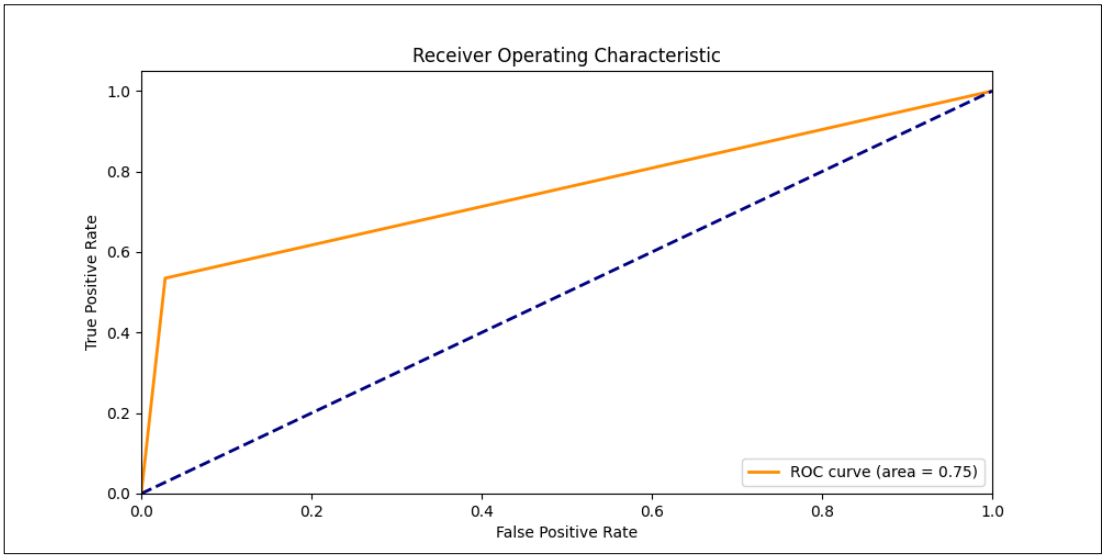
Şekil 5.1: DT Metrik Değerler Isı Haritası

Modelin genel doğruluğunu gösteren Doğruluk (Accuracy) değeri 0.9407 olarak oldukça yüksek bir değere sahiptir. Bu, modelin tahminlerinin çoğunun gerçek değerlerle uyumlu olduğunu göstermektedir.

Modelin pozitif olarak sınıflandırdığı örneklerin ne kadarının gerçekten pozitif olduğunu gösteren Kesinlik (Precision) değeri 0.5937 olarak oldukça düşük bir değere sahiptir. Bu da modelin pozitif tahminlerinin önemli bir kısmının yanlış pozitif (hatalı alarm) olduğunu gösterir.

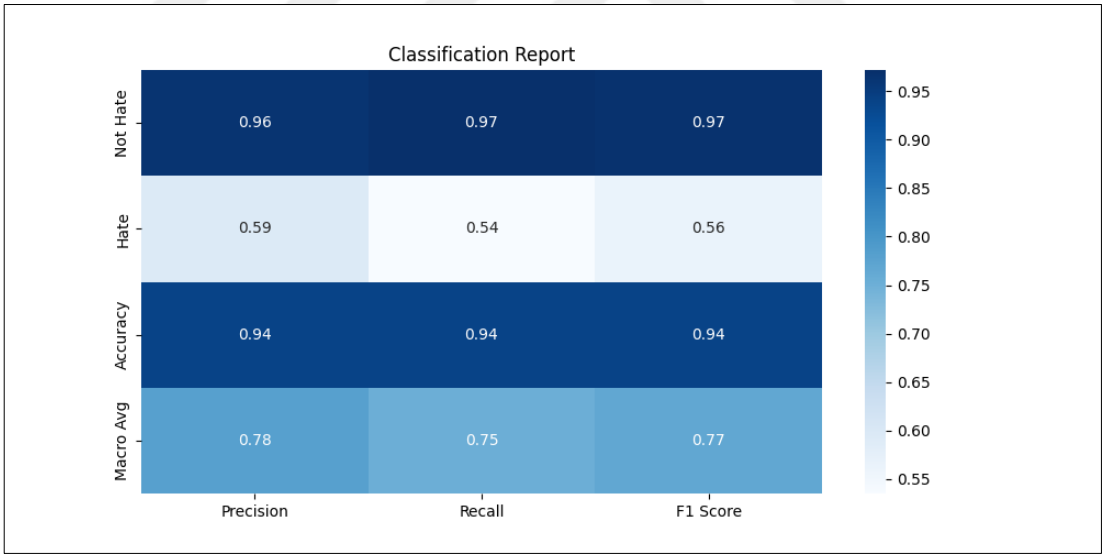
Gerçekte pozitif olan örneklerin ne kadarının model tarafından doğru olarak pozitif olarak sınıflandırıldığını gösteren Duyarlılık (Recall) değeri 0.5351 olarak orta düzeyde olduğu görülmektedir. Yani bazı gerçek pozitiflerin kaçırılmış olduğu söylenebilir.

Kesinlik ve duyarlılığın harmonik ortalaması olan ve dengeli bir metrik olarak kabul edilen F1 Skoru (F1 Score) değeri 0.5629 olarak ne kesinlik ne de duyarlılık açısından mükemmel olmayan ancak kabul edilebilir bir denge sağladığını göstermektedir. Modelin metrik değerlerini içeren ısı haritası Şekil 5.1 de gösterilmiştir.



Şekil 5.2: DT AUC-ROC Skoru (AUC-ROC Score) Eğrisi

Modelin sınıflandırma performansını genel olarak değerlendiren AUC-ROC Skoru (AUC-ROC Score) değeri 0.7535 olarak oldukça iyi olduğu söylenebilir. Ancak mükemmel değil. AUC-ROC Skoru (AUC-ROC Score) eğrisi Şekil 5.2 de gösterilmiştir.



Şekil 5.3: DT Sınıflandırma Raporu (Classification Report) Isı Haritası

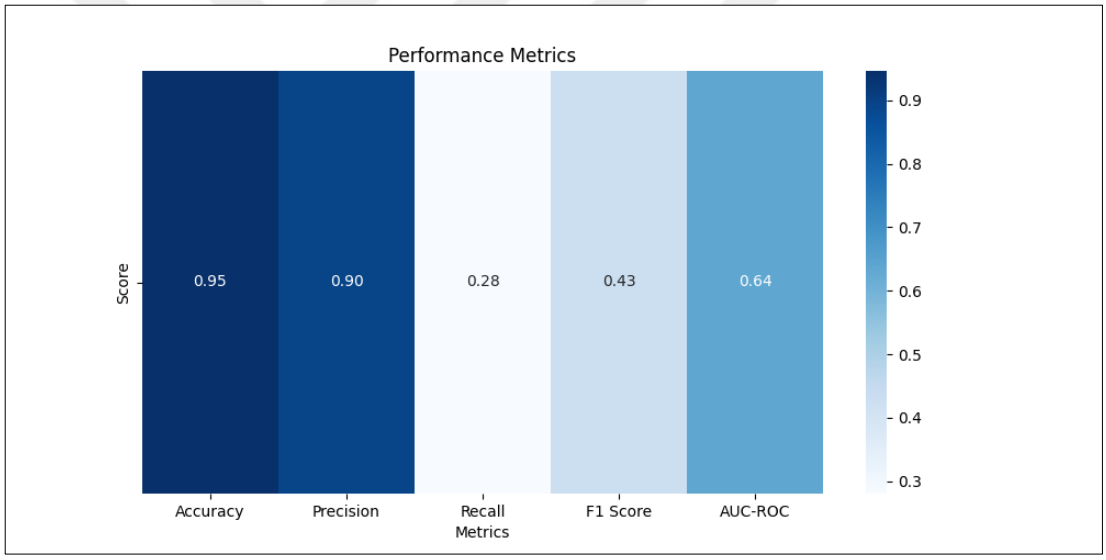
Genel olarak, bu modeli oldukça yüksek bir doğruluk oranına sahip olduğu ancak kesinlik, duyarlılık ve F1 skoru olarak düşük olduğu söylenebilir. Modelin Sınıflandırma Raporu (Classification Report) değerlerini içeren ısı haritası Şekil 5.3 de gösterilmiştir.

5.2 Gradyan Artırma Performans Değerlendirme

Gradyan artırma algoritması ile twitter-hate-speech veriseti üzerinde yapılan performans denemesi sonucunda ulaşılan metrik değerler Çizelge 5.2 de gösterilmektedir.

Çizelge 5.2: Gradyan Artırma Metrik Değerler

Metrik	**Değer**
Doğruluk (Accuracy)	0.9463475676521195
Kesinlik (Precision)	0.8951048951048951
Duyarlılık (Recall)	0.2807017543859649
F1 Skoru (F1 Score)	0.42737896494156924
AUC-ROC Skoru (AUC-ROC Score)	0.6390876129180961



Şekil 5.4: Gradyan Artırma Metrik Değerler Isı Haritası

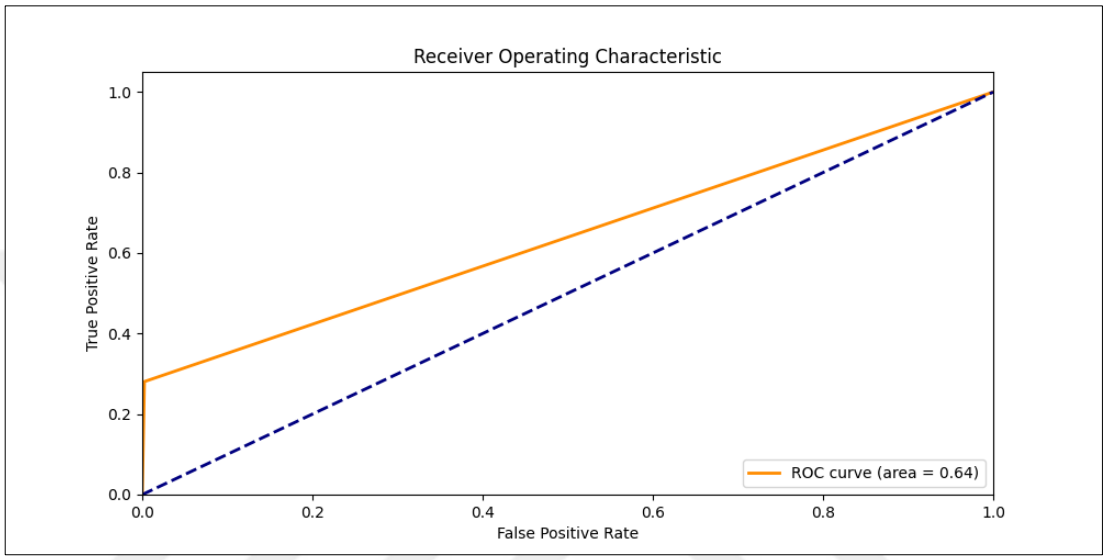
Modelin genel doğruluğunu gösteren Doğruluk (Accuracy) değeri 0.9463 olarak oldukça yüksek bir değere sahiptir. Bu, modelin tahminlerinin çoğunun gerçek değerlerle uyumlu olduğunu göstermektedir.

Modelin pozitif olarak sınıflandırdığı örneklerin ne kadarının gerçekten pozitif olduğunu gösteren Kesinlik (Precision) değeri 0.8951 olarak iyi bir değere sahiptir. Bu da modelin pozitif tahminlerinin %89.51'inin gerçekten pozitif olduğunu gösterir.

Gerçekte pozitif olan örneklerin ne kadarının model tarafından doğru olarak pozitif olarak sınıflandırıldığını gösteren Duyarlılık (Recall) değeri 0.2807 olarak

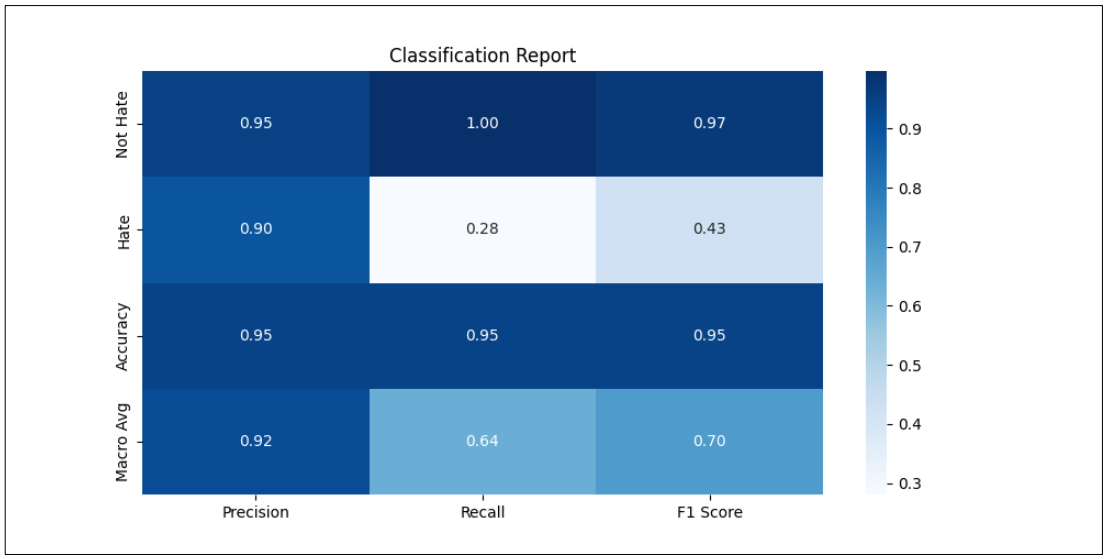
oldukça düşük düzeyde olduğu görülmektedir. Bu da modelin gerçek pozitif vakaları tespit etmede zayıf olduğunu gösterir.

Kesinlik ve duyarlılığın harmonik ortalaması olan ve dengeli bir metrik olarak kabul edilen F1 Skoru (F1 Score) değeri 0.4274 olarak modelin kesinlikle iyi performans gösterse de, duyarlılıkta zayıf olduğunu gösterir. Modelin metrik değerlerini içeren ısı haritası Şekil 5.4 de gösterilmiştir.



Şekil 5.5: Gradyan Artırma AUC-ROC Skoru (AUC-ROC Score) Eğrisi

Modelin sınıflandırma performansını genel olarak değerlendiren AUC-ROC Skoru (AUC-ROC Score) değeri 0.6391 olarak rastgele tahminden daha iyi olduğunu, ancak mükemmel olmadığını gösterir. AUC-ROC Skoru (AUC-ROC Score) eğrisi Şekil 5.5 de gösterilmiştir.



Şekil 5.6: Gradyan Artırma Sınıflandırma Raporu Isı Haritası

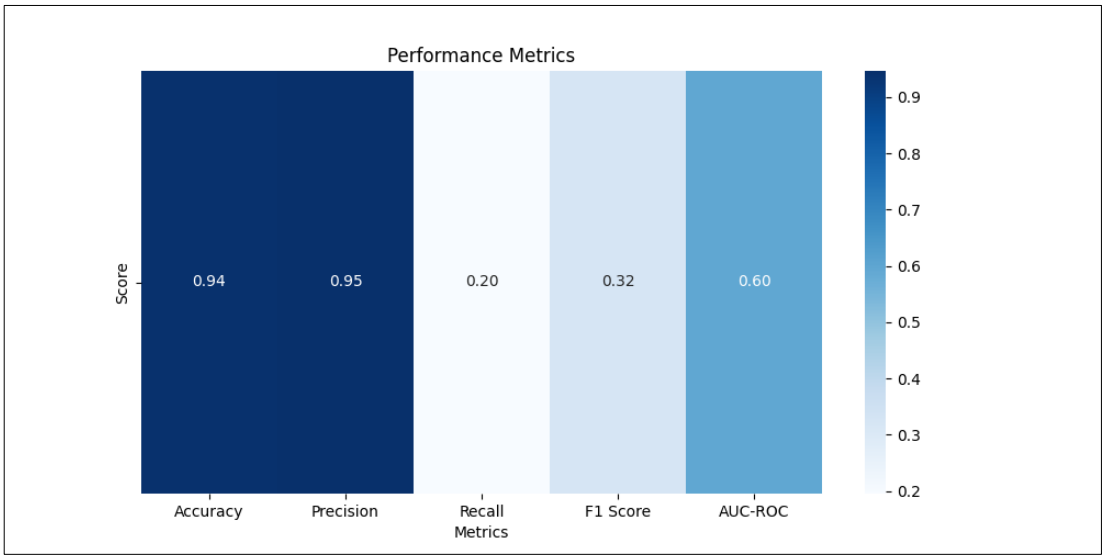
Genel olarak, bu model yüksek doğruluk ve kesinliğe sahipken, duyarlılık (recall) ve F1 skorunda zayıf performans göstermektedir. Bu, modelin gerçek pozitif vakaları tespit etmede zorlanabileceği anlamına gelir. Bu tür durumlar, modelin sadece çoğunluk sınıfına iyi performans gösterdiği, ancak azınlık sınıfını (bu durumda pozitif vakaları) göz ardı ettiği anlamına gelebilir. Modelin Sınıflandırma Raporu (Classification Report) değerlerini içeren ısı haritası Şekil 5.6 da gösterilmiştir.

5.3 KNN Performans Değerlendirme

K-En Yakın Komşu (KNN) algoritması ile twitter-hate-speech veriseti üzerinde yapılan performans denemesi sonucunda ulaşılan metrik değerler Çizelge 5.3 de gösterilmektedir.

Çizelge 5.3: KNN Metrik Değerler

Metrik	**Değer**
Doğruluk (Accuracy)	0.9418113561708118
Kesinlik (Precision)	0.9468085106382979
Duyarlılık (Recall)	0.19517543859649122
F1 Skoru (F1 Score)	0.3236363636363636
AUC-ROC Skoru (AUC-ROC Score)	0.5971666312066168



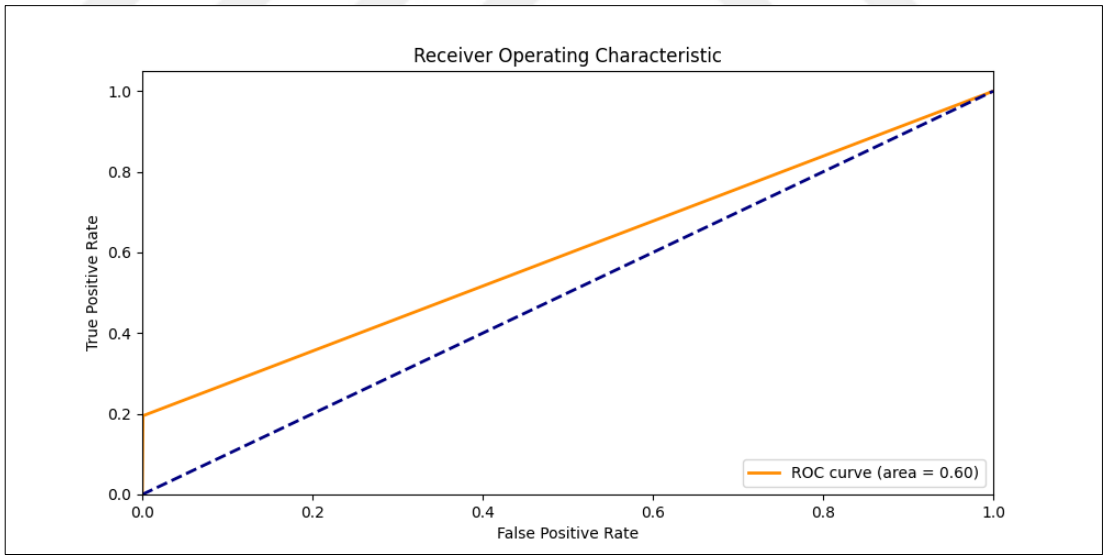
Şekil 5.7: KNN Metrik Değerler Isı Haritası

Modelin genel doğruluğunu gösteren Doğruluk (Accuracy) değeri 0.9418 olarak oldukça yüksek bir değere sahiptir. Bu, modelin tahminlerinin çoğunun gerçek değerlerle uyumlu olduğunu göstermektedir.

Modelin pozitif olarak sınıflandırdığı örneklerin ne kadarının gerçekten pozitif olduğunu gösteren Kesinlik (Precision) değeri 0.9468 olarak yüksek bir değere sahiptir. Bu da modelin yanlış pozitifleri (yanlış alarm) az sayıda tutma yeteneğini ifade eder.

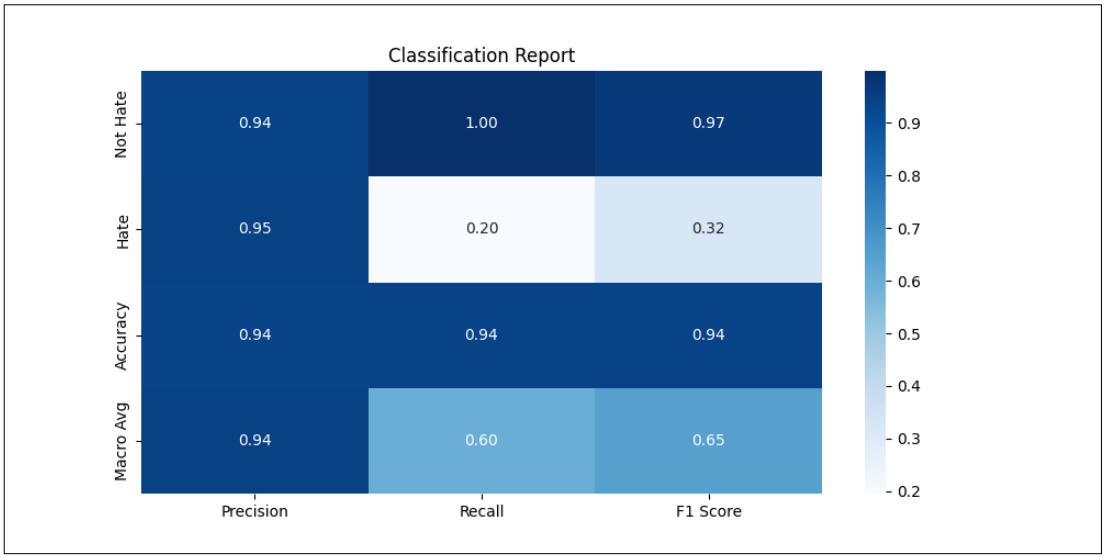
Gerçekte pozitif olan örneklerin ne kadarının model tarafından doğru olarak pozitif olarak sınıflandırıldığını gösteren Duyarlılık (Recall) değeri 0.1952 olarak çok düşük düzeyde olduğu görülmektedir. Bu da modelin gerçek pozitif vakaları tespit etmede çok zayıf olduğunu gösterir.

Kesinlik ve duyarlılığın harmonik ortalaması olan ve dengeli bir metrik olarak kabul edilen F1 Skoru (F1 Score) değeri 0.3236 olarak kesinlik ve duyarlılık arasında önemli bir dengesizliği göstermektedir. Bu durumda, yüksek kesinlik ve düşük duyarlılığın birleşimi nedeniyle düşük bir F1 skoru elde edilmiştir. Modelin metrik değerlerini içeren ısı haritası Şekil 5.7 de gösterilmiştir.



Şekil 5.8: KNN AUC-ROC Skoru (AUC-ROC Score) Eğrisi

Modelin sınıflandırma performansını genel olarak değerlendiren AUC-ROC Skoru (AUC-ROC Score) değeri 0.5972 olarak rastgele tahmin yapmaktan biraz daha iyi olduğunu gösterir. AUC-ROC Skoru (AUC-ROC Score) eğrisi Şekil 5.8 de gösterilmiştir.



Şekil 5.9: KNN Sınıflandırma Raporu (Classification Report) Isı Haritası

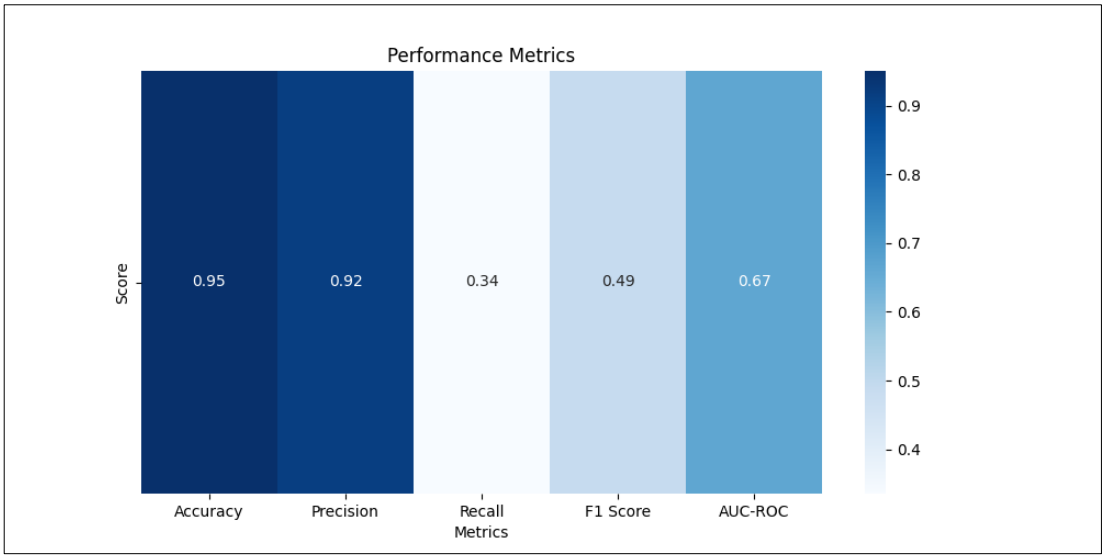
Genel olarak, model yüksek doğruluk ve kesinlik sunarken, duyarlılık ve F1 skorları oldukça düşüktür. Bu, modelin pozitif sınıfı doğru bir şekilde tespit etmede zorluk çektiğini gösterir. Ayrıca, AUC-ROC Skoru (AUC-ROC Score) değeri, modelin toplam performansının orta seviyede olduğuna işaret eder. Modelin Sınıflandırma Raporu (Classification Report) değerlerini içeren ısı haritası Şekil 5.9 da gösterilmiştir.

5.4 LR Performans Değerlendirme

Lojistik Regresyon (LR) algoritması ile twitter-hate-speech veriseti üzerinde yapılan performans denemesi sonucunda ulaşılan metrik değerler Çizelge 5.4 de gösterilmektedir.

Çizelge 5.4: LR Metrik Değerler

Metrik	**Değer**
Doğruluk (Accuracy)	0.9504145158767402
Kesinlik (Precision)	0.9161676646706587
Duyarlılık (Recall)	0.3355263157894737
F1 Skoru (F1 Score)	0.4911717495987159
AUC-ROC Skoru (AUC-ROC Score)	0.6665841112381763



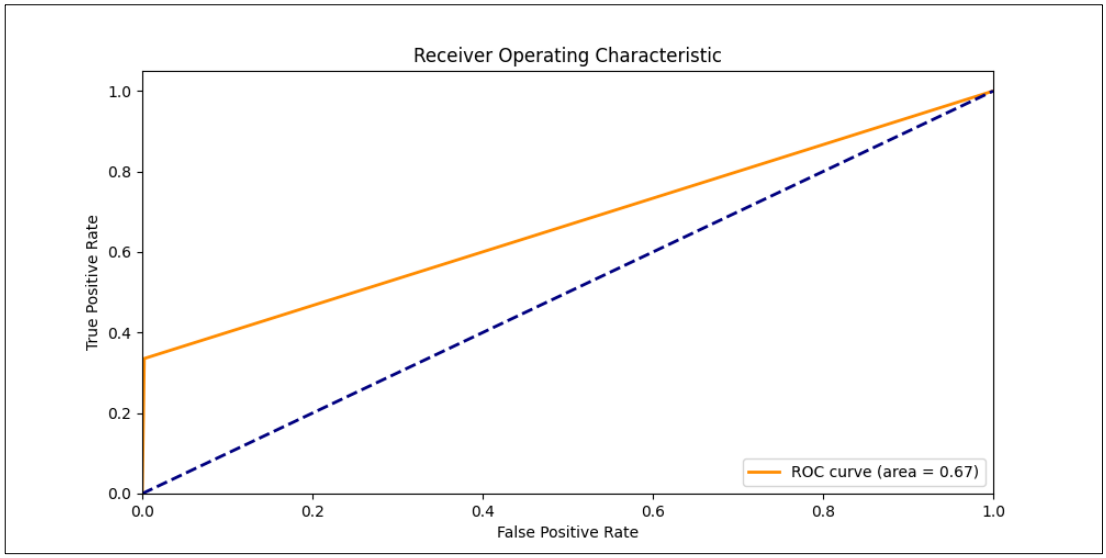
Şekil 5.10: LR Metrik Değerler Isı Haritası

Modelin genel doğruluğunu gösteren Doğruluk (Accuracy) değeri 0.9504 olarak oldukça yüksek bir değere sahiptir. Bu, modelin tahminlerinin çoğunun gerçek değerlerle uyumlu olduğunu göstermektedir.

Modelin pozitif olarak sınıflandırdığı örneklerin ne kadarının gerçekten pozitif olduğunu gösteren Kesinlik (Precision) değeri 0.9162 olarak oldukça iyi bir değere sahiptir. Bu da modelin bir şeyi pozitif olarak işaretlediğinde, bu durumun doğru olma olasılığının %91.62 olduğunu gösterir.

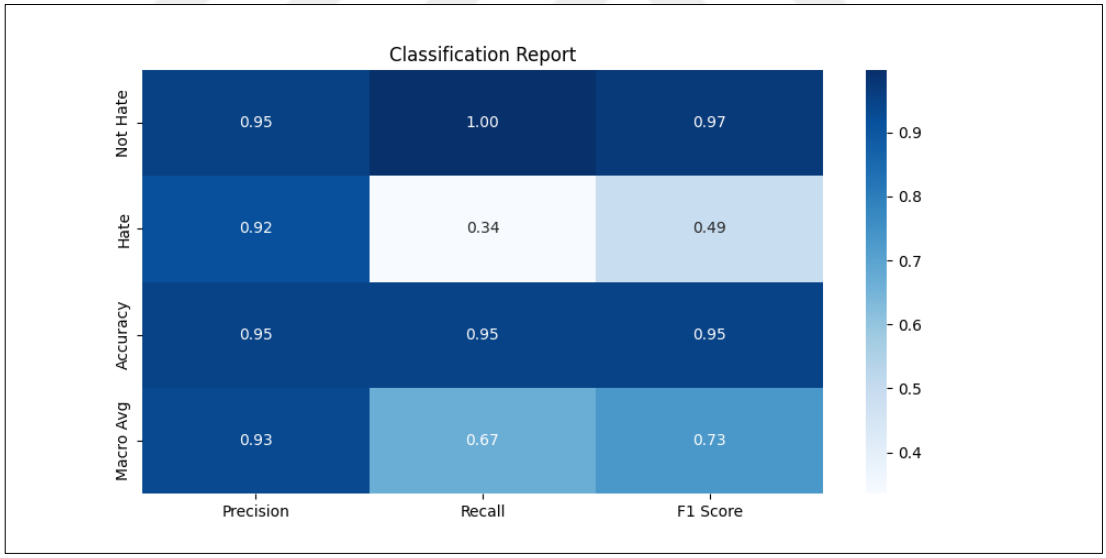
Gerçekte pozitif olan örneklerin ne kadarının model tarafından doğru olarak pozitif olarak sınıflandırıldığını gösteren Duyarlılık (Recall) değeri 0.3355 olarak oldukça düşük düzeyde olduğu görülmektedir. Yani model gerçek pozitif vakaların sadece üçte birini doğru olarak tespit edebiliyor. Bu, modelin pozitif vakaları kaçırdığı anlamına gelir.

Kesinlik ve duyarlılığın harmonik ortalaması olan ve dengeli bir metrik olarak kabul edilen F1 Skoru (F1 Score) değeri 0.4912 olarak orta seviyededir. Bu, modelin kesinliği yüksek olsa da, tüm pozitif vakaları tespit etme konusunda iyi olmadığını gösterir. Modelin metrik değerlerini içeren ısı haritası Şekil 5.10 da gösterilmiştir.



Şekil 5.11: LR AUC-ROC Skoru (AUC-ROC Score) Eğrisi

Modelin sınıflandırma performansını genel olarak değerlendiren AUC-ROC Skoru (AUC-ROC Score) değeri 0.6666 olarak ortalama üzerinde bir performans gösterdiği söylenebilir. Ancak mükemmel değil. AUC-ROC Skoru (AUC-ROC Score) eğrisi Şekil 5.11 de gösterilmiştir.



Şekil 5.12: LR Sınıflandırma Raporu (Classification Report) Isı Haritası

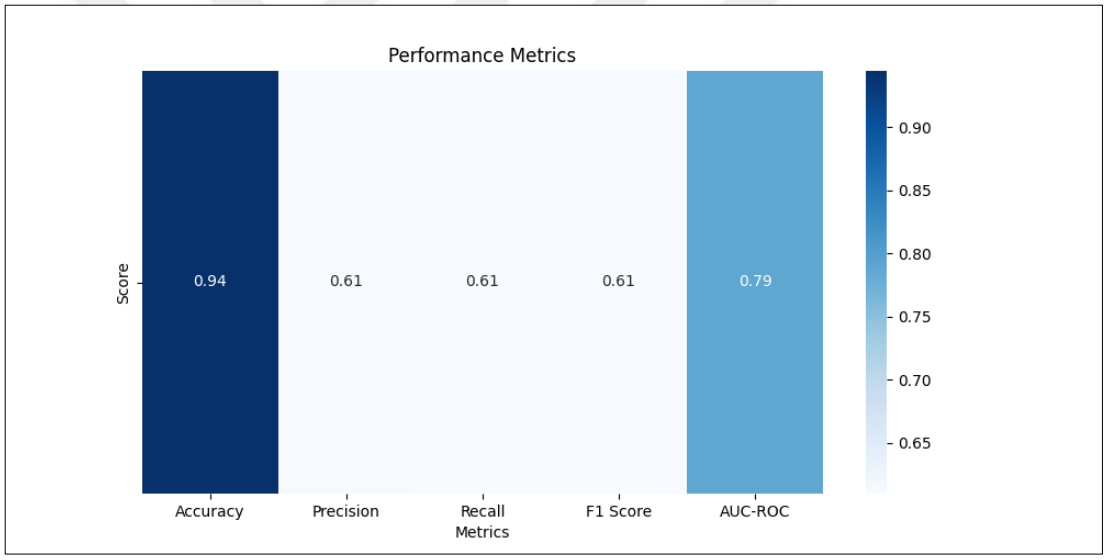
Genel olarak, model yüksek doğruluk ve kesinlik sunarken, duyarlılık (recall) oldukça düşük. Bu, özellikle pozitif vakaların tespitinde modelin zayıf olduğunu gösterir. Bu durum, modelin belirli bir sınıf üzerinde çok iyi performans gösterirken diğer sınıfı göz ardı etme eğiliminde olduğunu işaret edebilir. Modelin Sınıflandırma Raporu (Classification Report) değerlerini içeren ısı haritası Şekil 5.12 de gösterilmiştir.

5.5 MLP Performans Değerlendirme

Katmanlı Algılayıcılar (MLP) algoritması ile twitter-hate-speech veriseti üzerinde yapılan performans denemesi sonucunda ulaşılan metrik değerler Çizelge 5.5 de gösterilmektedir.

Çizelge 5.5: MLP Metrik Değerler

Metrik	**Değer**
Doğruluk (Accuracy)	0.9444705146253715
Kesinlik (Precision)	0.610989010989011
Duyarlılık (Recall)	0.6096491228070176
F1 Skoru (F1 Score)	0.610318331503842
AUC-ROC Skoru (AUC-ROC Score)	0.7899180429598504



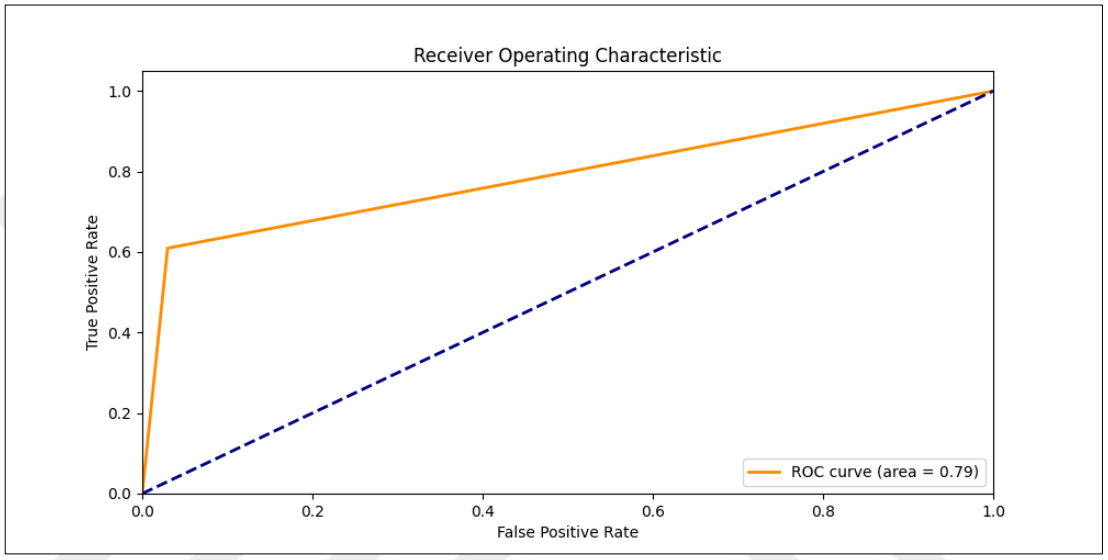
Şekil 5.13: MLP Metrik Değerler Isı Haritası

Modelin genel doğruluğunu gösteren Doğruluk (Accuracy) değeri 0.9444 olarak oldukça yüksek bir değere sahiptir. Bu, modelin tahminlerinin çoğunun gerçek değerlerle uyumlu olduğunu göstermektedir.

Modelin pozitif olarak sınıflandırdığı örneklerin ne kadarının gerçekten pozitif olduğunu gösteren Kesinlik (Precision) değeri 0.6109 olarak düşük bir değere sahiptir. Bu da modelin pozitif tahminlerinde hatalar yaptığını gösterir. Yani, model bazen negatif durumları pozitif olarak yanlış sınıflandırmaktadır.

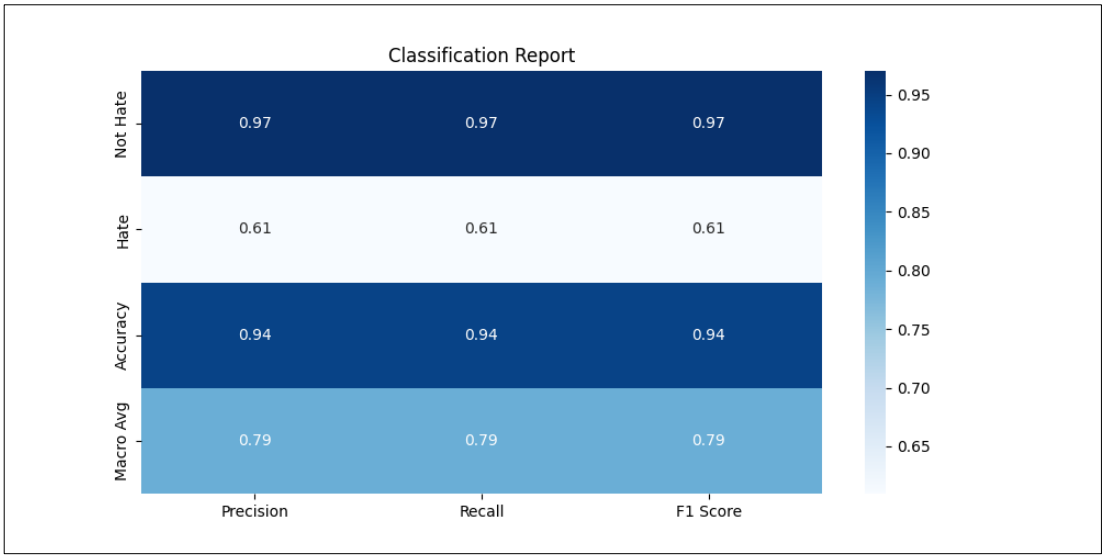
Gerçekte pozitif olan örneklerin ne kadarının model tarafından doğru olarak pozitif olarak sınıflandırıldığını gösteren Duyarlılık (Recall) değeri 0.6096 olarak orta düzeyde olduğu görülmektedir.

Kesinlik ve duyarlılığın harmonik ortalaması olan ve dengeli bir metrik olarak kabul edilen F1 Skoru (F1 Score) değeri 0.6103 olarak modelin kesinlik ve duyarlılık arasında dengeli, ancak mükemmel olmayan bir performans sergilediğini gösterir. Modelin metrik değerlerini içeren ısı haritası Şekil 5.13 de gösterilmiştir.



Şekil 5.14: MLP AUC-ROC Skoru (AUC-ROC Score) Eğrisi

Modelin sınıflandırma performansını genel olarak değerlendiren AUC-ROC Skoru (AUC-ROC Score) değeri 0.7899 olarak modelin iyi, ancak mükemmel olmayan bir sınıflandırma yeteneğine sahip olduğunu gösterir. AUC-ROC Skoru (AUC-ROC Score) eğrisi Şekil 5.14 de gösterilmiştir.



Şekil 5.15: MLP Sınıflandırma Raporu (Classification Report) Isı Haritası

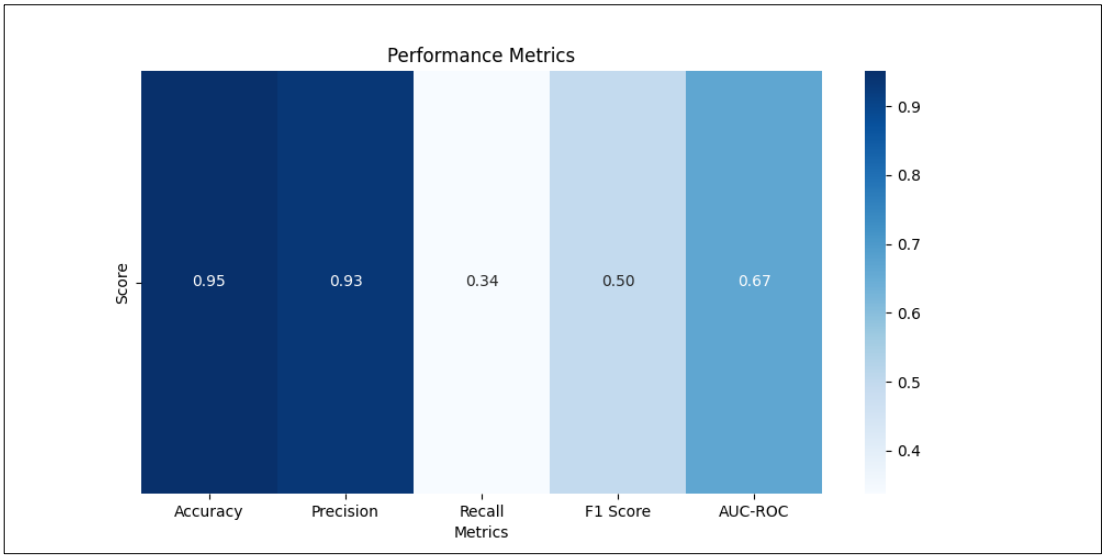
Genel olarak, model yüksek doğruluk oranına sahip ancak kesinlik, duyarlılık ve F1 skorları daha düşük, bu da modelin bazı sınıflandırma hataları yaptığını gösteriyor. AUC-ROC Skoru (AUC-ROC Score) değeri de modelin iyi bir performans gösterdiğini, ancak mükemmel olmadığını gösteriyor. Bu metrikler, modelin belirli türdeki hatalara daha yatkın olabileceğini gösterir. Modelin Sınıflandırma Raporu (Classification Report) değerlerini içeren ısı haritası Şekil 5.15 de gösterilmiştir.

5.6 MNB Performans Değerlendirme

Çok Terimli Naif Bayes (MNB) algoritması ile twitter-hate-speech veriseti üzerinde yapılan performans denemesi sonucunda ulaşılan metrik değerler Çizelge 5.6 da gösterilmektedir.

Çizelge 5.6: MNB Metrik Değerler

Metrik	**Değer**
Doğruluk (Accuracy)	0.9510402002189895
Kesinlik (Precision)	0.9333333333333333
Duyarlılık (Recall)	0.33771929824561403
F1 Skoru (F1 Score)	0.49597423510466987
AUC-ROC Skoru (AUC-ROC Score)	0.6679332553212238



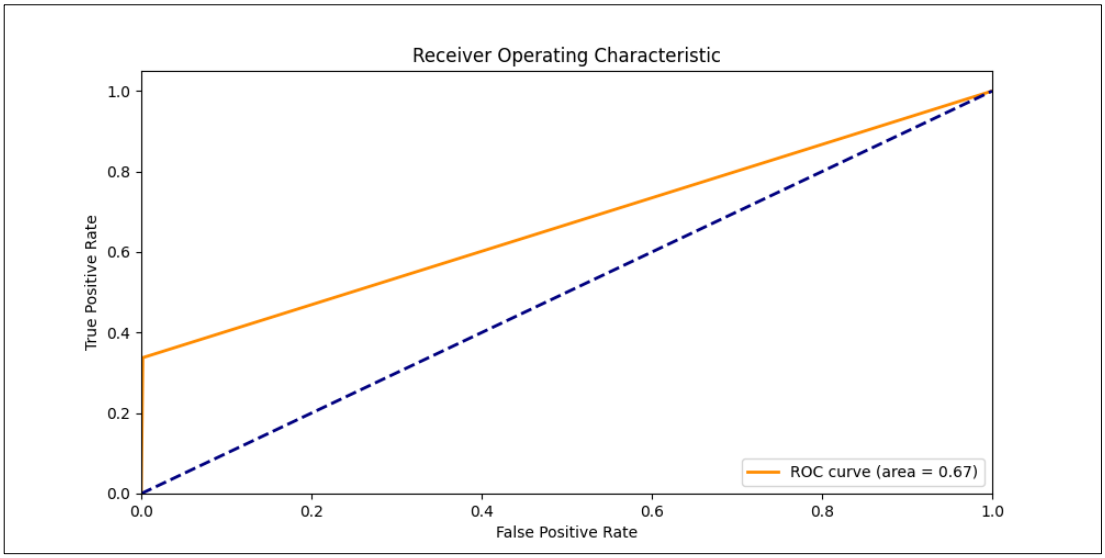
Şekil 5.16: MNB Metrik Değerler Isı Haritası

Modelin genel doğruluğunu gösteren Doğruluk (Accuracy) değeri 0.9510 olarak oldukça yüksek bir değere sahiptir. Bu, modelin tahminlerinin çoğunun gerçek değerlerle uyumlu olduğunu göstermektedir.

Modelin pozitif olarak sınıflandırdığı örneklerin ne kadarının gerçekten pozitif olduğunu gösteren Kesinlik (Precision) değeri 0.9333 olarak oldukça yüksek bir değere sahiptir. Bu da modelin pozitif tahminlerinin büyük çoğunluğunun doğru olduğunu gösterir.

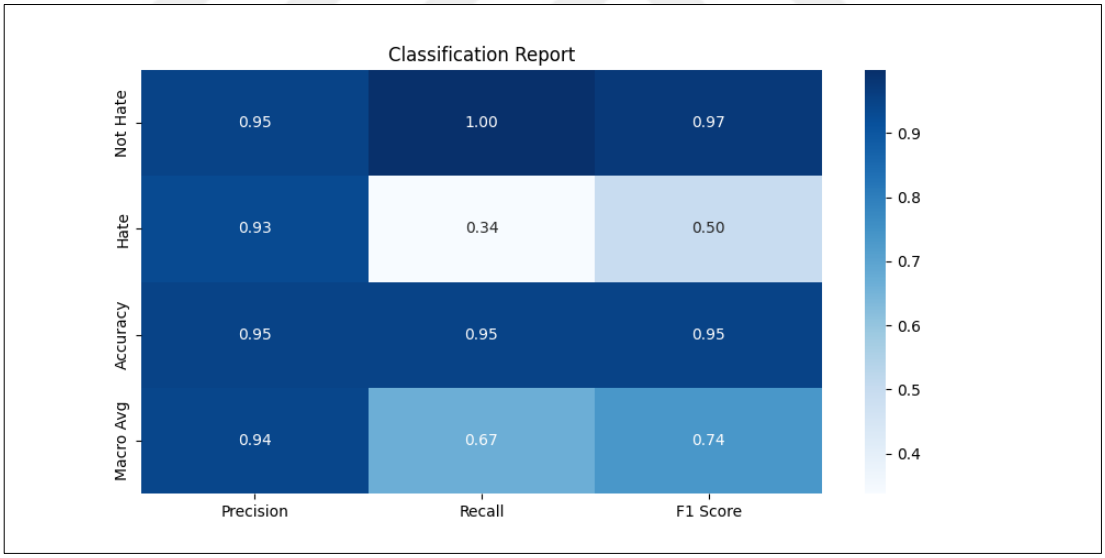
Gerçekte pozitif olan örneklerin ne kadarının model tarafından doğru olarak pozitif olarak sınıflandırıldığını gösteren Duyarlılık (Recall) değeri 0.3377 olarak oldukça düşük düzeyde olduğu görülmektedir. Yani modelin gerçekte pozitif olan durumların çoğunu kaçırdığını gösterir. Bu, modelin belirli bir sınıfı tespit etmede zayıf olduğunu işaret edebilir.

Kesinlik ve duyarlılığın harmonik ortalaması olan ve dengeli bir metrik olarak kabul edilen F1 Skoru (F1 Score) değeri 0.4959 olarak modelin kesinlik (precision) ve duyarlılık (recall) arasında orta düzeyde bir denge sağladığını ancak her iki metrikte de mükemmel olmadığını gösterir. Modelin metrik değerlerini içeren ısı haritası Şekil 5.16 da gösterilmiştir.



Şekil 5.17: MNB AUC-ROC Skoru (AUC-ROC Score) Eğrisi

Modelin sınıflandırma performansını genel olarak değerlendiren AUC-ROC Skoru (AUC-ROC Score) değeri 0.6679 olarak iyi olduğu söylenebilir. Ancak mükemmel değil. AUC-ROC Skoru (AUC-ROC Score) eğrisi Şekil 5.17 de gösterilmiştir.



Şekil 5.18: MNB Sınıflandırma Raporu (Classification Report) Isı Haritası

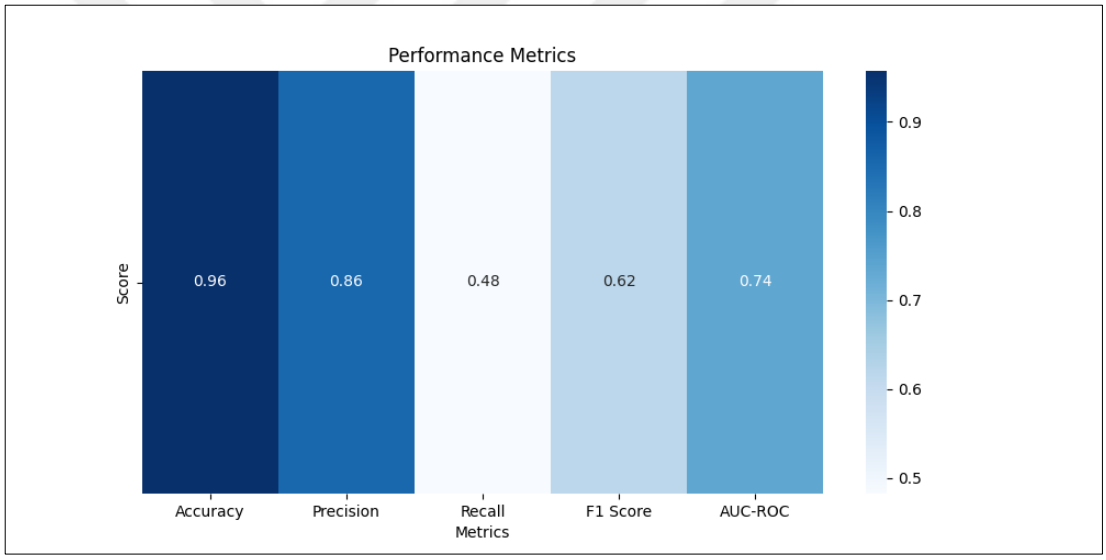
Genel olarak, bu modelin yüksek doğruluk ve kesinlik değerlerine sahip olduğu ancak duyarlılık (recall) ve AUC-ROC Skoru (AUC-ROC Score) değerlerinin nispeten düşük olduğu görülüyor. Bu, modelin belirli bir sınıfı tahmin etmede zorlanabileceği anlamına gelebilir. Modelin Sınıflandırma Raporu (Classification Report) değerlerini içeren ısı haritası Şekil 5.18 de gösterilmiştir.

5.7 RF Performans Değerlendirme

Rastgele Orman (RF) algoritması ile twitter-hate-speech veriseti üzerinde yapılan performans denemesi sonucunda ulaşılan metrik değerler Çizelge 5.7 de gösterilmektedir.

Çizelge 5.7: RF Metrik Değerler

Metrik	**Değer**
Doğruluk (Accuracy)	0.9572970436414828
Kesinlik (Precision)	0.8560311284046692
Duyarlılık (Recall)	0.4824561403508772
F1 Skoru (F1 Score)	0.6171107994389902
AUC-ROC Skoru (AUC-ROC Score)	0.7381120182973857



Şekil 5.19: RF Metrik Değerler Isı Haritası

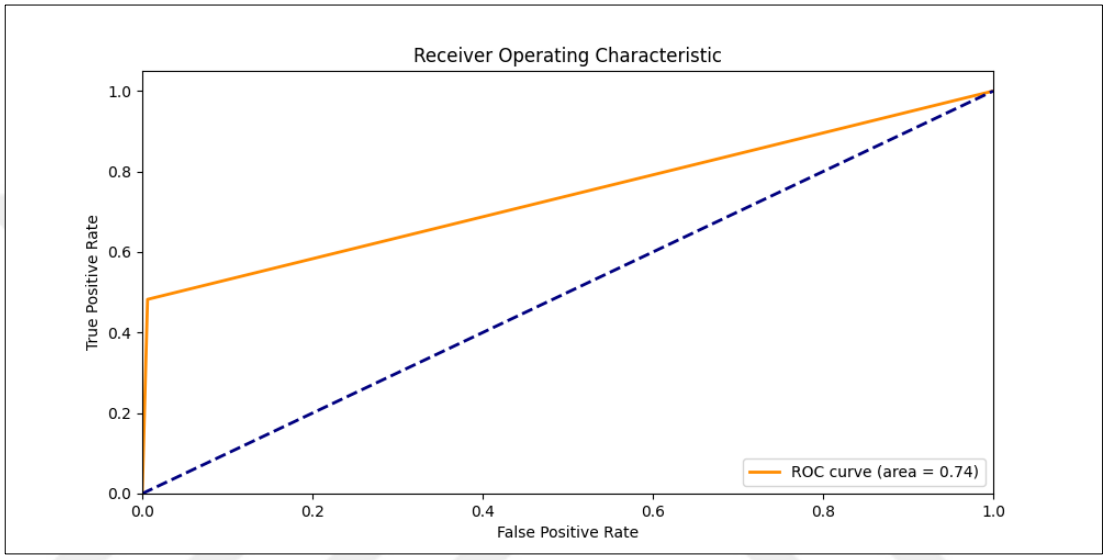
Modelin genel doğruluğunu gösteren Doğruluk (Accuracy) değeri 0.9573 olarak oldukça yüksek bir değere sahiptir. Bu, modelin tahminlerinin çoğunun gerçek değerlerle uyumlu olduğunu göstermektedir.

Modelin pozitif olarak sınıflandırdığı örneklerin ne kadarının gerçekten pozitif olduğunu gösteren Kesinlik (Precision) değeri 0.8560 olarak oldukça iyi bir değere sahiptir. Bu da modelin pozitif tahminlerinin önemli bir kısmının gerçekten pozitif olduğunu gösterir.

Gerçekte pozitif olan örneklerin ne kadarının model tarafından doğru olarak pozitif olarak sınıflandırıldığını gösteren Duyarlılık (Recall) değeri 0.4825 olarak

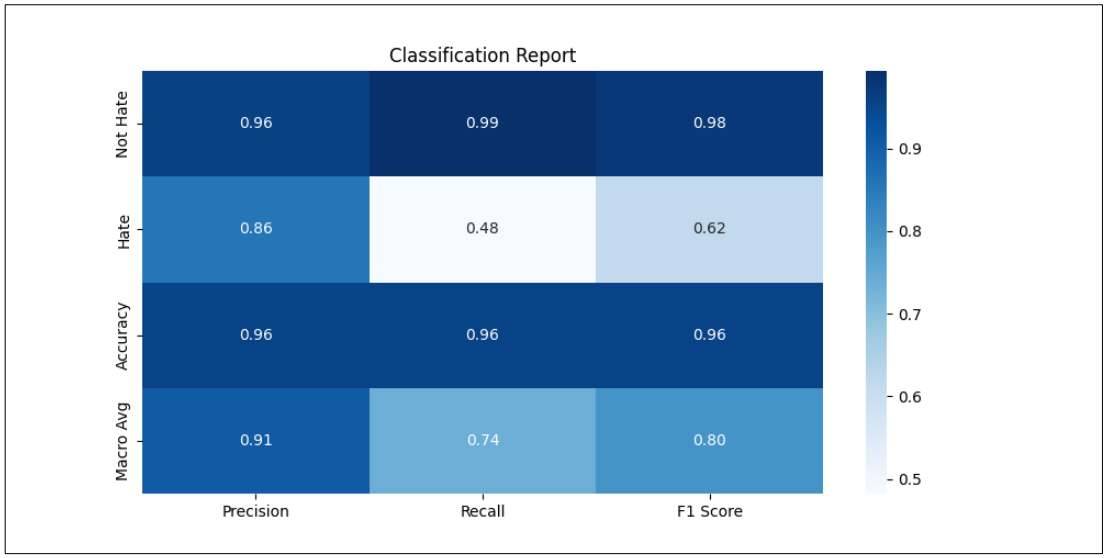
ortalamanın altı bir düzeyde olduğu görülmektedir. Yani modelin birçok gerçek pozitif vakayı kaçırdığı söylenebilir.

Kesinlik ve duyarlılığın harmonik ortalaması olan ve dengeli bir metrik olarak kabul edilen F1 Skoru (F1 Score) değeri 0.6171 olarak ne kesinlik ne de duyarlılık açısından mükemmel olmayan ancak orta seviyede bir denge sağladığını göstermektedir. Modelin metrik değerlerini içeren ısı haritası Şekil 5.19 da gösterilmiştir.



Şekil 5.20: RF AUC-ROC Skoru (AUC-ROC Score) Eğrisi

Modelin sınıflandırma performansını genel olarak değerlendiren AUC-ROC Skoru (AUC-ROC Score) değeri 0.7381 olarak oldukça iyi olduğu söylenebilir. Ancak mükemmel değil. AUC-ROC Skoru (AUC-ROC Score) eğrisi Şekil 5.20 de gösterilmiştir.



Şekil 5.21: RF Sınıflandırma Raporu (Classification Report) Isı Haritası

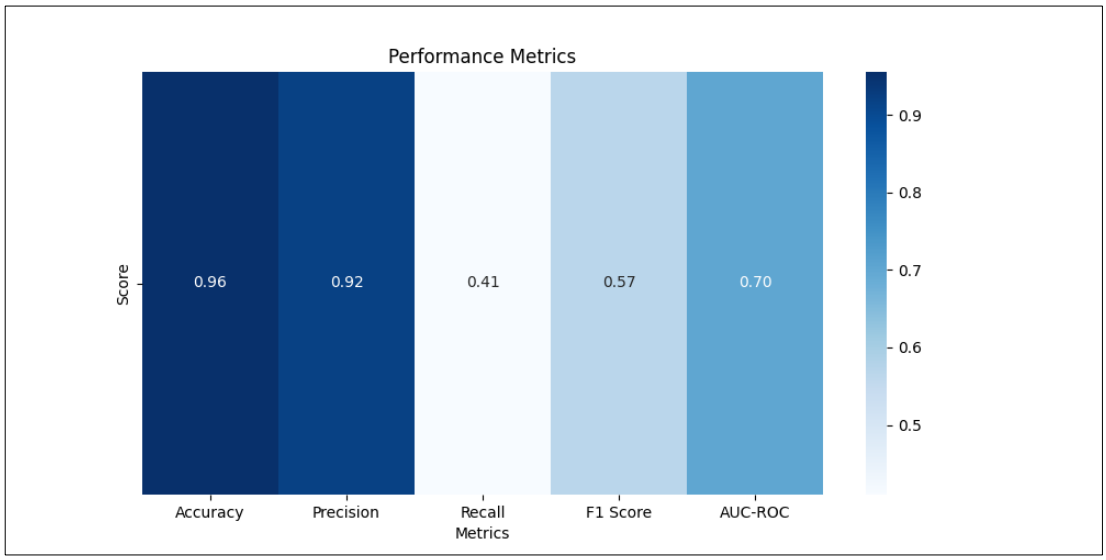
Genel olarak, bu model yüksek doğruluk oranına sahip olmakla birlikte, özellikle duyarlılık (recall) konusunda iyileştirilmesi gereken alanlar olduğunu göstermektedir. Özellikle, modelin gerçek pozitif vakaları kaçırma oranı (düşük duyarlılık) dikkate alınmalı ve bu yönde iyileştirmeler yapılmalıdır. Modelin Sınıflandırma Raporu (Classification Report) değerlerini içeren ısı haritası Şekil 5.21 de gösterilmiştir.

5.8 SVM Performans Değerlendirme

Destek Vektör Makineleri (SVM) algoritması ile twitter-hate-speech veriseti üzerinde yapılan performans denemesi sonucunda ulaşılan metrik değerler Çizelge 5.8 de gösterilmektedir.

Çizelge 5.8: SVM Metrik Değerler

Metrik	**Değer**
Doğruluk (Accuracy)	0.9554199906147348
Kesinlik (Precision)	0.9211822660098522
Duyarlılık (Recall)	0.4100877192982456
F1 Skoru (F1 Score)	0.56752655538695
AUC-ROC Skoru (AUC-ROC Score)	0.7036963777559109



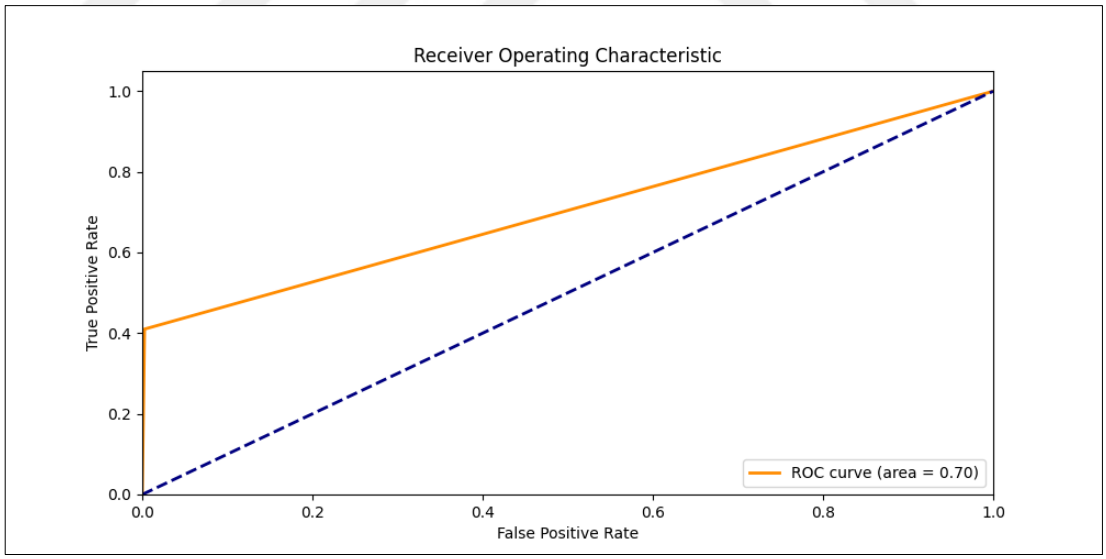
Şekil 5.22: SVM Metrik Değerler Isı Haritası

Modelin genel doğruluğunu gösteren Doğruluk (Accuracy) değeri 0.9554 olarak oldukça yüksek bir değere sahiptir. Bu, modelin tahminlerinin çoğunun gerçek değerlerle uyumlu olduğunu ve genel anlamda modelin başarılı olduğunu gösterir.

Modelin pozitif olarak sınıflandırdığı örneklerin ne kadarının gerçekten pozitif olduğunu gösteren Kesinlik (Precision) değeri 0.9212 olarak oldukça yüksek bir değere sahiptir. Bu da modelin pozitif tahminlerinin büyük çoğunluğunun doğru olduğunu gösterir.

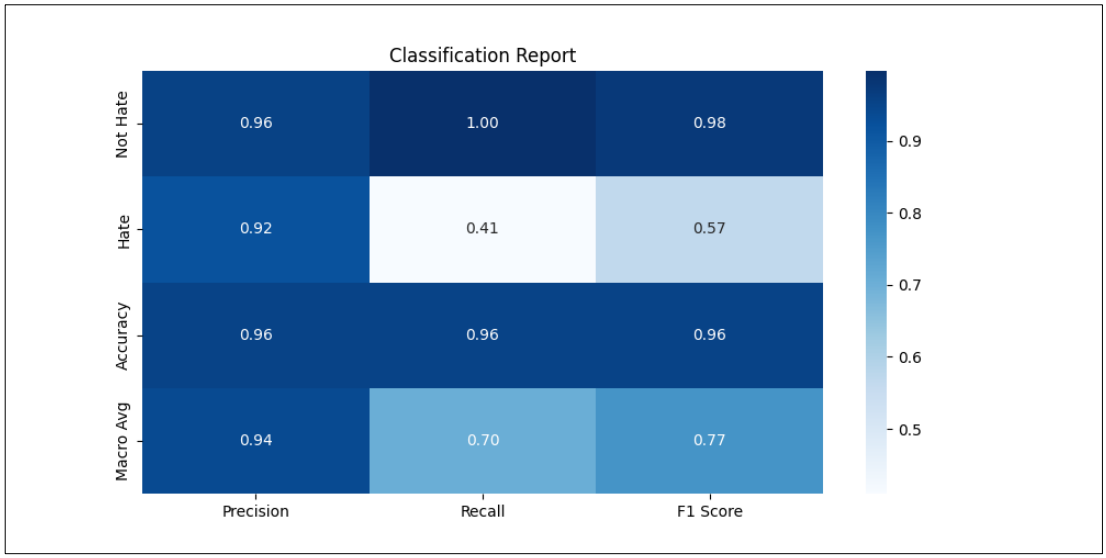
Gerçekte pozitif olan örneklerin ne kadarının model tarafından doğru olarak pozitif olarak sınıflandırıldığını gösteren Duyarlılık (Recall) değeri 0.4101 olarak düşük düzeyde olduğu görülmektedir. Yani modelin pozitif durumları tespit etmede zayıf olduğunu gösteriyor.

Kesinlik ve duyarlılığın harmonik ortalaması olan ve dengeli bir metrik olarak kabul edilen F1 Skoru (F1 Score) değeri 0.5675 olarak modelin kesinlik ve duyarlılık arasında orta düzeyde bir dengesi olduğunu gösterir. Yüksek kesinliğe rağmen düşük duyarlılık, F1 skorunu düşürmüştür. Modelin metrik değerlerini içeren ısı haritası Şekil 5.22 de gösterilmiştir.



Şekil 5.23: SVM AUC-ROC Skoru (AUC-ROC Score) Eğrisi

Modelin sınıflandırma performansını genel olarak değerlendiren AUC-ROC Skoru (AUC-ROC Score) değeri 0.7037 olarak oldukça iyi olduğu söylenebilir. Ancak mükemmel değil. AUC-ROC Skoru (AUC-ROC Score) eğrisi Şekil 5.23 de gösterilmiştir.

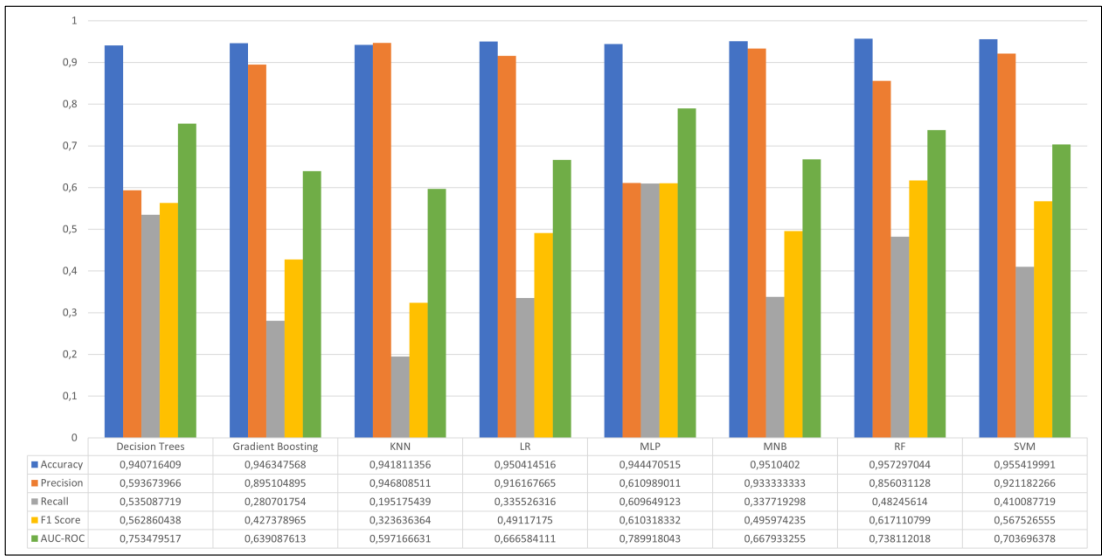


Şekil 5.24: SVM Sınıflandırma Raporu (Classification Report) Isı Haritası

Sonuç olarak, bu SVM modeli yüksek doğruluk ve kesinlikle genel olarak iyi performans gösterse de, düşük duyarlılık oranı bazı pozitif durumların kaçırılmasına yol açabilir. Özellikle duyarlılığın önemli olduğu durumlarda, modelin iyileştirilmesi gerekebilir. Modelin Sınıflandırma Raporu (Classification Report) değerlerini içeren ısı haritası Şekil 5.24 de gösterilmiştir.

5.9 Tüm Algoritmalar Genel Değerlendirme

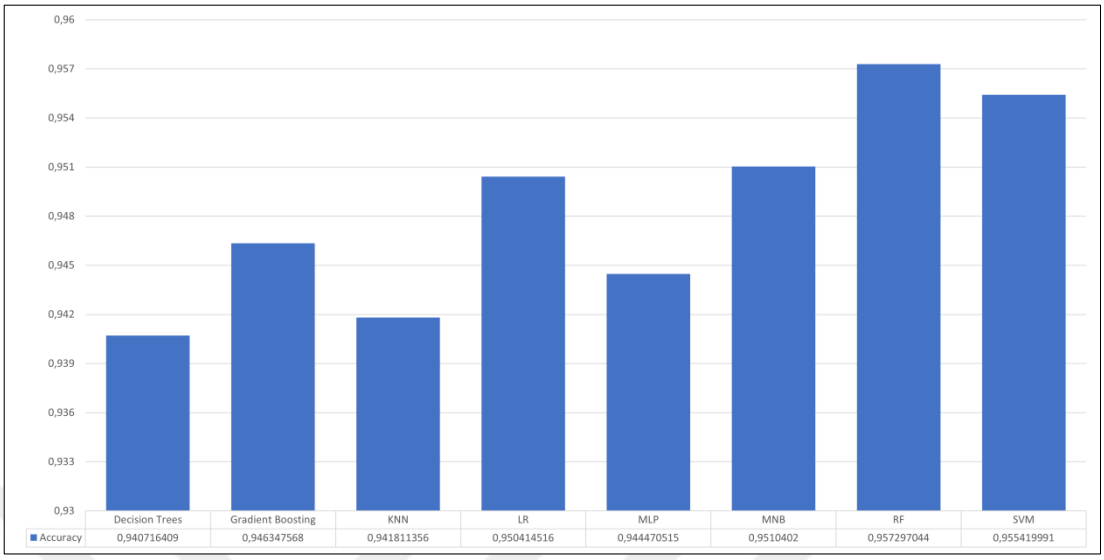
Performans değerlendirme testlerinde kullanılan tüm algoritmaların metrik değerleri Şekil 5.25 de gösterilmiştir.



	Decision Trees	Gradient Boosting	KNN	LR	MLP	MNB	RF	SVM
Accuracy	0,940716409	0,946347568	0,941811356	0,950414516	0,944470515	0,9510402	0,957297044	0,955419991
Precision	0,593673966	0,895104895	0,946808511	0,916167665	0,610989011	0,933333333	0,856031128	0,921182266
Recall	0,535087719	0,280701754	0,195175439	0,335526316	0,609649123	0,337719298	0,48245614	0,410087719
F1 Score	0,562860438	0,427378965	0,323636364	0,49117175	0,610318332	0,495974235	0,617110799	0,567526555
AUC-ROC	0,753479517	0,639087613	0,597166631	0,666584111	0,789918043	0,667933255	0,738112018	0,703696378

Şekil 5.25: Tüm Algoritmaların Metrik Değerleri

Performans değerlendirme testlerinde kullanılan tüm algoritmaların Doğruluk (Accuracy) değerleri Şekil 5.26 da gösterilmiştir.



Şekil 5.26: Tüm Algoritmaların Doğruluk (Accuracy) Değerleri

Rastgele Orman (RF) en yüksek doğruluk değerine sahipken en düşük doğruluk değeri Karar Ağaçları (DT) algoritmasına aittir.

6. SONUÇ

Bu çalışmada, çeşitli makine öğrenimi algoritmalarının, Twitter'da nefret söylemi içeren tweetleri sınıflandırma performanslarının değerlendirilmesi üzerine odaklanmıştır. Bu algoritmalar arasında Karar Ağaçları (DT), Gradyan Artırma (Gradient Boosting), K-En Yakın Komşu (KNN), Lojistik Regresyon (LR), Çok Katmanlı Algılayıcılar (MLP), Çok Terimli Naif Bayes (MNB), Rastgele Orman (RF) ve Destek Vektör Makineleri (SVM) yer almaktadır. Bu algoritmaların her biri için Doğruluk (Accuracy), Kesinlik (Precision), Duyarlılık (Recall), F1 Skoru (F1 Score) ve AUC-ROC Skoru (AUC-ROC Score) metriklerine göre performansları değerlendirilmiştir. Her algoritma için ön işleme, model tanımlama, eğitme, değerlendirme ve sonuçların yorumlanması aşamaları gerçekleştirilmiştir.

Bulgular, genel olarak, bu algoritmaların oldukça yüksek doğruluk (accuracy) oranlarına sahip olduğunu göstermektedir. Bu, modellerin tahminlerinin çoğunun gerçek değerlerle uyumlu olduğunu gösterir ve genel olarak modellerin başarılı olduğunu işaret eder. Öte yandan, duyarlılık (recall) ve F1 skorlarının genellikle daha düşük olduğu gözlemlenmiştir, bu da modellerin gerçek pozitif durumları tespit etmede bazı zorluklar yaşadığını gösterir.

Kesinlik (Precision) metrikleri, modellerin yanlış pozitifleri (yanlış alarm) oldukça düşük oranda tuttuğunu gösterirken, düşük duyarlılık (recall) değerleri, gerçek pozitif vakaların önemli bir kısmının kaçırıldığını göstermektedir. Bu, özellikle pozitif vakaların tespiti açısından modellerin iyileştirilmesi gerektiğini işaret etmektedir. F1 skorları, modellerin kesinlik (precision) ve duyarlılık (recall) arasında dengeli ancak mükemmel olmayan bir performans sergilediğini göstermektedir.

AUC-ROC Skoru (AUC-ROC Score) değerleri, modellerin sınıflandırma performanslarının ortalama üzerinde olduğunu gösterse de, hiçbir modelin mükemmel bir performans sergilemediğini işaret etmektedir. Bu durum, modellerin belirli türdeki hatalara daha yatkın olabileceğini ve bu yüzden belirli senaryolarda dikkatli kullanılması gerektiğini göstermektedir.

Tüm algoritmalar ortak çalıştırıldığında, genel olarak yüksek doğruluk ve kesinlik elde edilmiş, ancak düşük duyarlılık (recall) değerleri bazı pozitif durumların kaçırılmasına yol açmıştır.

Her algoritmanın performansını kısaca özetlemek gerekirse:

Karar Ağaçları (DT): Yüksek Doğruluk (Accuracy) ile başarılı tahminler yapmış ancak düşük Kesinlik (Precision) ve Duyarlılık (Recall) değerleriyle belirli sınıflandırmalarda zayıf kalmıştır.

Gradyan Artırma (Gradient Boosting): Yüksek Doğruluk (Accuracy) ve Kesinlik (Precision) ile dikkat çekmiş, ancak düşük Duyarlılık (Recall) değeri bazı gerçek pozitif vakaların kaçırılmasına yol açmıştır.

K-En Yakın Komşu (KNN): Yüksek Doğruluk (Accuracy) ve Kesinlik (Precision) göstermiş, fakat çok düşük Duyarlılık (Recall) değeriyle gerçek pozitif vakaların çoğunu kaçırmıştır.

Lojistik Regresyon (LR): Yüksek Doğruluk (Accuracy) ve iyi Kesinlik (Precision) ile güvenilir tahminler yapmış, ancak düşük Duyarlılık (Recall) oranıyla bazı pozitif vakaları gözden kaçırmıştır.

Çok Katmanlı Algılayıcılar (MLP): Yüksek Doğruluk (Accuracy) göstermiş ancak Kesinlik (Precision), Duyarlılık (Recall) ve F1 Skoru (F1 Score) değerleri daha düşük kalmıştır.

Çok Terimli Naif Bayes (MNB): Çok yüksek Doğruluk (Accuracy) ve Kesinlik (Precision) ile başarılı tahminler yapmış, ancak yine düşük Duyarlılık (Recall) değeriyle bazı pozitif durumları tespit edememiştir.

Rastgele Orman (RF): En yüksek Doğruluk (Accuracy) ve iyi Kesinlik (Precision) ile iyi genel performans göstermiş, ancak düşük Duyarlılık (Recall) oranıyla bazı pozitif vakaları kaçırmıştır.

Destek Vektör Makineleri (SVM): İkinci en yüksek Doğruluk (Accuracy) ve Kesinlik (Precision) göstermiş ancak düşük Duyarlılık (Recall) oranıyla bazı pozitif vakaları gözden kaçırmıştır.

Bu çalışmanın sonuçları, nefret söylemi sınıflandırmasının karmaşık doğası ve makine öğrenimi modellerinin bu alandaki potansiyel etkinlikleri üzerine önemli içgörüler sağlamaktadır. Özellikle, çeşitli algoritmaların farklı türdeki veri yapısına

ve ifade biçimlerine olan duyarlılıkları, bu alanın zorluklarını ve algoritmaların iyileştirilmesi için potansiyel alanları ortaya koymaktadır.

Model Seçimi ve İyileştirme: Bulgular, her algoritmanın farklı senaryolarda farklı derecede etkili olduğunu göstermektedir. Örneğin, Rastgele Orman ve Gradyan Artırma algoritmaları genel olarak yüksek doğruluk ve F1 skorları ile dikkat çekerken, Çok Terimli Naif Bayes ve K-En Yakın Komşu algoritmaları daha düşük performans sergilemiştir. Bu, model seçiminin önemini ve spesifik veri setlerine uygun algoritmaların seçilmesi gerekliliğini vurgulamaktadır.

Veri Ön İşleme ve Özellik Mühendisliği: Nefret söylemi sınıflandırmasında, veri ön işleme ve özellik mühendisliği oldukça önemlidir. Metin verilerinin temizlenmesi, önemli özelliklerin çıkarılması ve veri dengesizliğinin giderilmesi gibi faktörler, model performansını önemli ölçüde etkileyebilir. Bu çalışma, bu faktörlerin her bir modelin performansını nasıl etkilediğini detaylı bir şekilde incelenmesini gerektirir.

Gerçek Dünya Uygulamaları ve Etik Hususlar: Nefret söylemi sınıflandırması, hem toplumsal hem de etik açıdan önemli bir konudur. Bu modellerin gerçek dünya senaryolarında nasıl kullanılabileceği, özellikle yanlış pozitif ve yanlış negatif sonuçların potansiyel sonuçları açısından dikkatlice değerlendirilmelidir. Örneğin, yanlış pozitif bir sınıflandırma, masum bir kullanıcının ifade özgürlüğünün kısıtlanmasına yol açabilirken, yanlış negatif bir sonuç, zararlı içeriğin gözden kaçmasına neden olabilir.

Model Karşılaştırması ve Hiperparametre Optimizasyonu: Bu çalışma, farklı makine öğrenimi algoritmalarının karşılaştırılmasını içermekte ve her bir modelin hiperparametrelerinin optimizasyonunun önemini vurgulamaktadır. Hiperparametre ayarlarının incelikli bir şekilde yapılması, modelin veriye daha iyi uyum sağlamasına ve dolayısıyla daha yüksek performans elde etmesine olanak sağlar. Özellikle, nefret söylemi gibi hassas ve dinamik bir konuda, modelin veriye uygunluğunu artırmak için sürekli ayarlamalar yapılması gerekmektedir.

Veri Kaynaklarının Çeşitliliği ve Kapsamı: Nefret söylemi sınıflandırmasında kullanılan veri setlerinin çeşitliliği ve kapsamı, modelin genel uygulanabilirliğini ve güvenilirliğini önemli ölçüde etkiler. Farklı demografik gruplardan, coğrafi bölgelerden ve sosyal medya platformlarından elde edilen verilerin kullanılması,

modelin daha kapsayıcı ve objektif olmasını sağlar. Bu tür çalışmalarda, kullanılan veri setinin sınırlamalarını ve bunların model sonuçları üzerindeki potansiyel etkileri de göz önünde bulundurulmalıdır.

Algılanan Riskler ve Sosyal Sorumluluk: Nefret söylemi sınıflandırmasının potansiyel riskleri ve sosyal sorumlulukları da bu çalışmanın önemli bir parçasıdır. Otomatik sınıflandırma sistemlerinin yanlışlıkla ifade özgürlüğünü kısıtlama veya belirli gruplara karşı önyargı oluşturma riskleri göz önünde bulundurulmalıdır. Bu riskleri azaltmak için, modelin karar süreçlerinin şeffaf ve adil olması sağlanmalı ve insan denetimi mekanizmaları entegre edilmelidir.

Uygulama ve Entegrasyon: Bu algoritmalara dayalı sistemlerin pratikte nasıl uygulanacağı ve mevcut sosyal medya platformlarına nasıl entegre edileceği önemlidir. Gerçek zamanlı sınıflandırma sistemleri, kullanıcı deneyimini etkilemeden, zararlı içeriği etkili bir şekilde engelleyebilmelidir. Bu, teknolojik altyapı, kullanıcı arayüzü tasarımı ve performans optimizasyonu gibi bir dizi teknik ve kullanıcı deneyimi sorunlarını da beraberinde getirir.

Sürekli Öğrenme ve Adaptasyon: Son olarak, nefret söylemi ve sosyal medya platformlarının sürekli değişen doğası göz önünde bulundurulduğunda, makine öğrenimi algoritmalarının sürekli öğrenme ve adaptasyon yetenekleri hayati öneme sahiptir. Algoritmaların güncel verilere dayalı olarak kendilerini sürekli olarak yeniden eğitmeleri ve yeni türlerdeki nefret söylemine adapte olmaları gerekir. Bu, hem teknolojik gelişmeleri hem de sosyal değişimleri yansıtacak şekilde modellerin sürekli güncellenmesini gerektirir.

Sonuç olarak, bu çalışma, Twitter üzerinde nefret söylemi içeren içeriklerin tespiti için çeşitli makine öğrenimi algoritmalarının etkili bir şekilde kullanılabileceğini göstermektedir. Ancak, her bir algoritmanın kendi avantajları ve sınırlılıkları olduğu ve bu yüzden dikkatli bir şekilde seçilip uygulanması gerektiği de açıktır. Algoritmaların daha etkin ve sorumlu bir şekilde kullanılabilmesi için sürekli iyileştirilmesi ve etik standartlara uygun şekilde uygulanması gerekmektedir. Bu algoritmaların daha da iyileştirilmesi ve farklı veri kümeleri üzerinde test edilmesi, nefret söylemi içeren içeriklerin tespiti ve engellenmesi konusunda daha etkili sonuçlar doğurabilir. Bu nedenle, daha dengeli ve etkin bir model geliştirmek için algoritmaların birleştirilmesi veya geliştirilmesi gerektiği sonucuna varılmıştır.

Bu tespitler, Twitter'da nefret söylemini tespit edebilecek daha güçlü ve dengeli modellerin geliştirilmesi için yol gösterici olabilir.

Bu tezde incelenen algoritmalar, nefret söylemi tespiti alanında önemli bir adım atsa da, bu alanda sürekli gelişim ve dikkatli düşünce gereklidir. Nefret söylemi tespiti, sadece teknolojik bir mesele değil, aynı zamanda toplumsal, kültürel ve etik bir sorumluluktur.

Gelecek Çalışmalar: Bu çalışma, bu alanda gelecekte yapılacak çalışmalar için önemli bir temel oluşturmakta ve makine öğrenimi uygulamalarının sosyal medya içeriğinin yönetimindeki potansiyelini ortaya koymaktadır. Gelecek çalışmalar, farklı dillerdeki veri kümeleri, sosyal medyanın değişen doğası ve nefret söyleminin evrimi üzerine odaklanabilir. Ayrıca, bu algoritmaların sınıflandırma kararlarının yorumlanabilirliği ve şeffaflığı üzerine çalışmalar da büyük önem taşımaktadır. Model kararlarının nedenleri ve bu kararların sosyal etkileri hakkında daha fazla anlayış kazanmak, bu teknolojinin daha sorumlu bir şekilde kullanılmasını sağlayacaktır.

KAYNAKÇA

Agresti, A. (2007). "*An introduction to categorical data analysis*". John Wiley & Sons.

Alatawi, R. A., & Lee, K. (2020). Hate speech detection on Twitter using deep learning: A comparative analysis. *IEEE Access, 8*, 170752-170760.

Albright, R., Lanfranchi, A., Fredriksen, A., Styve, G., & Warner, C. (2019). Error analysis in an automated machine learning (AutoML) framework. In Proceedings of the 2019 AAAI/ACM Conference on AI, *Ethics, and Society* (ss. 339-344). ACM.

Alfina, I., Mulia, R. A., & Fanany, M. I. (2020). Active Learning for Hate Speech Detection on Twitter. *In 2020 International Conference on Advanced Computer Science and Information Systems (ICACSIS)* (ss. 105-112). IEEE.

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2), 211-236.

Allport, G. W. (1954). *The nature of prejudice*. Cambridge, MA: Addison-Wesley.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82-115.

Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). *Deep learning for hate speech detection in tweets.* Proceedings of the 26th International Conference on World Wide Web Companion, 759-760.

Banks, J. A. (2006). R*ace, culture, and education: The selected works of James A. Banks.* Routledge.

Barret, Z., Santoro, A., Montgomery, H., Lillicrap, T., & Lerchner, A. (2020). *Open-ended learning in symmetric zero-sum games*. arXiv preprint arXiv:2007.13544.

Bartlett, J., & Krasodomski-Jones, A. (2015). *Counter-speech: Examining content that challenges extremism online*. Demos.

Bengio, Y., Courville, A., & Vincent, P. (2016). Representation learning: a review and new perspectives. IEEE transactions on pattern analysis and machine intelligence, 35(8), 1798-1828.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(Feb), 281-305.

Bleich, E. (2011). *The freedom to be racist? How the United States and Europe struggle to preserve freedom and combat racism*. Oxford University Press. London.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *Proceedings of the fifth annual workshop on Computational learning theory*, 144-152.

Bostrom, N., & Yudkowsky, E. (2014). The ethics of artificial intelligence. *In The Cambridge handbook of artificial intelligenc*e (ss. 316-334). Cambridge University Press.

Boyd, D.M. ve Ellison, N.B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1), 210-230.

Branković, S. (2018). Fighting hate speech or restricting freedom of expression: Limits and possibilities. *Filozofija I Društvo*, 29(3), 316-327.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R.A. (1986). *Classification and regression trees*. CRC Press.

Brown, A. (2017). What is hate speech? Part 1: The myth of hate. *Law and Philosophy*, 36(4), 419-468.

Brown, A. (2018). What is hate speech? Part 1: The myth of hate. *Law and Philosophy*, 37(4), 419-468.

Bryson, J. J. (2018). Patiency is not a virtue: the design of intelligent systems and systems of ethics. *Ethics and Information Technology*, 20(1), 15-26.

Burnap, P., & Williams, M. L. (2015). Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2), 223-242.

Castells, M. (2012). *Networks of Outrage and Hope: Social Movements in the Internet Age.* Polity.

Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., & Floridi, L. (2018). Artificial intelligence and the 'good society': the US, EU, and UK approach. *Science And Engineering Ethics*, 24(2), 505-528.

Chen, Y., Zhou, Y., Zhu, S., & Xu, H. (2018). Detecting offensive language in social media to protect adolescent online safety. Proceedings of the 2012 ASE/IEEE International Conference on Privacy, Security, *Risk and Trust*, 71-80.

Chetty, N., & Alathur, S. (2018). Hate speech review in the context of online social networks. *Aggression and Violent Behavior*, 40, 108-118.

Coppola, R., Morisio, M., Torchiano, M., & Vardanega, T. (2020). A survey on teaching artificial intelligence in higher education. *Journal of Ambient Intelligence and Humanized Computing*, 11(4), 1615-1634.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1), 21-27.

Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, 88(11), 2783-2792.

Darling, K. (2016). *Extending legal rights to social robots*. We Robot Conference.

Daugherty, P. R., & Wilson, H. J. (2018). *Human+ machine: reimagining work in the age of AI.* Harvard Business Review Press.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *In Proceedings of the 11th International Conference on Web and Social Media*, ICWSM 2017, ss. 512-515.

Debatin, B., Lovejoy, J. P., Horn, A. K., & Hughes, B. N. (2009). Facebook and online privacy: Attitudes, behaviors, and unintended consequences. *Journal of Computer-Mediated Communication*, 15(1), 83-108.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding*. arXiv preprint arXiv:1810.04805.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: *Human Language Technologie*s, Volume 1 (Long and Short Papers) (ss. 4171-4186).

Dietterich, T. G. (2000). Ensemble methods in machine learning. In International workshop on multiple classifier systems. *Springer*, (ss. 1-15). Berlin, Heidelberg.

Dixon, L., Li, J., Sorensen, J., Thain, N., & Vasserman, L. (2018). Measuring and mitigating unintended bias in text classification. In Proceedings of the 2018 AAAI/ACM Conference on AI, *Ethics, and Society* (ss. 67-73). ACM.

Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., & Bhamidipati, N. (2015). Hate speech detection with comment embeddings. *In Proceedings of the 24th International Conference on World Wide Web*, ss. 29-30. ACM.

Ellison, N.B., Steinfield, C. ve Lampe, C. (2007). The benefits of Facebook "friends:" Social capital and college students' use of online social network sites. *Journal of Computer-Mediated Communication*, 12(4), 1143-1168.

ElSherief, M., Kulkarni, V., Nguyen, D., Wang, W. Y., & Belding, E. (2018). Hate lingo*: A target-based linguistic analysis of hate speech in social media.* Proceedings of the International AAAI Conference on Web and Social Media, 12(1), 292-302.

ElSherief, M., Nilizadeh, S., Nguyen, D., Vigna, G., & Belding, E. (2018). Peer to peer hate: *Hate speech instigators and their targets*. In Twelfth International Conference on Web and Social Media, ICWSM 2018, ss. 52-61.

Fairclough, N. (1989). *Language and power*. Longman.

Fogel, J., & Nehmad, E. (2009). Internet social network communities: Risk taking, trust, and privacy concerns. *Computers in Human Behavior*, 25(1), 153-160.

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys* (CSUR), 51(4), 1-30.

Founta, A. M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., ... & Kourtellis, N. (2018). Large scale crowdsourcing and characterization of Twitter abusive behavior. *In Proceedings of the Twelfth International Conference on Web and Social Media*, ICWSM 2018, ss. 609-661.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1), 119-139.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.

Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015*). Countering online hate speech*. UNESCO Publishing.

Gambäck, B., & Sikdar, U. K. (2017). Using convolutional neural networks to classify hate-speech. *In Proceedings of the First Workshop on Abusive Language Online*, ss. 85-90.

Gao, L., Huang, R., & Huang, L. (2017). Detecting online hate speech using context aware models. *In Proceedings of the International Conference on Recent Advances in Natural Language Processing*, RANLP 2017, ss. 260-266.

Gelber, K. (2013). Free speech versus hate speech: Policy arguments beyond the harm principle. *Australian Journal of Political Science*, 48(2), 151-166.

Gelber, K., & McNamara, L. J. (2016). Efficacy of the use of online hate speech laws in Australia and Germany. *Journal of Media Law*, 8(2), 199-223.

Goldberg, D. E. (1989). *Genetic algorithms in search, optimization, and machine learnin*g. Addison-Wesley.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Green, D. P., McFalls, L. H., & Smith, J. K. (2001). Hate crime: An emergent research agenda. *Annual Review of Sociology*, 27(1), 479-504.

Greenhow, C., Robelia, B., & Hughes, J. E. (2009). Learning, teaching, and scholarship in a digital age: Web 2.0 and classroom research: What path should we take now? *Educational Researcher*, 38(4), 246-259.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Survey*s (CSUR), 51(5), 1-42.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). T*he elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, 770-778.

Hechenbichler, K., & Schliep, K. (2004). *Weighted k-nearest-neighbor techniques and ordinal classification.* Ludwig Maximilian University Munich, Munich, Germany.

Heinze, E. (2016). H*ate speech and democratic citizenship*. Oxford University Press.

Hobbs, R. (2010). *Digital and media literacy: A plan of action*. The Aspen Institute.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5), 359-366.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.

Junco, R. (2012). The relationship between frequency of Facebook use, participation in Facebook activities, and student engagement. *Computers & Education*, 58(1), 162-171.

Kaplan, A.M. ve Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of social media. *Business Horizons*, 53(1), 59-68.

Keats Citron, D. (2014). *Hate crimes in cyberspace*. Harvard University Press.

Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2015). Fu*ndamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studie*s. MIT Press.

Kohavi, R. (1995). *A study of cross-validation and bootstrap for accuracy estimation and model selection.* In Proceedings of the 14th International Joint Conference on Artificial Intelligence, Morgan Kaufmann, (ss. 1137-1143).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. I*n Advances in neural information processing systems* (ss. 1097-1105).

Kuss, D.J. ve Griffiths, M.D. (2011). Online social networking and addiction—a review of the psychological literature. *International Journal of Environmental Research and Public Health*, 8(9), 3528-3552.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4), 541-551.

LeCun, Y., Bottou, L., Orr, G. B., & Müller, K. R. (1998). Efficient backprop. In Neural networks: Tricks of the trade. *Springer*, Berlin, Heidelberg. (ss. 9-50).

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, 2(3), 18-22.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arXiv preprint arXiv:1907.11692.

Livingstone, S. (2004). Media literacy and the challenge of new information and communication technologies. *The Communication Review*, 7(1), 3-14.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

Matamoros-Fernández, A. (2017). Platformed racism: The mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube. *Information, Communication & Society*, 20(6), 930-946.

McCallum, A., & Nigam, K. (1998). A comparison of event models for naive Bayes text classification. *AAAI-98 workshop on learning for text categorization*, 752, 41-48.

McCarthy, J. (1959). *Programs with common sense. Proceedings of the Teddington Conference on the Mechanization of Thought Processes*, 75, 77-84.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space.* In Proceedings of the International Conference on Learning Representations.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *In Advances in Neural Information Processing Systems*, ss. 3111-3119.

Minsky, A., Cymberknop, L., & Pustilnik, F. (2021). The AI classroom: teaching artificial intelligence to primary and secondary education students. In AIED 2021: Artificial Intelligence in Education. *Springer*, Cham., (ss. 485-498).

Mitchell, T. M. (1997). *Machine learning*. McGraw Hill.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... & Petersen, S. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540), 529-533.

Möller, I., & Krahé, B. (2009). Exposure to violent video games and aggression in German adolescents: A longitudinal analysis. *Aggressive Behavior*, 35(1), 75-89.

Newell, A., & Simon, H. A. (1963). *GPS, a program that simulates human thought. Computers and thought*, 279-293.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Prentice-Hall.

Ng, A. (2017). *Why AI is the new electricity*. Stanford Business.

Nicolelis, M. A., & Lebedev, M. A. (2009). Principles of neural ensemble physiology underlying the operation of brain-machine interfaces. *Nature Reviews Neuroscience*, 10(7), 530-540.

Nilsson, N. J. (1998). *Artificial intelligence: a new synthesis*. San Francisco: Morgan Kaufmann Publishers.

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. IEEE Transactions on knowledge and data engineering, 22(10), 1345-1359.

Park, J. H., & Fung, P. (2017). One-step and two-step classification for abusive language detection on Twitter. *Proceedings of the First Workshop on Abusive Language Online*, 41-45.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (ss. 1532-1543).

Perry, B. (2001). *In the name of hate: Understanding hate crimes*. New York: Routledge.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: *Human Language Technologies*, 1, 2227-2237.

Pettigrew, T. F., & Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. *Journal of Personality and Social Psychology*, 90(5), 751-783.

Poole, D. L., & Mackworth, A. K. (2017). *Artificial Intelligence: Foundations of Computational Agents*. Cambridge University Press.

Quinlan, J.R. (1986). Induction of decision trees. *Machine learning*, 1(1), 81-106.

Ribeiro, M. H., Calais, P. H., Santos, Y. G., Almeida, V. A., & Meira Jr, W. (2018). Characterizing and detecting hateful users on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1), 676-679.

Richardson, J. E. (2010). *Analysing newspapers: An approach from critical discourse analysis*. Palgrave Macmillan.

Ruder, S. (2019). *Neural transfer learning for natural language processing*. Ph.D. thesis, National University of Ireland, Galway.

Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel distributed processing: explorations in the microstructure of cognition*, vol. 1: foundations. MIT Press.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088), 533-536.

Rummery, G. A., & Niranjan, M. (1994). *On-line Q-learning using connectionist systems* (Vol. 37). University of Cambridge, Department of Engineering.

Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: a modern approach. Malaysia*; Pearson Education Limited.

Saha, K., Patwa, P., & Solorio, T. (2018). *A hierarchically-labeled Twitter dataset for online harassment research.* In Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, ss. 663-666.

Salminen, J., Almerekhi, H., Milenković, M., Jung, S. G., An, J., Kwak, H., & Jansen, B. J. (2018). *Anatomy of online hate: Developing a taxonomy*

and machine learning models for identifying and classifying hate in online news media. In Twelfth International Conference on Web and Social Media, ICWSM 2018, ss. 330-339

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The risk of racial bias in hate speech detection. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1668-1678.

Sap, M., Park, G., Eichstaedt, J. C., Kern, M. L., Stillwell, D. J., Kosinski, M., ... & Schwartz, H. A. (2019). *Developing age and gender predictive lexica over social media.* In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, ss. 1146-1151.

Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. *In Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media* (ss. 1-10).

Schölkopf, B., & Smola, A. J. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press.

Sherif, M. (1966). *Group conflict and co-operation: Their social psychology*. Routledge.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., ... & Lillicrap, T. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362(6419), 1140-1144.

Singh, R., Rajput, S., & Bhatt, R. (2020). Detection of hate speech in social media using BERT-based deep learning models. *International Journal of Knowledge Discovery in Bioinformatics*, 2(1), 1-11.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427-437.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Suzor, N., Van Geelen, T., & West, S. (2019). *The role of intermediaries in tackling hate speech online.* In L. A. Hagen & R. Kuhn (Eds.), Intermediary liability and freedom of expression in the EU: From concepts to safeguards (ss. 281-303). Edward Elgar Publishing.

Tai, K. S., Socher, R., & Manning, C. D. (2015). Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks. *In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (Volume 1: Long Papers) (ss. 1556-1566).

Tajfel, H., & Turner, J. C. (1986). The social identity theory of inter-group behavior. In S. Worchel & L. W. Austin (Eds.), *Psychology of Intergroup Relations* (ss. 7-24). Nelson-Hall.

Tufekci, Z. (2018). Algorithms of oppression: How search engines reinforce racism. Information, *Communication & Society*, 21(3), 412-426.

van Dijk, T. A. (1993). Principles of critical discourse analysis. *Discourse & Society*, 4(2), 249-283.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer.

Vidgen, B., & Derczynski, L. (2020). Directions in abusive language training data: *Garbage in, garbage out. Proceedings of the Fourth Workshop on Online Abuse and Harms*, 62-70.

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., & diğ,. (2016). Matching networks for one shot learning. *In Advances in neural information processing systems* (ss. 3630-3638).

Waldron, J. (2012). *The harm in hate speech*. Harvard University Press.

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. *In Proceedings of the NAACL student research workshop* (ss. 88-93).

Watkins, C. J., & Dayan, P. (1992). Q-learning. *Machine learning*, 8(3-4), 279-292.

Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G., & Vaughan, T. M. (2002). Brain-computer interfaces for communication and control. *Clinical neurophysiology*, 113(6), 767-791.

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? *In Advances in neural information processing systems* (ss. 3320-3328).

Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning-based natural language processing. *IEEE Computational Intelligence Magazine*, 13(3), 55-75.

Zhang, H. (2004). The optimality of naive Bayes. *AA*, 1(2), 3.

Zhang, X., & Wallace, B. C. (2015). *A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification*. arXiv preprint arXiv:1510.03820.

Zhang, Y., Qi, P., & Manning, C. D. (2017). *Graph Convolution over Pruned Dependency Trees Improves Relation Extraction.* In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (ss. 2205-2215).

Zhang, Z., & Luo, L. (2018). Hate speech detection: A solved problem? The challenging case of long tail on Twitter. *Semantic Web*, 10(5), 925-945.

Zhang, Z., Robinson, D., & Tepper, J. (2018). *Detecting hate speech on Twitter using a convolution-GRU based deep neural network*. In European Semantic Web Conference, ss. 745-760.

Zhao, Y., & Mao, Y. (2020). Identifying Hate Speech in Social Networks Using Machine Learning Algorithms. *In International Conference on Big Data and Blockchain* (ss. 117-128). Springer, Singapore.

ÖZGEÇMİŞ

ÖĞRENİM DURUMU

- İstanbul Gedik Üniversitesi, Yapay Zeka Mühendisliği Tezli YL, 2022-…
- Eskişehir Anadolu Üniversitesi, MIS-Yönetim Bilişim Sistemleri, 2017-2020
- Sakarya Üniversitesi, Eğitim Fakültesi, 2003-2008

MESLEKİ DENEYİM

- İstanbul Gedik Üniversitesi, Bilgi İşlem Daire Başkanlığı, 2014-…
- Çemsan End. Proses ve Otomasyon A.Ş., Bilgi İşlem Departmanı, 2011-2013
- Arifiye İlçe Milli Eğitim Müdürlüğü, Ücretli Öğretmenlik, 2010-2011