Van Yüzüncü Yıl Üniversitesi Sosyal Bilimler Enstitüsü Dergisi
Van Yüzüncü Yıl University The Journal of Social Sciences Institute
e-ISSN: 2822 - 3136
Yıl / Year: 2025 - Sayı / Issue: 67

Sosyal Medyada Dezenformasyonla Mücadelede Yapay Zekâ: Olanaklar ve Sınırlılıklar

Artificial Intelligence in Combating Disinformation on Social Media: Possibilities and Limitations

Ergin SARI
Lecturer Dr., Van Yüzüncü Yıl University, Department of Audiovisual Techniques and Media Production, Van, Türkiye
Öğretim Görevlisi Dr., Van Yüzüncü Yıl Üniversitesi, Van, Türkiye
Orcid: 0000-0002-7956-1390 erginsari@yyu.edu.tr

Article Information/Makale Bilgisi

Cite as/Atıf: Sarı, E. (2025). Artificial intelligence in combating disinformation on social media: Opportunities and limitations. *Van Yüzüncü Yıl University The Journal of Social Sciences Institute*, 67, 49-61
Sarı, E. (2025). Sosyal medyada dezenformasyonla mücadelede yapay zekâ: Olanaklar ve sınırlılıklar. *Van Yüzüncü Yıl Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 67, 49-61

Article Types / Makale Türü: Research Article/Araştırma Makalesi
Received/Geliş Tarihi: November 11, 2024/11 Kasım 2024
Accepted/Kabul Tarihi: December 29, 2024/29 Aralık 2024
Published/Yayın Tarihi: March 18, 2025/18 Mart 2025
Pub Date Season/Yayın Sezonu: March/Mart
Issue/Sayı: 67
Pages/Sayfa: 49-61
Plagiarism/İntihal: This article has been reviewed by at least two referees and scanned via a plagiarism software./ Bu makale, en az iki hakem tarafından incelendi ve intihal içermediği teyit edildi.
Published by/Yayıncı: Van Yüzüncü Yıl University of Social Sciences Institute/Van Yüzüncü Yıl Üniversitesi Sosyal Bilimler Enstitüsü
Ethical Statement/Etik Beyan: It is declared that scientific and ethical principles have been followed while carrying out and writing this study and that all the sources used have been properly cited/ Bu çalışmanın hazırlanma sürecinde bilimsel ve etik ilkelere uyulduğu ve yararlanılan tüm çalışmaların kaynakçada belirtildiği beyan olunur (Ergin SARI).
Conflict of Interest/Çıkar Beyanı
There are no conflicts of interest./Bu çalışma kapsamında herhangi bir kurum, kuruluş, kişi ile çıkar çatışması yoktur.
Declaration of Authors' Contribution/Yazarların Katkı Oran Beyanı
This article has one author and the contribution rate of the author is 100 %/Bu makale tek yazarlıdır ve yazarın katkı oranı yüzde 100'dür.

Öz

Bu çalışma, sosyal medyada yalan haberlerin önlenmesinde yapay zekâ destekli sistemlerin rolünü, sunduğu olanakları ve sınırlılıklarını kapsamlı bir şekilde incelemektedir. Dijitalleşme ve sosyal medya platformlarının yaygınlaşması, bilgiye erişimi kolaylaştırırken aynı zamanda yanıltıcı ve manipülatif içeriklerin geniş kitlelere hızla ulaşmasına da zemin hazırlamaktadır. Yapay zekâ tabanlı makine öğrenimi (ML) ve doğal dil işleme (NLP) teknikleri, dezenformatif içeriklerin tespiti ve yayılmasının engellenmesi için güçlü araçlar sunmaktadır. Ancak, bu sistemlerin uygulanmasında etik, tarafsızlık, şeffaflık eksiklikleri ve yanlış pozitifler gibi önemli sınırlılıklar da öne çıkmaktadır. Çalışmamız, yapay zekâ tabanlı dezenformasyon tespit sistemlerinin sunduğu olanakları ve karşılaştığı etik, sosyal ve teknik zorlukları ele almakta; toplumsal güvenin sağlanması ve bilgi ekosisteminin sürdürülebilirliği için daha şeffaf ve hesap verebilir yapılar geliştirilmesi gerektiğini vurgulamaktadır. Çalışmanın bulguları, yapay zekânın toplumsal bilgi güvenliği ve doğru bilgiye erişim sağlama açısından önemini ortaya koymakta ve bu teknolojilerin daha sorumlu ve etik bir şekilde geliştirilmesine yönelik öneriler sunmaktadır.

Anahtar Kelimeler

Dezenformasyon, sosyal medya, yapay zekâ, yankı odaları, algoritmalar

Abstract

This study comprehensively examines the role, opportunities and limitations of artificial intelligence-supported systems in preventing fake news in social media. While digitalization and the proliferation of social media platforms facilitate access to information, they also pave the way for misleading and manipulative content to reach large audiences rapidly. Artificial intelligence-based machine learning (ML) and natural language processing (NLP) techniques offer powerful tools for detecting and preventing the spread of disinformative content. However, there are significant limitations in the application of these systems, such as ethics, impartiality, lack of transparency and false positives. Our study explores the opportunities offered by AI-based disinformation detection systems and the ethical, social and technical challenges they face, emphasizing the need to develop more transparent and accountable structures to ensure public trust and sustainability of the information ecosystem. The findings of the study reveal the importance of artificial intelligence for societal information security and access to accurate information, and provide recommendations for more responsible and ethical development of these technologies.

Keywords

Disinformation, social media, artificial intelligence, echo chambers, algorithms.

Giriş

Dijital çağ, bireylerin bilgiye erişiminde devrim niteliğinde değişiklikler yaratmış, ancak bu dönüşüm beraberinde ciddi riskleri de getirmiştir. Sosyal medya platformlarının sağladığı hızlı ve yaygın bilgi paylaşımı, toplumların bilgiye ulaşma biçimlerini kökten dönüştürürken, dezenformasyon gibi küresel düzeyde kaygı uyandıran sorunları da derinleştirmiştir. Manipülatif içeriklerin, toplumsal değerleri, bireysel algıları ve kolektif karar alma süreçlerini etkileme gücü, bu platformları yalnızca iletişim aracı olmaktan çıkarıp karmaşık bir sosyal olgu haline getirmiştir. Bu bağlamda, dezenformasyonun toplumsal yapı üzerindeki etkilerini yalnızca medya okuryazarlığı veya geleneksel doğrulama mekanizmalarıyla kontrol etmek mümkün görünmemektedir. Yapay zekâ tabanlı teknolojiler, dezenformasyonun tespiti ve yayılmasının önlenmesi için umut verici çözümler sunmakla birlikte, bu teknolojilerin uygulanabilirliği hem etik hem de teknik açılardan tartışmalıdır. Sosyal medyada dezenformasyonun etkilerinin ve yapay zekânın bu sorunları azaltmadaki rolünün derinlemesine incelenmesi, dijital çağın toplumsal dinamiklerini anlamak ve daha sağlam çözüm önerileri geliştirebilmek adına kritik bir gerekliliktir.

Dijital devrim; bilgi ve haberlerin üretimi, dağıtımı ve tüketiminde köklü değişiklikler yaratarak, bilginin yayılma hızını ve biçimini radikal şekilde dönüştürmüştür. Bilgiye erişimin demokratikleşmesi olarak görülen bu dönüşüm, özgür bilgi akışını sağlarken aynı zamanda dezenformasyonun hızla yayılmasını kolaylaştırarak toplumsal kutuplaşmayı derinleştirebilmektedir. Sosyal medya platformları, bireylerin bilgiye hızla erişmesini ve paylaşmasını sağlarken aynı zamanda dezenformasyonun yayılması için de elverişli bir ortam sunmaktadır. Dezenformasyon; kasıtlı olarak yanıltıcı ya da tamamen yanlış bilgilerin yayıldığı, genellikle politik, ekonomik veya ideolojik çıkarlar doğrultusunda manipülatif bir araç olarak kullanılan bir olgudur (Tandoc, Lim & Ling, 2018).

Cambridge Analytica skandalı ve COVID-19 pandemisi sırasında yanlış bilgilerin hızla yayılması, dezenformasyonun toplumsal düzeydeki ciddi sonuçlarını açıkça göstermiştir. Cambridge Analytica skandalı, sosyal medya platformu Facebook'tan izinsiz şekilde toplanan milyonlarca kullanıcının kişisel verilerinin, siyasi kampanyalarda manipülatif amaçlarla kullanıldığını ortaya çıkarmıştır. Bu skandal, 2016 ABD Başkanlık Seçimleri ve Brexit referandumu gibi önemli siyasi olaylarda seçmen davranışlarını manipüle ederek, bu olayların sonuçlarını etkilemiştir (M. Figlia vd., 2022). COVID-19 pandemisi süresince sosyal medyada yayılan yanlış bilgiler de toplum sağlığını doğrudan tehdit etmiştir. Aşıların zararlı olduğuna dair yanlış inançlar ve sahte tedavi yöntemleri gibi dezenformatif içerikler, bireylerin sağlık hizmetlerine duyduğu güveni sarsmıştır (Iddianto & Azi, 2022).

Uluslararası savaşlar ve toplumsal olaylar sırasında, dezenformasyonun yayılması önemli ölçüde artmaktadır. Bu tür durumlarda, sosyal medya platformları, yanlış bilgilerin ve propagandaların hızla geniş kitlelere ulaşması için güçlü bir araç haline gelmektedir. Örneğin; devam eden Rusya-Ukrayna savaşı sırasında, dezenformasyon uluslararası kamuoyunu etkilemek için etkin bir şekilde kullanılmaktadır. Yanıltıcı videolar ve propaganda içerikleri, savaşın gerçekliğini çarpıtarak kamuoyunu yanıltmakta ve savaşın algılanış biçimini manipüle etmektedir (Sharevski vd., 2022). Bu örnekler, dezenformasyonun yalnızca bireyler üzerinde değil, toplumsal ve uluslararası düzeyde de ciddi sonuçlar doğurabileceğini göstermektedir. Sosyal medya platformlarının yaygın kullanımı, manipülatif içeriklerin hızla geniş kitlelere ulaşmasını sağlamakta, bu da bilgi kirliliğini artırmaktadır. Yapay zekâ tabanlı çözümler, bu tür dezenformasyonu tespit etmek ve yayılmasını önlemek için kritik bir rol oynamaktadır. Ancak, bu süreçte teknik sınırlamalar ve etik zorluklar önemli engeller oluşturmaktadır (Chen vd., 2022). Sosyal medyanın algoritmik yapısı, kullanıcıların ilgi alanlarına uygun içeriklerle daha sık karşılaşmalarını ve bu içeriklerle etkileşimlerini artırmalarını sağlamaktadır. Bu durum, *yankı odaları (echo chambers)* ve *bilgi balonları (filter bubbles)* yaratarak yanlış bilgilerin daha da güçlenmesine ve yayılmasına zemin hazırlamaktadır. Algoritmalar, kullanıcıların önceki etkileşimlerine dayanarak benzer içerikler önerir, bu da kullanıcıların yalnızca belirli bir perspektif veya doğruluk derecesine sahip içeriklerle karşılaşmasına neden olabilir. Böylece, dezenformasyon hem bireylerin algılarını şekillendirmekte hem de toplumsal kutuplaşmayı artırmaktadır.

Bu bağlamda, yapay zekânın dezenformasyon akışlarını belirleme ve engelleme kapasitesi kritik bir önem taşımaktadır. Büyük veri analizi yetenekleri sayesinde yapay zekâ, kullanıcı davranışlarını ve içerik yayılımını inceleyerek yalan haberlerin tespiti ve önlenmesi sürecinde merkezi bir rol üstlenebilir. Özellikle doğal dil işleme (*Natural Language Processing - NLP*) ve makine öğrenimi (*Machine Learning - ML*) gibi teknolojiler, haber metinlerini analiz ederek içerikteki tutarsızlıkları, manipülatif dilleri ve güvenilir olmayan kaynakları tespit edebilme potansiyeline sahiptir (Shu vd., 2017). Bu ileri teknolojiler, sosyal medya platformlarında yalan haberlerin hızlı bir şekilde tanımlanmasını sağlayarak zararlı içeriklerin yayılmasını engellemede etkili bir araç olarak kullanılabilir. Facebook, X (Twitter) ve YouTube gibi platformlar, yapay zekâ destekli araçları kullanarak zararlı veya yanlış bilgi içeriklerini tespit etmeye ve yayılmalarını sınırlamaya çalışmaktadır (Vosoughi, vd., 2018). Bu platformlar, kullanıcıların karşılaştıkları içeriklerin doğruluğunu artırmak için çeşitli doğrulama ve filtreleme

mekanizmaları geliştirmektedir. Ancak, yapay zekânın yalan haberle mücadelede sunduğu olanaklar sadece teknik çözümlerle sınırlı değildir. Bu mücadele biçimi etik ve sosyal zorlukları da beraberinde getirmektedir. Yapay zekâ algoritmalarının tarafsızlık eksikliği, yanlış pozitif sonuçlara (örneğin, doğru haberlerin yalan olarak etiketlenmesi) yol açabilir. Bu durum, ifade özgürlüğü ve sansür arasındaki dengeyi tartışmaya açmaktadır. Yapay zekâ sistemlerinin bu bağlamda nasıl yönlendirileceği, hangi kriterlere göre çalışacağı ve karar mekanizmalarının nasıl şeffaflaştırılacağı gibi önemli etik sorunlar da ortaya çıkmaktadır (Brennen vd., 2020).

Yukarıdaki bilgilere ek olarak sosyal medyayla birlikte gelişim gösteren bir başka unsur vatandaş gazeteciliğidir. Vatandaş gazeteciliği, bireylerin sosyal medya üzerinden profesyonel gazetecilere benzer şekilde haber paylaşma potansiyelini ifade etmektedir. Vatandaş gazeteciliği, demokratikleşme açısından bilgiye erişimde büyük bir fırsat sunsa da doğrulama süreçlerinin eksikliği ve etik sorunlar nedeniyle yanlış bilgilerin yayılmasına zemin hazırlamaktadır. Vatandaş gazeteciliği, özellikle kriz anlarında hızlı bilgi aktarımı sağlasa da doğruluğu teyit edilmemiş bilgilerin yayılması, toplumsal yanlış anlamalara ve karar verme süreçlerinde yanıltıcı etkilere neden olmaktadır. Bu bağlamda, vatandaş gazeteciliğinin etik sorumlulukları ve doğrulama süreçleri eksikliği de ciddi bir sorunlar teşkil etmektedir.

Bu çalışma, yapay zekânın sosyal medyada dezenformasyonun tespiti ve önlenmesindeki rolünü, sunduğu olanaklar ve karşı karşıya olduğu sınırlılıklarla birlikte incelemeyi amaçlamaktadır.

1. Dezenformasyon ve Sosyal Medya

Dezenformasyon, kasıtlı olarak yanlış veya yanıltıcı bilgilerin yayılması olarak tanımlanır ve genellikle toplumu manipüle etmek veya kamuoyunu yanıltmak amacı taşır. Dezenformasyon, yalnızca bireysel yanlış bilgilendirme değil aynı zamanda toplumsal düzeyde algı yönetimini hedefleyen geniş çaplı bir stratejidir. Bu bağlamda Claire Wardle ve Hossein Derakhshan'a (2017) göre dezenformasyonun genel anlamda üç türü bulunmaktadır:

1. *Dezenformasyon (Disinformation)*: Kasıtlı olarak yanlış veya yanıltıcı bilgilerin yayılmasıdır ve genellikle manipülatif amaçlarla kullanılır.

2. *Mezenformasyon (Misinformation)*: Yanlış olan ancak kasıt içermeyen bilgi yayılımını ifade eder. Bilginin doğru sanılarak paylaşılması durumunu kapsar.

3. *Malenformasyon (Malinformation)*: Bilginin doğru olması ancak zarar verme amacıyla, manipülatif bir şekilde paylaşılmasıdır.

Bu üç kavram, bilgi akışında doğruluk kadar niyetin de önem taşıdığını göstermektedir. Dezenformasyon kasıtlı olarak yanıltma amacı güderken, mezenformasyon bilgi yanlış olsa da niyetin kötü olmadığı, genellikle bilgisizlikten kaynaklanan bir bilgi yayılımıdır. Malenformasyon ise doğru bilginin kötü niyetle, manipülatif veya zarar verme amacıyla kullanılmasıdır. Bu ayrımlar, özellikle sosyal medya ve çevrimiçi platformlarda, bilginin neden ve nasıl yayıldığını anlamak açısından kritik öneme sahiptir.

Bektaş da (2002) dezenformasyonu, kitle iletişim araçlarının özellikle kriz dönemlerinde kara propaganda amacıyla yanlış bilgi yaymak için kullanılması olarak tanımlar. Bu anlamda dezenformasyon, bilgi asimetrisi yaratır ve genellikle kriz dönemlerinde toplumsal kaosa yol açar.

Sosyal medya, dezenformasyonun hızla yayılmasına olanak sağlayan en etkili platformlardan biridir. Algoritmalar, kullanıcıların ilgisine göre içerikleri öne çıkararak, yankı odaları ve bilgi balonları oluşturur. Bu, kullanıcıların yalnızca kendi inançlarını destekleyen bilgilere maruz kalmasına ve farklı görüşlere erişimin kısıtlanmasına yol açar. Vosoughi ve arkadaşları (2018), sosyal medya üzerinde yanlış bilgilerin doğru bilgilere göre çok daha hızlı yayıldığını ve daha geniş kitlelere ulaştığını belirtmektedir. Bu durum nihayetinde toplumsal güvensizlik ve kutuplaşmayı derinleştirmektedir.

Sosyal medya, anonimlik ve hız faktörleriyle, doğruluk kontrolü yapılmadan bilgi paylaşımını kolaylaştırır. Erkan ve Ayhan (2018), sosyal medya platformlarının bu özelliğinin, dezenformasyonun yayılması için elverişli bir zemin hazırladığını ve kullanıcıların bilgi doğrulama süreçlerini zorlaştırdığını vurgulamaktadırlar. Örneğin; pandemi dönemi, dezenformasyonun etkisinin en belirgin olduğu dönemlerden biri olmuştur. COVID-19 sırasında sosyal medya, aşı karşıtlığı gibi yanlış bilgilerin yayılmasında önemli bir rol oynamıştır (Akyüz, 2021). Bu dönemde, doğru bilgiye ulaşmanın zorlukları ve halkın sağlık konularında yanlış yönlendirilmesi, dezenformasyonun ne kadar tehlikeli olabileceğini göstermiştir. Eroğlu (2023) ise sosyal medyada dezenformasyonun, özellikle seçim dönemlerinde toplumsal algıyı ve seçmen davranışlarını manipüle etmek için kullanıldığını, bu süreçte bilgi kirliliğinin seçmenlerin doğru karar vermesini zorlaştırdığını ifade etmektedir.

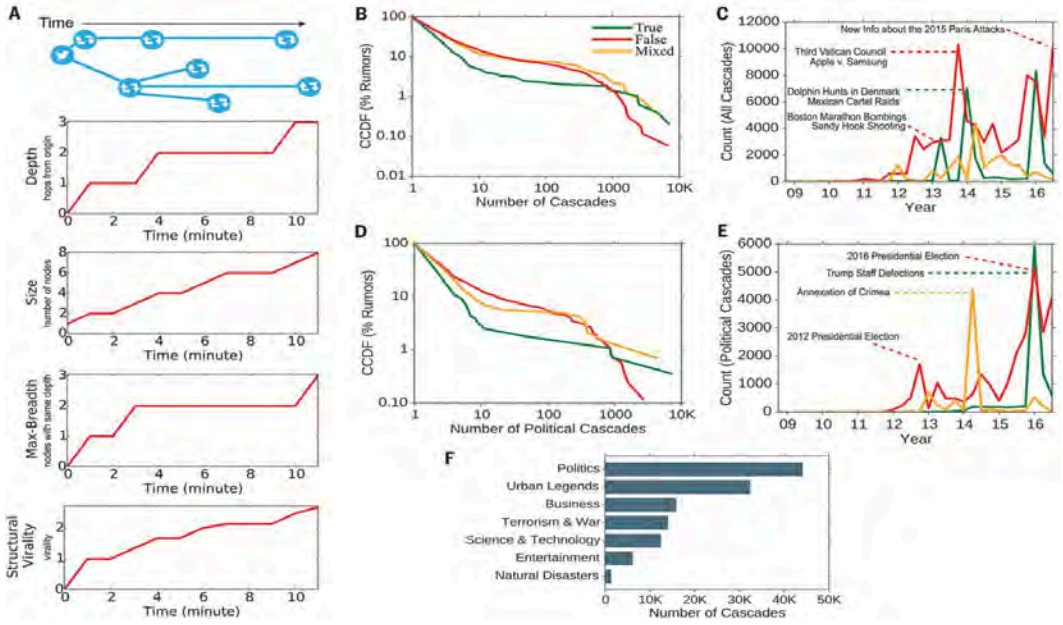
2. Sosyal Medya Platformlarında Dezenformasyonun Dinamikleri: Yankı Odaları ve Algoritmaların Rolü

Sosyal medya platformları, bilgi akışını hızlandırma ve geniş kitlelere erişme kapasiteleriyle modern iletişim ekosisteminin temel unsurları haline gelmiştir. Ancak platformların bu özellikleri aynı zamanda dezenformasyonun yayılmasında da kritik bir rol oynamaktadır. Dezenformasyonun yayılma mekanizmaları arasında yankı odaları ve algoritmaların oynadığı rol özellikle dikkat çekicidir. Sosyal medya platformları, bilgi akışını hızlandırmak ve kullanıcıların ilgisini çekmek amacıyla algoritmalar kullanır. Bu algoritmalar, kullanıcıların geçmiş davranışlarına dayanarak içerikleri kişiselleştirir, böylece yankı odaları ve filtre balonları oluşur. Yankı odaları, kullanıcıların yalnızca kendi inançlarını ve görüşlerini destekleyen içeriklerle etkileşimde bulunduğu, farklı bakış açılarına kapalı bir bilgi ekosistemi yaratır (Pariser, 2011). Bu durum, bireylerin sadece belirli bir perspektiften bilgi almasına ve diğer alternatif bilgilere ulaşmalarının zorlaşmasına neden olur (Ertürk, 2020).

Algoritmaların dezenformasyon üzerindeki etkisi, yanlış bilgilerin daha hızlı ve geniş kitlelere ulaşmasını sağlar. Vosoughi ve arkadaşları (2018) tarafından yapılan bir çalışma, sosyal medyada yanlış bilgilerin doğru bilgilere göre daha hızlı yayıldığını ve daha geniş bir kitleye ulaştığını ortaya koymuştur. Özellikle çarpıcı ve duygu yüklü dezenformatif içerikler, kullanıcıların dikkatini daha fazla çektiği için algoritmalar tarafından daha çok öne çıkarılmaktadır. Bu durum, toplumsal kutuplaşmayı derinleştirmekte ve demokratik süreçlere zarar verebilmektedir.

Tablo 1.

Söylenti şelalesi: Sosyal medyada doğru ve sahte haberlerin yayılma hızları, derinlikleri ve etkileri

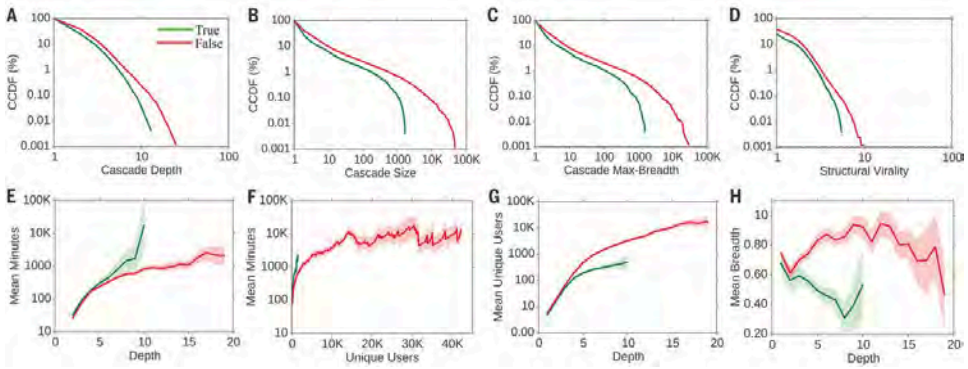


(Vosoughi vd., 2018)

Vosoughi ve arkadaşlarının (2018) çalışmasında yer alan bu tablo, sosyal medyada doğru ve sahte haberlerin yayılma hızlarını, derinliğini ve genişliğini kıyaslamalı olarak ortaya koymaktadır. Verilere göre, sahte haberler doğru haberlerden daha hızlı, daha derin ve daha geniş bir kitleye ulaşmaktadır. Özellikle B ve D grafiklerinde görüldüğü üzere sahte haberlerin cascade (yayılım) sayısı ve derinliği, doğru haberlere kıyasla belirgin şekilde daha yüksektir. C ve E grafiklerinde belirli olaylar (örneğin seçimler veya büyük krizler) etrafında sahte haberlerin yoğunlaştığı yani bu dönemlerde yanlış bilgilerin yayılma hızının arttığı dikkat çekmektedir. F grafiği ise, sahte haberlerin en çok siyaset, kentsel efsaneler ve iş dünyası gibi konular etrafında yoğunlaştığını göstermektedir. Bu bulgular, sahte haberlerin doğru haberlerden daha fazla dikkat çektiğini ve sosyal medya algoritmalarının bu tür içerikleri öne çıkarma eğiliminde olduğunu ortaya koymaktadır. Yine sözü edilen çalışmada yer alan başka bir tabloda ise sahte ve doğru haberlerin sosyal medyada yayılma dinamiklerini ayrıntılı olarak şu şekilde karşılaştırmaktadır:

Tablo 2.

Doğru ve yanlış söylenti basamaklarının tamamlayıcı kümülatif dağılım fonksiyonları



(Vosoughi vd., 2018)

Tablo 2'de yer alan grafiklere göre, sahte haberler, yayılma derinliği ve boyutu açısından daha yüksek değerlere ulaşmakta yani daha fazla kişi ve daha derin katmanlara yayılmaktadır. Ayrıca, yapısal virallik ve maksimum genişlik açısından da sahte haberlerin doğru haberlerden daha geniş bir alanda ve daha uzun süre boyunca dolaşımda kaldığı gözlemlenmektedir. Bu eğilim, sahte haberlerin sosyal medya platformlarında doğru bilgilere kıyasla daha hızlı, daha geniş kitlelere yayıldığını ve kullanıcılar arasında daha kalıcı bir etkileşim yarattığını göstermektedir.

Yankı odaları (echo chambers) da bireylerin kendi inançlarını sürekli pekiştiren bir çevrede kalmalarına neden olarak, toplumsal kutuplaşmayı artırır. Bu durum, bireylerin yanlış bilgileri daha kolay kabul etmesine ve bu bilgileri doğru gibi paylaşmasına yol açar (Sunstein, 2001). Bu süreçte, sosyal medya platformlarının tasarımı ve işleyişi, dezenformasyonun geniş çapta ve hızla yayılmasını kolaylaştırır. Dezenformasyonun sosyal medya aracılığıyla yayılması, sadece bireysel bilgi alımını değil toplumsal yapıyı ve demokratik süreçleri de etkiler. Bilgi kirliliği, bireylerin doğru karar vermesini zorlaştırır ve kamuoyunun yanlış yönlendirilmesine yol açar. Narin (2018) ve Açıkgöz & Sarı (2021), bu süreçlerin toplum üzerindeki etkilerini detaylandırarak, dezenformasyonun demokratik yapılar üzerindeki tehditlerini vurgulamaktadırlar.

Türkiye'de dezenformasyon ile mücadelede gönüllülük esasıyla çalışan ve bu konuda düzenli aralıklarla raporlar hazırlayan teyit.org'un 2023 Cumhurbaşkanlığı ve Milletvekili Seçimleri sürecine dair raporu, seçim döneminde sosyal medya platformlarında yayılan dezenformatif içeriklerin niteliğini kapsamlı bir şekilde ele almaktadır. Raporda, özellikle sosyal medyada hızla yayılan ve toplumun büyük kesimlerine ulaşan yanlış bilgilerin seçmen davranışları üzerindeki potansiyel etkileri analiz edilmiştir. Seçim sürecinde sosyal medyadaki haber akışının derinlemesine çözümlendiği raporda, *'deepfake'* gibi ileri seviyede manipülasyon teknikleri yerine daha basit dijital montaj ve yanıltıcı içerik üretim teknikleri olan *'cheapfake'* yöntemlerinin yaygın olarak kullanıldığına dikkat çekilmektedir. Teyit.org'un raporunda ayrıca bu tür içeriklerin sosyal medya aracılığıyla hızlı yayılım göstererek toplumsal kutuplaşmayı artırdığı ve seçmenlerin karar alma süreçlerini etkilediği vurgulanmaktadır. Teyit.org'un bu raporu, sosyal medya ve geleneksel medya kanallarında dezenformasyonun yayılım hızına dair veri sağlarken, dezenformasyonla mücadele kapsamında yapay zekâ destekli teknolojilerin eksikliklerini de gözler önüne sermektedir. Sözü edilen rapor, Türkiye'deki seçim sürecinde sıkça rastlanan dezenformatif içerik türlerini sınıflandırarak, bu içeriklerin toplum üzerinde yaratabileceği riskleri detaylandırmaktadır. Derin analizlere yer verilen raporda; '*cheapfake*' içeriklerin, ileri düzey teknik gerektirmeden, basit montaj veya bağlamdan koparılmış görüntü ve videolarla oluşturulmasının, bunların hızlıca üretilip yayılmasını kolaylaştırdığı vurgulanmaktadır. Raporda, bu içeriklerin seçmenlerin düşüncelerini şekillendirme potansiyeline sahip olduğu ve sosyal medya algoritmaları sayesinde geniş kitlelere hızla ulaştığına vurgu yapılmaktadır. Teyit.org'un detaylı çalışmasında yapay zekânın dezenformasyonla mücadeledeki yetersizlikleri de ele alınarak, toplumun yanıltıcı içeriklere karşı daha savunmasız hale geldiğinin altı çizilmektedir (Teyit.org, 2023).

3. Yapay Zekâ Tabanlı Dezenformasyon Tespit Yöntemleri

Yapay Zekâ, dezenformasyonun tespit edilmesi ve yayılmasının önlenmesi sürecinde kritik bir araç haline gelmiştir. Bu teknoloji, büyük veri setlerini analiz ederek yanlış bilgileri tanımlayabilir ve doğru bilgi akışını destekleyebilir. Makine öğrenimi, doğal dil işleme ve OLAP (*Online Analytical Processing*) küpü gibi yöntemler, sosyal medya platformlarında dolaşan içerikleri analiz ederek dezenformasyonun etkili bir şekilde tespit edilmesine yardımcı olabilir.

3.1. Makine Öğrenimi Tekniği

Makine öğrenimi (*machine learning*), dezenformasyonun tespitinde yaygın olarak kullanılan bir teknolojidir. Bu yöntem, büyük veri setlerinden öğrenme yaparak dezenformatif içeriği tespit eder. Makine öğrenimi algoritmaları, haber metinlerinin dil özelliklerini analiz ederek yanlış bilgiyi ayırt eder. Kai Shu ve arkadaşlarının (2017) çalışması, sosyal medya platformlarında dezenformasyon tespiti için kullanılan veri madenciliği yöntemlerini detaylandırmıştır. Bu çalışma, makine öğreniminin

dezenformasyonun dilsel ve içeriksel kalıplarını nasıl öğrendiğini göstermektedir. Michat Choraś ve arkadaşları da (2020), makine öğreniminin sahte haberlerin tespitinde etkin bir araç olduğunu vurgulamış ve bu alandaki mevcut teknikleri incelemiştir. Araştırma, sahte haberlerin yayılmasını önlemek için kullanılan makine öğrenimi metodolojilerini kapsamlı bir şekilde değerlendirmekte ve özellikle, algoritmaların verimliliğini artırmak için farklı tekniklerin entegrasyonuna dikkat çekmektedir. Michail Tsikerdekis ve Şerali Zeadally (2023), derin öğrenme yöntemlerini incelemiş ve bu teknolojilerin dezenformasyon tespitindeki etkinliğini ele almıştır. Derin öğrenme modelleri makine öğreniminin alt dalıdır ve karmaşık dil yapılarının anlaşılmasında ve manipülatif içeriklerin tespitinde önemli avantajlar sunmaktadır. Çalışmada, derin öğrenmenin özellikle hızlı değişen dezenformasyon biçimlerini tespit etmede nasıl etkili olduğu vurgulanmıştır.

3.2. Doğal Dil İşleme Teknikleri (NLP)

Doğal Dil İşleme (NLP) teknikleri, dezenformasyonun tespitinde giderek daha önemli bir rol oynamaktadır. Bu teknikler, metinlerdeki dil yapısını ve içerik özelliklerini analiz ederek, yanlış bilginin tespitine olanak tanır. NLP, denetim süreçlerinde büyük hacimli verilerin hızlı ve doğru bir şekilde analiz edilmesini sağlar. NLP, makinelerin insan dilini doğal olarak anlamasını sağlayan bir yapay zekâ teknolojisidir ve konuşma kalıplarını, söz dizimi yapılarını ve sözcük anlamlarını analiz ederek bilgisayarların insan dilleriyle etkileşim kurmasına olanak tanır (Merter & Özer, 2023, s. 273).

NLP yöntemi ile dezenformasyonun tespit edilebileceğini ortaya koyan çalışmalar da bulunmaktadır. Örneğin Katrina J. Ward ve arkadaşlarının (2022) çalışmaları, NLP'nin retorik araçları kullanarak dezenformasyonu nasıl tespit edebileceğini göstermektedir. Boris A. Galitsky (2015) ise, web madenciliği ve dil teknolojisini entegre ederek, metinlerin anlamsal yapılarını karşılaştırma yoluyla dezenformasyonun tespitine odaklanmıştır. Bu yöntem, sosyal ağ analizleriyle birlikte çalışarak dezenformatif içeriklerin daha geniş bağlamlarda tanımlanmasına imkân sağlamaktadır. Başka bir çalışmada da Haidar Rasyid ve arkadaşları (2023), 2024 Endonezya Cumhurbaşkanlığı Seçimleri sırasında Twitter'da yayılan dezenformasyonu tespit etmek amacıyla NLP tabanlı IndoBERT modelini kullanmışlardır. Araştırma, geleneksel makine öğrenimi teknikleri ile modern dönüştürücü tabanlı modelleri karşılaştırmış ve IndoBERT'in, %95 doğruluk oranıyla diğer modelleri geride bırakarak, dezenformasyon tespiti için en etkili model olduğunu göstermiştir. Çalışma, NLP'nin, sosyal medya platformlarında yayılan yanıltıcı içerikleri tanımlama ve dezenformasyonun yayılmasını önleme potansiyelini vurgulamaktadır.

Yukarıdaki çalışmalarda da görüleceği gibi, Doğal Dil İşleme teknikleri, dezenformasyonun tespitinde ve yayılmasının önlenmesinde kritik bir rol oynamaktadır. Bu teknikler, metinlerdeki dilsel ve anlamsal tutarsızlıkları belirleyerek, yanıltıcı içeriklerin hızla ve etkin bir şekilde tespit edilmesini sağlar. Sosyal medya gibi yoğun bilgi akışının olduğu platformlarda, NLP'nin sağladığı bu avantajlar, doğru bilgiye erişimi kolaylaştırarak toplumsal bilgi güvenliğini artırmakta ve dijital bilgi ortamının daha güvenilir bir hale gelmesine katkıda bulunmaktadır.

Bu iki tekniğin yanı sıra, OLAP (*Online Analytical Processing*) teknolojisi de dezenformasyonun çok boyutlu analizi için güçlü bir araç sunmaktadır. OLAP küpleri; zaman, yer ve içerik gibi farklı boyutlardan veri analizini mümkün kılarak dezenformatif içeriklerin kaynağını ve yayılma dinamiklerini daha derinlemesine inceleme imkânı tanır. Çimen ve Yüksel (2018), OLAP sistemlerinin hızlı veri analizi ve özetleme özellikleri sayesinde özellikle büyük ölçekli veri kümelerinde hızlı ve doğru sonuçlar ürettiğini vurgulamışlardır. Dolayısıyla, bu tür teknolojiler, yapay zekâ tabanlı dezenformasyon tespit sistemlerinin etkisini artırmak için kullanılabilir.

3.3. Yapay Zekânın Dezenformasyon Tespitindeki Başarı Örnekleri

Yapay zekâ, dezenformasyonun tespitinde ve yayılmasının önlenmesinde önemli başarılar elde ederek, bilgi ekosisteminin güvenilirliğini artırmada kilit bir rol oynamaktadır. Gelişmiş algoritmalar ve modeller, sosyal medya platformlarında hızla yayılan yanlış bilgileri tespit ederek, kriz dönemlerinde kamuoyunun doğru bilgilendirilmesini sağlamaktadır. Özellikle, seçim süreçleri ve sağlık krizleri gibi hassas dönemlerde, yapay zekâ tabanlı sistemler doğru bilgiyi öne çıkararak yanıltıcı içeriklerin etkisini minimize etmektedir. Bu başarılar, teknolojinin toplumsal bilgilendirme süreçlerinde etkinliğini net bir şekilde ortaya koymaktadır.

Justina Mandravickaite ve arkadaşları (2023), dezenformasyonu izlemek ve analiz etmek için geliştirdikleri gösterge panosunda, RoBERTa modeliyle sahte haberlerin sınıflandırılmasını sağlamışlardır. Bu sistem, görsel analiz araçlarıyla birleştirilen çeşitli analitik araçlarla dezenformasyonun daha etkili bir şekilde izlenmesine olanak tanımaktadır. Bu çalışma, yapay zekânın büyük veri setlerini işleyerek dezenformatif içerikleri hızlı ve doğru bir şekilde tespit etme kapasitesini göstermiştir. Bandi Sravani Reddy ve A.P. Siva Kumar (2023) ise, sahte haberlerin tespitinde çok modelli bir yaklaşım tekniği kullanmışlardır. Araştırmacılar sahte haberlerin tanımlanmasında çok alanlı görsel veri bilgilerinden yararlanmak için makine öğrenimini, görseldeki metnin tanımlaması için de OCR (Optical Character Recognition) yöntemlerini bir arada kullanmış ve

%97 doğruluk oranıyla etkili sonuçlar elde etmişlerdir. Bu yaklaşım, metinsel ve görsel içeriklerin birlikte analiz edilmesinin dezenformasyon tespitini nasıl geliştirdiğini göstermesi bakımından önemlidir.

Himanshi Rathore'nin (2021), COVID-19 pandemisi sırasında halk arasında yayılan yanlış bilgilerin tespitini hedefleyerek geliştirdiği yapay zekâ destekli sahte haber tespit programı da bu alandaki başarılı çalışmalara örnek olarak gösterilebilir. Program, kullanıcıların doğru bilgiyi yanlış olandan ayırt etmesini sağlayarak halk arasında paniği azaltmayı ve yanlış bilginin yayılmasını önlemeyi amaçlamıştır. Çalışma, özellikle sağlık krizleri gibi acil durumlarda yapay zekâ tabanlı araçların ne kadar etkili olabileceğini göstermektedir.

Raquiba Sultana ve Tetsuro Nishino (2022) ise, dönüştürücü tabanlı ve topluluk modeliyle sosyal medyada dezenformasyonu %97 doğruluk oranıyla tespit eden bir sistem geliştirmiştir. Model, yüksek F1 skoru ile doğru bilgi tanımlama sürecinde önemli bir başarı sağlamıştır. Bu çalışma, yapay zekâ modellerinin, özellikle sosyal medya platformlarında yayılan yanlış bilgilere karşı nasıl etkili bir çözüm sunduğunu gözler önüne sermektedir.

4. Yapay Zekânın Dezenformasyonla Mücadeledeki Sınırlılıkları

Yapay zekâ, dezenformasyonun tespitinde önemli bir araç olarak kullanılmasına rağmen, bu teknolojinin uygulanmasında ciddi etik ve operasyonel sınırlılıklar bulunmaktadır. Etik ve tarafsızlık sorunları, yapay zekâ algoritmalarının eğitildiği veri setlerindeki önyargılardan kaynaklanır ve bu, belirli grupların sistematik olarak yanlış değerlendirilmesine yol açabilir. Örneğin, sosyal medyada yayılan içeriklerin otomatik olarak sınıflandırılması sırasında, hassas gruplar veya konular yanlışlıkla dezenformatif olarak etiketlenebilir, bu da toplumsal eşitlik ve ifade özgürlüğü üzerinde olumsuz etkiler yaratabilir.

Yanlış pozitifler (*false positives*) ve içerik denetimi (*content moderation*) konuları da yapay zekâ tabanlı sistemlerin karşılaştığı diğer önemli sınırlılıklardır. Doğru bilgilerin yanlış bir şekilde dezenformatif olarak işaretlenmesi, bilginin serbest dolaşımını kısıtlayabilir ve kullanıcıların güvenini sarsabilir. Bu tür hatalar, özellikle politik veya sosyal olarak hassas içeriklerde, sansür ve bilgiye erişim konusunda ciddi tartışmalara yol açabilir. Bunun yanı sıra, yapay zekâ algoritmalarının karar alma süreçlerindeki şeffaflık eksikliği, kullanıcıların bu sistemlere olan güvenini azaltabilir. Şeffaflık ve hesap verebilirlik eksikliği, toplumsal düzeyde bilgi güvenliğine yönelik kaygıları artırır ve bu teknolojilere duyulan güvensizliği pekiştirir. Dolayısıyla, yapay zekâ sistemlerinin daha adil, şeffaf ve hesap verebilir bir şekilde geliştirilmesi, dezenformasyonla mücadelede etik sorunların üstesinden gelmek için hayati önem taşımaktadır.

4.1. Etik ve Tarafsızlık Sorunları

Yapay zekâ, dezenformasyonla mücadelede güçlü bir araç olmasına rağmen, etik ve tarafsızlık konularında önemli zorluklarla karşı karşıyadır. Nimesh Gupta (2023), yapay zekâ sistemlerinin adalet ve tarafsızlıkla ilgili sorunlarını ele alır ve bu sistemlerin genellikle toplumsal önyargıları yansıttığını belirtir. Gupta, bu önyargıların özellikle veri setlerinin dengesizliği ve algoritmaların geliştiricilerinin bilinçsiz veya bilinçli taraflılıkları nedeniyle ortaya çıktığını vurgulamakta ve yapay zekâ uygulamalarında adil sonuçlar elde etmek için şeffaflık ve hesap verebilirliğin artırılması gerektiğini savunmaktadır. Ayrıca, Gupta, adil algoritmalar geliştirmenin, kullanıcı güvenini artırmanın ve etik uygulamaları teşvik etmenin yollarını detaylandırmaktadır.

Mengyi Wei ve Zhixuan Zhou (2022) ise, yapay zekâ sistemlerinin etik sorunlarını daha geniş bir çerçevede ele alır ve bu teknolojilerin gerçek dünyadaki uygulamalarında ortaya çıkan problemlere dikkat çeker. Adı geçen araştırmacılar özellikle, ırk ayrımcılığı ve toplumsal eşitsizlikleri artırabilecek algoritmaların, kullanıcıların güvenini sarsabileceğini ve toplumsal huzursuzluğa neden olabileceğini tartışırlar. Wei ve Zhou, etik rehberlerin sınırlılıklarını vurgulayarak, daha etkili bir yapay zekâ etiği uygulaması için pratik bir taksonomi önerirler. Buna göre bu taksonomi, etik sorunların sınıflandırılmasına ve bunlara yönelik çözümlerin geliştirilmesine katkıda bulunabilir.

Bernd Stahl (2021) da SHERPA projesinden elde edilen bulgularla yapay zekânın gelişimi ve kullanımındaki etik endişeleri detaylandırmaktadır. Stahl, yapay zekâ teknolojilerinin toplumsal etkilerini analiz ederken, bu sistemlerin şeffaflık eksikliğinin ve hesap verebilirlik sorunlarının, kullanıcıların güvenini nasıl etkilediğini tartışmaktadır. Ayrıca, yapay zekânın etik açıdan daha sorumlu bir şekilde geliştirilmesi için öneriler sunan Stahl, bu süreçlerin yalnızca teknik düzeyde değil, aynı zamanda toplumsal ve politik düzeyde de ele alınması gerektiğini savunmaktadır.

Dolayısıyla yapay zekâ, dezenformasyonla mücadelede güçlü bir araç olmakla birlikte, etik ve tarafsızlık sorunları, bu teknolojinin güvenilirliği açısından ciddi bir endişe yaratmaktadır. Yapay zekâ sistemlerinin toplumsal önyargıları yansıtma eğilimi, özellikle veri setlerinin dengesizliği ve algoritma geliştiricilerinin farkında olmadan ya da bilinçli bir şekilde bu

önyargılara katkı sunmasıyla daha da belirgin hale gelmektedir. Bu durum, yapay zekânın yalnızca teknik bir araç olarak değil, aynı zamanda etik ilkelere uyum sağlaması gereken bir yapı olarak ele alınmasını zorunlu kılmaktadır.

4.2. Yanlış Pozitifler (false positives) ve İçerik Denetimi (content moderation) Sorunları

Yapay zekâ tabanlı dezenformasyon tespit sistemlerinde karşılaşılan en yaygın sınırlılıklardan biri olan yanlış pozitifler (false positives), doğru içeriklerin yanlışlıkla dezenformatif olarak sınıflandırılması durumudur. Yanlış pozitiflerin yaygınlaşması, bilgiye erişim özgürlüğünü kısıtlayabilir ve kullanıcıların bu sistemlere olan güvenini zedeleyebilir. Divya Tiwari ve Surbhi Thorat (2021), mevcut sahte haber tespit modellerinde yanlış alarm oranlarının yüksekliğine dikkat çekerek, sosyal bağlam ve içerik analizine dayalı yeni sınıflandırma yaklaşımlarının yanlış pozitifleri azaltmada daha etkili olduğunu belirtmektedir.

Yanlış pozitiflerin kullanıcı algısı ve bireysel inançlardan nasıl etkilendiğine dair detaylı bir analiz sunan Marco Hameleers (2023), yanlış pozitiflerin, içeriğin yanlı ve manipülatif algılanmasından kaynaklanabileceğini vurgulamaktadır. Bu bağlamda Hameleers, yanlış pozitiflerin yalnızca algoritmik sınırlılıklarla değil, aynı zamanda bireysel önyargılar ve algısal yanılgılarla da şekillendiğini söylemektedir.

Dezenformasyon stratejileri arasında 'falsification' (gerçeği çarpıtma veya sahte bilgi üretme) ve 'misleading' (yanıltıcı yönlendirme veya manipülasyon) gibi yanıltma stratejilerinin rolünü değerlendiren Svitlana Volkova ve Jae-wook Jang'a (2018) göre, yanlış pozitifler özellikle manipülatif yanıltma (misleading) stratejilerinde daha sık meydana gelmektedir. Volkova ve Jang, manipülatif içeriklerin tespitinde ortaya çıkan yanlış pozitif oranlarının, kullanıcı algısına göre değişen stratejik içeriklerle baş etmede daha yüksek olduğunu ifade etmektedir.

Buna ek olarak, Niall Conroy, Victoria L. Rubin ve Yimin Chen (2015)'in çalışmalarında ise, yanlış pozitiflerin içerik çeşitliliği ve karmaşıklığından kaynaklanan teknik zorlukları artırdığı belirtilmektedir. Sosyal medya ve çevrimiçi platformlardaki sahte haberlerin otomatik tespiti üzerine odaklanan çalışma, sosyal medya platformlarındaki içerik türlerinin çok yönlülüğünün yanlış pozitif oranlarını artırabileceğini ve algoritmaların bu karmaşıklık karşısında sınırlılıklarla karşılaştığını ortaya koymaktadır.

Yanlış pozitifler, yalnızca algoritmik sınırlamalardan değil, aynı zamanda kullanıcı algıları ve dezenformasyon stratejilerinin çeşitliliği gibi daha karmaşık etkenlerden de etkilenmektedir. Bireylerin kişisel inançları ve dünya görüşleri, içeriklerin yanlış sınıflandırılmasına zemin hazırlayabilirken, dezenformasyon stratejilerinin 'falsification' (gerçeği çarpıtma) ve 'misleading' (yanıltıcı yönlendirme) gibi karmaşık biçimleri, bu hataların ortaya çıkmasını daha da kolaylaştırmaktadır. Özellikle, manipülatif içerikler ve sosyal bağlamdan bağımsız değerlendirilen bilgi parçaları, algoritmaların yanılma payını artırmaktadır.

Bu doğrultuda, yanlış pozitif oranlarını minimize etmek amacıyla, yapay zekâ tabanlı sistemlerin yalnızca veri odaklı değil, aynı zamanda sosyal bağlamı anlamlandırmaya yönelik çok yönlü yaklaşımlar geliştirmesi gereklidir. Kullanıcıların bilgiye serbest erişimini sınırlamadan, doğru ve yanlış bilgiyi daha hassas bir şekilde ayırt edebilen algoritmaların geliştirilmesi hem toplumsal güveni artıracak hem de dezenformasyonla mücadelede daha etkin sonuçlar elde edilmesini sağlayacaktır.

4.3. Şeffaflık, Hesap Verebilirlik ve Toplumsal Etkiler

Yapay zekâ sistemlerinin toplumsal kabul görmesi ve kullanıcı güvenini sağlaması, şeffaflık ve hesap verebilirlik ilkeleriyle yakından ilişkilidir. Örneğin Ida Varošanec (2022), Avrupa Birliği yapay zekâ düzenlemeleri kapsamında şeffaflık yükümlülüklerinin net bir şekilde belirlenmesinin, bireylerin haklarını koruma ve şirketlerin suistimal riskini azaltma açısından kritik olduğunu vurgulamaktadır. Bu şeffaflık ilkeleri, kullanıcıların yapay zekâ sistemlerinin karar alma süreçlerini anlamalarına olanak tanır ve demokratik hakların korunmasına katkıda bulunur.

Martha C. Correa, Moreno ve Gina L. González Castro (2023), şeffaflık ve hesap verebilirlik eksikliklerinin, toplumdaki güven duygusunu zayıflattığını ve bunun yapay zekâ sistemlerinin toplumsal etkilerini olumsuz etkileyebileceğini belirtmektedir. Buna göre, yapay zekâ tabanlı sistemlerin şeffaflık politikaları ve hesap verebilirlik standartları geliştirilmediği sürece, özellikle kırılgan gruplar yanlış bilgilendirilmekte ve dolayısıyla dezenformasyon riskine daha fazla maruz kalmaktadırlar. Correa ve González, toplumun bilgiye erişim ve sistem işleyişi hakkında daha fazla bilgiye sahip olmasının, güven ilişkisini güçlendireceğini de savunmaktadırlar.

Kashyap Haresamudram, Stefan Larsson ve Fredrik Heintz (2023) ise, yapay zekâ şeffaflığının algoritmik, etkileşimsel ve toplumsal olmak üzere üç düzeyde ele alınması gerektiğini öne sürmüşlerdir. Araştırmacılar bu üçlü modelin, şeffaflığın hem sistem içi işleyişi hem de kullanıcı etkileşimlerini iyileştirebileceğini ve toplumsal etkilerin daha kapsamlı anlaşılmasını

sağlayabileceğini vurgulamışlardır. Çalışma, şeffaflık kavramının parçalı bir şekilde ele alınmasının sosyal etkileri zayıflattığını, bütüncül bir yaklaşımın ise güvenilirliği artıracağını ortaya koymaktadır.

Bu çalışmalar, yapay zekâ tabanlı sistemlerin şeffaflık ve hesap verebilirlik eksikliklerinin, kullanıcı güvenini zayıflattığını ve toplumsal etkilerini sınırlandırdığını göstermektedir. Özellikle dezenformasyonla mücadelede, yapay zekâ algoritmalarının işleyişini ve karar alma süreçlerini açıklığa kavuşturmak, bu sistemlerin etkinliğini artırmanın yanı sıra toplumda güven tesis etmenin de anahtarı olarak görülmektedir.

Sonuç

Bu çalışmada, yapay zekâ tabanlı teknolojilerin sosyal medya platformlarında dezenformasyonun yayılmasını önleme konusundaki potansiyeli ele alınmıştır. Sosyal medya, bilgiye hızlı erişim ve kolay paylaşım imkânı sunan bir mecra olarak toplumsal iletişimde önemli bir yer edinmiştir. Ancak, aynı zamanda manipülatif ve yanıltıcı bilgilerin hızla yayılmasına da zemin hazırlamaktadır. Bu sorunla mücadelede yapay zekâ tabanlı sistemlerin sunduğu çözümler, karmaşık bilgi ortamında umut verici sonuçlar ortaya koymaktadır. Özellikle makine öğrenimi ve doğal dil işleme gibi yapay zekâ teknikleri, dezenformatif içeriklerin tespiti için etkili araçlar olarak öne çıkmaktadır. Bu teknolojiler, dilin yapısal ve retorik özelliklerini analiz ederek manipülatif içerikleri ayırt etme kapasitesine sahiptir. Çeşitli çalışmalar, yapay zekâ modellerinin doğru sınıflandırma yapma becerisiyle dezenformasyonu tespit etmede kayda değer başarılar elde ettiğini göstermektedir.

Bu çalışmada referans alınan başarılı uygulama örnekleri, yapay zekâ algoritmalarının bilgi kirliliğiyle mücadelede umut verici sonuçlar verdiğini göstermektedir. Ancak bu sistemlerin sosyal medyadaki tüm dezenformasyon çeşitliliğini kapsayabilmesi için sürekli olarak geliştirilmesi ve bağlam analizine dayalı daha hassas algoritmaların tasarlanması gerekmektedir. Bununla birlikte sözü edilen teknoloji ve programların uygulanabilmesi için çözülmesi gereken çok sayıda sınırlılık ve etik sorun da bulunmaktadır.

Çalışmada da detaylandırıldığı gibi; yapay zekâ algoritmalarının dezenformasyon tespitinde önyargılardan etkilenebileceğine ve bu önyargıların doğru içeriklerin yanlış bir şekilde dezenformatif olarak etiketlenmesine yol açabileceğine yönelik kaygılar bulunmaktadır. Bu yanlış pozitiflerin yüksek olması, kullanıcıların doğru bilgiye erişim özgürlüğünü tehdit etmekte ve toplumsal düzeyde güvensizliğe neden olmaktadır. Özellikle sosyal medya gibi kullanıcı çeşitliliğinin yüksek olduğu ortamlarda, yanlış pozitifler ciddi olumsuz sonuçlara yol açabilmektedir. Örneğin, doğru içeriklerin engellenmesi veya yanlışlıkla dezenformatif olarak sınıflandırılması, kullanıcılar arasında bilgiye güvenin azalmasına yol açabilmektedir.

Etik sorunlar da yapay zekâ sistemlerinin tarafsızlık ve adalet ilkeleri çerçevesinde uygulanabilmesinde en önemli engellerden biridir. Algoritmaların hangi kriterlere göre dezenformasyon tespitinde bulunduğu ve bu sürecin ne ölçüde tarafsız olduğu, toplumsal güven için kritik bir öneme sahiptir. Yapay zekâ teknolojilerinin hesap verebilir olmaması ve şeffaflık eksikliği, algoritmaların işleyişi hakkında kullanıcıların bilgi sahibi olamamasına neden olmaktadır. Özellikle, yapay zekâ algoritmalarının şeffaf olmaması, bu sistemlerin nasıl karar verdiği konusunda belirsizlikler yaratmakta ve dezenformasyonla mücadelede bu teknolojilere olan güveni sarsmaktadır.

Bu çalışmada ayrıca, sosyal medya platformlarının ve yapay zekâ sistemlerinin toplumsal etkileri de ele almış, dezenformasyonla mücadelede yapay zekâ tabanlı sistemlerin daha şeffaf, hesap verebilir ve adil bir yapıya sahip olması gerektiği sonucuna varılmıştır. Kullanıcılar, bu sistemlerin nasıl çalıştığını ve içeriklerin hangi nedenlerle dezenformatif olarak etiketlendiğini anlamalıdır. Bu doğrultuda, daha etkili hesap verebilirlik standartları ve kullanıcı bilgilendirme mekanizmalarının geliştirilmesi, toplumsal güvenin artırılmasına katkıda bulunacaktır.

Gelecekte yapılacak araştırmalar, yapay zekâ sistemlerinin sosyal bağlamı ve kullanıcı algısını daha hassas bir şekilde değerlendirebilecek gelişmiş algoritmalar üzerine odaklanmalıdır. Çok modelli analiz yöntemleri ve bağlamsal içerik analizine dayalı teknolojiler, yanlış pozitiflerin oranını düşürmek ve doğru içeriklerin yanlış bir şekilde engellenmesini önlemek için kritik öneme sahiptir. Bunun yanı sıra, yapay zekâ algoritmalarının şeffaflığını artırmaya yönelik standartların belirlenmesi, kullanıcıların bu sistemlere güvenini yeniden inşa etmede önemli bir adım olacaktır. Özellikle, platformların kullanıcı bilgilendirme politikalarını güçlendirmesi ve yapay zekâ tabanlı dezenformasyon tespit sistemlerini daha şeffaf hale getirmesi, dezenformasyonun etkilerini azaltmak ve bilgi ekosistemini daha güvenilir kılmak için hayati bir öneme sahiptir.

Sonuç olarak, yapay zekâ tabanlı dezenformasyon tespit sistemlerinin gelişimi, yalnızca teknolojik performans açısından değil, aynı zamanda toplumsal değerler ve etik standartlar doğrultusunda şekillendirilmelidir. Bu tür teknolojilerin toplum nezdinde kabul görmesi için şeffaflık, hesap verebilirlik ve etik değerler merkezinde ilerlenmeli; doğru bilgiye erişimi engellemeyen, adil ve kullanıcı dostu yapılar geliştirilmelidir.

Kaynakça | References

Akyüz, S. S. (2021). Koronavirüs komplo teorileri: dezenformasyon ve politik kimliklerin komplocu düşünüşe etkisi. İletişim ve Medya Alanında Uluslararası Araştırmalar II, 57, 86.

Allcott, H., & Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. Journal of Economic Perspectives, 31(2), 211-236. https://doi.org/10.1257/jep.31.2.211

Awais, I., Rahim, N. A., Alhossary, A. Z., & Rahman, Z. A. (2022). Israeli Arabic-speaking Facebook pages and its effects on the elements of Palestinian national identity. International Journal of Humanities Studies, 6(4), 11337. https://doi.org/10.53730/ijhs.v6ns4.11337

Bektaş, A. (2002). Siyasal Propaganda: Tarihsel Evrimi ve Demokratik Toplumdaki Uygulamaları, Bağlam Yayınları, İstanbul.

Brennen, J. S., Simon, F., Howard, P. N., & Nielsen, R. K. (2020). Types, sources, and claims of COVID-19 misinformation. Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk

Chen, S., Xiao, L., & Kumar, A. (2022). Spread of misinformation on social media: What contributes to it and how to combat it. Computers in Human Behavior, 107, 643. https://doi.org/10.1016/j.chb.2022.107643

Choraś, M., Demestichas, K., Giełczyk, A., Herrero, Á., Ksieniewicz, P., Remoundou, K., Urda, D., & Woźniak, M. (2020). Advanced Machine Learning Techniques for Fake News (Online Disinformation) Detection: A Systematic Mapping Study. Applied Soft Computing, 107, 643. https://doi.org/10.1016/j.asoc.2020.107050

Çimen, Ü., & Yüksel, H. (2018). Medya sektörü bağlamında iş zekâsı kavramı ve önemi. Tarih Okulu Dergisi (TOD), 11(37), 55-69. https://doi.org/10.14225/Joh1357

Conroy, N., Rubin, V. L., & Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. Proceedings of the Association for Information Science and Technology, 52(1), 1-4. https://doi.org/10.1002/pra2.2015.145052010082

Correa Moreno, M. C., & González Castro, G. L. (2023). Unveiling public information in the metaverse and AI era: Challenges and opportunities. Media Research Journal, 35, 187-203. https://doi.org/10.56294/mr202335

Çömlekçi, M. F. (2019). Sosyal Medyada Dezenformasyon ve Haber Doğrulama Platformlarının Pratikleri. Gümüşhane Üniversitesi İletişim Fakültesi Elektronik Dergisi, 7, 1549-1563.

Delgado, A., Glisson, W., Shashidhar, N., McDonald, J., Grispos, G., & Benton, R. (2021). Deception Detection Using Machine Learning. Proceedings of the Hawaii International Conference on System Sciences. https://doi.org/10.24251/HICSS.2021.857

Eroğlu, E. (2023). Seçim Dönemlerinde Sosyal Medya Dezenformasyonu: 2023 Genel Seçimleri Üzerine Bir İçerik Analizi. Elektronik Cumhuriyet İletişim Dergisi, 5(2), 142-151.

Erkan, G., & Ayhan, A. (2018). Siyasal iletişimde dezenformasyon ve sosyal medya: Bir doğrulama platformu olarak teyit. org. Akdeniz Üniversitesi İletişim Fakültesi Dergisi, (29. Özel Sayısı), 202-223.

Ertürk, H. A. (2022). YENİ MEDYA EKSENİNDE İDEOLOJİYİ ANLAMAK: FİLTRE BALONLARI VE YANKI ODALARI. Niğde Ömer Halisdemir Üniversitesi İletişim Fakültesi Akademik Dergisi, 1(2), 137-159.

Galitsky, B. A. (2015). Detecting Rumor and Disinformation by Web Mining. DBLP.

Gupta, N. (2023). Artificial Intelligence Ethics and Fairness: A study to address bias and fairness issues in AI systems, and the ethical implications of AI applications. Research Review International Journal of Multidisciplinary. https://doi.org/10.31305/rrijm2023.v03.n02.004

Hameleers, M. (2023). This is clearly fake! Mis- and disinformation beliefs and the (accurate) recognition of pseudo-information—Evidence from the United States and the Netherlands. American Behavioral Scientist, 77(5), 583-598. https://doi.org/10.1177/00027642231174334

Haresamudram, K., Larsson, S., & Heintz, F. (2023). Three levels of AI transparency. IEEE Computer, 56(3), 46-53. https://doi.org/10.1109/MC.2022.3213181

Iddianto, R. Azi. (2022). SOCIAL EFFECT OF SOCIAL MEDIA REVEALED IN THE SOCIAL DILEMMA DOCUMENTARY MOVIE: POST-TRUTH PERSPECTIVE. Seshiski Journal, 2(1), 3. https://doi.org/10.53922/seshiski.v2i1.3

Litvinova, O., Seredin, P., Litvinova, T., & Lyell, J. (2017). Deception detection in Russian texts. Proceedings of the Conference on Empirical Methods in Natural Language Processing. https://doi.org/10.18653/V1/E17-4005

Mandravickaite, J., Songailaitė, M., Kankevičiūtė, E., Volčok, A., & Krilavičius, T. (2023). Disinformation Analysis and Tracking Dashboard. Proceedings of the IEEE International Conference on Cyber Warfare and Security. https://doi.org/10.1109/ICMCIS59922.2023.10253590

Merter, A. K., & Özer, G. (2023). Denetimde yapay zeka: S. Z. İmamoğlu, S. Erat, & H. İnce (Editör.), Yönetim Biliminde Yapay Zekâ (Sayfa. 257-274). Nobel Bilimsel Eserler.

Pariser, E. (2011). The filter bubble: What the internet is hiding from you. Penguin Press.

Rasyid, H., Sibaroni, Y., & Ihsan, A. F. (2023). Classification of Disinformation Tweet on the 2024 Presidential Election in Indonesia Using Optimal Transformer Based Model. Proceedings of the IEEE International Conference on Cyber Defense and Secure Communications. https://doi.org/10.1109/ICoDSA58501.2023.1027710

Rathore, H. (2021). Detecting fake Covid-19 news. International Journal for Research in Applied Science & Engineering Technology. https://doi.org/10.22214/ijraset.2021.35271

Reddy, B. S., & Kumar, A. S. (2023). Multimodal approaches based on fake news detection. IEEE. https://doi.org/10.1109/ICAIS56108.2023.10073839

Sharevski, F., Devine, A., Pieroni, E., & Jachim, P. (2022). Folk models of misinformation on social media. arXiv. https://doi.org/10.48550/arXiv.2207.12589

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22-36. https://doi.org/10.1145/3137597.3137600

Siapera, E. (2014). Tweeting #Palestine: Twitter and the mediation of Palestine. Journal of Communication, 136, 787-791. https://doi.org/10.1177/1367877913503865

Stahl, B. (2021). Ethical issues of AI. In Ethics of Artificial Intelligence (pp. 77-95). Springer. https://doi.org/10.1007/978-3-030-69978-9_4

Sultana, R., & Nishino, T. (2022). Fake News Detection Using Transformer and Ensemble Learning Models. Proceedings of the IEEE Winter Conference on Applications of Computer Vision. https://doi.org/10.1109/IIAI-AAI-Winter58034.2022.00044

Sunstein, C. R. (2001). Republic.com. Princeton University Press, Princeton, NJ.

Tandoc, E. C., Lim, Z. W., & Ling, R. (2018). Defining “fake news”: A typology of scholarly definitions. Digital Journalism, 6(2), 137-153. https://doi.org/10.1080/21670811.2017.1360143

Teyit.org. (2023). Sahte Haber Karnesi: 2023 Cumhurbaşkanı ve Milletvekili Seçimleri.

Tiwari, D., & Thorat, S. (2021). An Analysis on False Positives in Fake News Detection Algorithms: Improving Content Classification with Contextual Approaches. International Journal of Computer Science and Information Technologies, 7(6). https://doi.org/10.32628/cseit217670

Tsikerdekis, M., & Zeadally, S. (2023). Misinformation Detection Using Deep Learning. IEEE Computer. https://doi.org/10.1109/MITP.2023.3314752

Varošanec, I. (2022). Transparency Obligations in AI: A European Perspective on AI Regulation. International Journal of Law and Information Technology, 30(2), 92-110. https://doi.org/10.1080/13600869.2022.2060471

Volkova, S., & Jang, J. (2018). Misleading and Falsification Strategies in Social Media: Evaluating False Positives in Detection Models. Proceedings of the 27th International Conference on World Wide Web Companion. https://doi.org/10.1145/3184558.3188728

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146-1151. https://doi.org/10.1126/science.aap9559

Ward, K., & Goodwin, J. (2022). Identifying disinformation using rhetorical devices in natural language models. arXiv. https://doi.org/10.2172/1891194

Wardle, C., & Derakhshan, H. (2017). Information Disorder: Toward an Interdisciplinary Framework for Research and Policymaking. Council of Europe.

Wei, M., & Zhou, Z. (2022). AI Ethics Issues in Real World: Evidence from AI Incident Database. arXiv. https://doi.org/10.48550/arXiv.2206.07635