Türkiye Teknoloji ve Uygulamalı Bilimler Dergisi

Turkish Journal of Applied Sciences and Technology

# BERT Tabanlı Sahte Haber Tespiti: Veri Sızıntısını Önleyen Yenilikçi Bir Yaklaşım ve Metodolojik Katkılar

## BERT-Based Fake News Detection: An Innovative Approach for Preventing Data Leakage and Methodological Contributions

Muhammed Yusuf KAYA[1], Bashar ALHAJAHMAD[2]

ISSN: 2528-8652
e-ISSN: 2822-2660

**Araştırma Makalesi**

**Doi:**10.70562/tubid.1870238

**Sorumlu yazar**
Muhammed Yusuf KAYA
myusuf.kaya@siirt.edu.tr

**Geliş tarihi:** 23.01.2026
**Kabul tarihi:** 30.03.2026
**Yayınlanma tarihi:** 17.07.2026

**Alıntı**
Kaya, M.Y. ve Alhajahmad, B. (2026). BERT Tabanlı Sahte Haber Tespiti: Veri Sızıntısını Önleyen Yenilikçi Bir Yaklaşım ve Metodolojik Katkılar. *Türkiye Teknoloji ve Uygulamalı Bilimler Dergisi*, 81-91.

**Öz:** Dijital çağda sahte haberlerin hızla yayılması, toplumsal güven ve demokratik süreçler açısından kritik bir tehdit unsuru oluşturmaya başlamıştır. Sunulan araştırma, BERT (Bidirectional Encoder Representations from Transformers) tabanlı yenilikçi bir sahte haber tespit yaklaşımı ortaya koymakta olup, veri sızıntısı (data leakage) problemini sistematik biçimde ele almayı hedeflemektedir. Araştırmanın temel amacı, literatürde sıklıkla karşılaşılan metodolojik hataları gidererek güvenilir bir performans değerlendirmesi gerçekleştirmektir. Materyal olarak herkesin erişimine açık Kaggle platformundan temin edilen ISOT Sahte Haber Veri Seti tercih edilmiştir. Yöntem olarak MD5 hash tabanlı tekrar tespiti, katmanlı örnekleme ile veri bölme ve BERT-base-uncased modeli üzerinde ince ayar (fine-tuning) stratejisi benimsenmiştir. Veri temizliği aşamasında 6.252 tekrarlı örnek (% 13.9) belirlenerek veri setinden çıkarılmış, eğitim-test seti örtüşmesi 1.950 örnekten 2'ye indirgenmiştir. Elde edilen bulgular, temizlenmiş veri seti üzerinde % 99.97 test doğruluğuna ulaşıldığını, sahte haber sınıfı için % 99.92 kesinlik ve % 100 duyarlılık değerlerinin sağlandığını ortaya koymaktadır. AUC-ROC skoru 0.9997 olarak hesaplanmıştır. Sonuç olarak, metodolojik titizlik gözetildiğinde BERT modellerinin sahte haber tespitinde üstün performans sergileyebildiği doğrulanmış; veri sızıntısı önleme stratejilerinin model değerlendirmesindeki belirleyici rolü açık biçimde gösterilmiştir.

**Anahtar Kelimeler:** Sahte haber tespiti, BERT, veri sızıntısı, transformer modelleri, derin öğrenme

**Abstract**
The rapid spread of fake news in the digital age has emerged as a critical threat to social trust and democratic processes. This research presents an innovative BERT-based (Bidirectional Encoder Representations from Transformers) fake news detection approach that systematically addresses the data leakage problem. The primary objective is to provide reliable performance evaluation by eliminating methodological errors commonly encountered in the literature. The ISOT Fake News Dataset, publicly available through the Kaggle platform, was utilized as material. MD5 hash-based duplicate detection, stratified sampling for data splitting, and fine-tuning strategy with BERT-base-uncased model were adopted as methodology. During data cleaning, 6,252 duplicate samples (13.9%) were identified and removed from the dataset, while train-test set overlap was reduced from 1,950 samples to 2. The findings demonstrate that 99.97% test accuracy was achieved on the cleaned dataset, with 99.92% precision and 100% recall obtained for the fake news class. The AUC-ROC score was calculated as 0.9997. In conclusion, it has been verified that BERT models can exhibit superior performance in fake news detection when methodological rigor is maintained; the decisive role of data leakage prevention strategies in model evaluation has been clearly demonstrated.

**Keywords:** Fake news detection, BERT, data leakage, transformer models, deep learning

[1*]Siirt Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı, Siirt, Türkiye
[2]Siirt Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Siirt, Türkiye

## 1. Giriş

Sosyal medya platformlarının günlük yaşamın ayrılmaz bir parçası haline gelmesiyle birlikte, bilgi tüketim alışkanlıkları köklü bir dönüşüm sürecine girmiştir. Facebook, Twitter ve WhatsApp gibi platformlar aracılığıyla haberler, geleneksel medya kanallarına kıyasla çok daha hızlı biçimde milyonlarca kullanıcıya ulaşabilmektedir (Lazer ve ark., 2018). Ancak bilgi paylaşımının tabana yayılması, beraberinde bilgi kirliliği sorunlarını da getirmiştir. Doğrulanmamış ya da kasıtlı olarak yanıltıcı üretilmiş içerikler, sosyal medya üzerinden viral şekilde yayılarak toplumsal düzeyde derin etkiler bırakabilmektedir.

Sahte haber kavramı, kasıtlı olarak yanıltıcı bilgi içeren ve haber formatında sunulan içerikleri tanımlamak amacıyla kullanılmaktadır. Shu ve ark. (2017), sahte haber tespitinde kullanılan veri madenciliği tekniklerini kapsamlı biçimde inceleyerek içerik tabanlı ile sosyal bağlam tabanlı yaklaşımları karşılaştırmış ve dilbilimsel özniteliklerin (n-gram, sözcük türü etiketleme, duygu skorları) sahte haberleri ayırt etmede etkili sonuçlar verdiğini ortaya koymuştur. Bu tür içeriklerin üretilmesinde siyasi manipülasyon, ekonomik çıkar elde etme ve toplumsal kargaşa yaratma gibi çeşitli motivasyonlar rol oynamaktadır. Allcott ve Gentzkow (2017), 2016 ABD başkanlık seçimleri döneminde sahte haber sitelerinden gelen içeriklerin Facebook üzerinde 30 milyonu aşkın paylaşım aldığını raporlamıştır. Söz konusu bulgu, sahte haberlerin demokratik süreçler üzerindeki potansiyel etkisini somut verilerle gözler önüne sermektedir.

Vosoughi ve ark. (2018), Science dergisinde yayımlanan kapsamlı araştırmalarında 2006-2017 yılları arasında Twitter platformunda yayılan 126.000 haber hikayesini incelemiştir. Analiz sonuçları, sahte haberlerin gerçek haberlere kıyasla altı kat daha hızlı yayıldığını ve % 70 oranında daha fazla paylaşım aldığını göstermiştir. Araştırmacılar, sahte haberlerin daha yüksek yenilik değeri taşıması ve daha güçlü duygusal tepkiler uyandırmasının bu yayılım hızını açıklayabileceğini ileri sürmüştür. Nitekim Horne ve Adali (2017) tarafından gerçekleştirilen dilbilimsel analiz de bu bulguları destekler niteliktedir; sahte haberlerin başlıklarının genellikle içerikle zayıf bir ilişki sergilediği ve tıklama tuzağı (clickbait) karakteristikleri taşıdığı tespit edilmiştir. Random Forest sınıflandırıcı kullanılarak % 87 F1-skoru elde eden bu çalışma, geleneksel makine öğrenmesi yaklaşımlarının temel kısıtlamalarını da ortaya koymuştur: kapsamlı öznitelik mühendisliği gerekliliği, bağlamsal anlama eksikliği ve alan transferinde düşük performans.

COVID-19 pandemisi sürecinde sahte haberlerin yayılımı, Dünya Sağlık Örgütü tarafından "infodemi" olarak nitelendirilen yeni bir fenomeni gündeme taşımıştır (Zarocostas, 2020). Pandemi döneminde yanlış sağlık bilgilerinin halk sağlığı müdahalelerini ciddi şekilde tehlikeye attığı belgelenmiştir. Zarocostas (2020), hatalı bilgilerin aşı kararsızlığını artırdığını ve tedavi protokollerine uyumu olumsuz yönde etkilediğini vurgulamıştır. Bu gelişmeler, sahte haber tespitinin salt akademik bir ilgi alanı olmaktan öte, acil bir toplumsal gereklilik niteliği kazandığını açıkça işaret etmektedir.

Geleneksel doğrulama mekanizmaları, günde milyonlarca içeriğin üretildiği dijital ekosistemde yetersiz kalmaya başlamıştır (Graves, 2018). İnsan tabanlı doğrulama sistemleri hem zaman hem de maliyet açısından ölçeklendirilebilir değildir. Doğrulama (fact-checking) kuruluşları, içeriklerin yalnızca küçük bir bölümünü doğrulayabilmekte ve bu süreç çoğunlukla viral yayılım gerçekleştikten sonra tamamlanmaktadır. Zhou ve Zafarani (2020), 200'den fazla

çalışmayı kapsayan literatür taramalarında otomatik tespit sistemlerinin bu boşluğu kapatmada kritik bir işlev üstlendiğini belirtmiş; veri setlerinin heterojenliği, değerlendirme ölçütlerinin tutarsızlığı ve tekrarlanabilirlik sorunlarını alandaki temel engeller olarak nitelendirmiştir.

Makine öğrenmesi ve derin öğrenme teknikleri, otomatik sahte haber tespiti için umut vadeden çözümler sunmaya başlamıştır. Wang (2017) tarafından tanıtılan LIAR veri seti, 12.836 kısa ifade içermekte ve PolitiFact doğrulama etiketleri ile eşleştirilmektedir; CNN-LSTM hibrit modeli ile elde edilen % 27.4 doğruluk oranı, çok sınıflı sınıflandırma probleminin karmaşıklığını yansıtmaktadır. (Yang ve ark., 2018), TI-CNN modeliyle metin ve görsel öznitelikleri birleştirerek % 95.4 doğruluk elde etmiş; bu çok kipli (multimodal) yaklaşım, haber içeriklerindeki metin-görsel tutarsızlıklarını yakalamak için etkili bir yöntem sunmuştur. Nasir ve ark. (2021) ise CNN-RNN hibrit mimarisi ile uzun metinlerdeki yerel ve küresel örüntüleri eş zamanlı olarak modelleyerek % 96 doğruluğa ulaşmıştır; ancak bu modeller uzun menzilli bağımlılıkları yakalamada ve paralel işlemede belirli kısıtlamalarla karşı karşıya kalmaktadır.

Özellikle transformer mimarisine dayalı modeller, doğal dil işleme görevlerinde paradigma değişimine yol açmıştır. Vaswani ve ark. (2017) tarafından 2017 yılında tanıtılan transformer mimarisi, dikkat mekanizması (attention mechanism) sayesinde dizideki tüm konumlar arasındaki ilişkileri eş zamanlı olarak modelleyebilmektedir. Bu mimari üzerine inşa edilen BERT modeli (Devlin ve ark., 2019), çift yönlü bağlam anlama kapasitesiyle metin sınıflandırma görevlerinde dikkat çekici başarılar yakalamıştır. (Kaliyar ve ark., 2021), FakeBERT modelini geliştirerek BERT'i farklı çekirdek boyutlarına sahip tek katmanlı CNN blokları ile birleştirmiş ve % 98.90 doğruluk elde etmiştir. Kula ve ark. (2020) ise ince ayarlı BERT modeli ile % 99.2 doğruluk bildirmiştir. Ancak bu çalışmalarda veri sızıntısı kontrolünün yeterli düzeyde gerçekleştirilmediği dikkat çekmektedir.

Öte yandan, literatürdeki mevcut çalışmalar incelendiğinde kritik bir metodolojik sorun göze çarpmaktadır: veri sızıntısı (data leakage). Kapoor ve Narayanan (2023), makine öğrenmesi tabanlı bilimsel çalışmalarda sızıntı ve tekrarlanabilirlik krizini sistematik olarak analiz etmiştir. Horne ve ark. (2019) ise sahte haber tespit modellerinin zaman içindeki dayanıklılığını ve saldırılara karşı direncini incelemiştir. Pek çok araştırma, veri setlerindeki tekrarlı örnekleri temizlemeden ya da eğitim-test bölmelerindeki örtüşmeleri kontrol etmeden % 99-100 aralığında doğruluk değerleri bildirmektedir. Ahmed ve ark. (2017) tarafından tanıtılan ISOT veri setinde yaklaşık 6.000-7.000 tekrarlı örnek bulunduğu tespit edilmiş olup, bu tekrarlar modelin eğitim setinde gördüğü örneklerin test setinde de yer almasına neden olmakta ve yapay olarak yüksek performans değerlerine yol açmaktadır. Bu durum, modellerin gerçek genelleme kapasitesini değerlendirmeyi güçleştirmekte ve yanıltıcı bir başarı algısı yaratmaktadır.

Sunulan araştırma, sahte haber tespiti çalışmalarında sıklıkla göz ardı edilen veri sızıntısı problemini sistematik biçimde ele almayı amaçlamaktadır. Bu doğrultuda çalışmanın temel katkıları aşağıdaki şekilde özetlenebilir:

• Sahte haber veri setlerinde ortaya çıkabilen veri sızıntısı problemini azaltmaya yönelik, MD5 tabanlı tekrar tespiti ve temizliği ile desteklenen yeniden üretilebilir bir veri temizleme protokolü önerilmiştir.

• Veri setindeki tekrarların model performansı üzerindeki etkisini ortaya koymak amacıyla, veri temizliği öncesi ve sonrası değerlendirme süreçleri analiz edilerek metodolojik açıdan daha güvenilir bir deneysel çerçeve sunulmuştur.

• BERT tabanlı transformer mimarisinin temizlenmiş veri seti üzerinde gerçek genelleme performansı sistematik olarak değerlendirilmiştir.

• Tüketici sınıfı donanımlarda uygulanabilir olacak şekilde, karma hassasiyetli eğitim ve gradyan biriktirme tekniklerini içeren verimli bir ince ayar (fine-tuning) stratejisi sunulmuştur.

• Çalışmanın tüm veri işleme, eğitim ve değerlendirme adımları ayrıntılı biçimde raporlanarak tekrarlanabilir bir deneysel metodoloji sağlanmıştır.

## 2. Materyal ve Yöntem

### 2.1. Veri seti

Araştırma kapsamında, Ahmed ve ark. (2017) tarafından oluşturulan ve herkesin erişimine açık Kaggle platformundan temin edilen ISOT Sahte Haber Veri Seti kullanılmıştır. Söz konusu veri seti, sahte haber tespiti araştırmalarında yaygın olarak tercih edilen bir kıyaslama ölçütü niteliği taşımaktadır. PolitiFact tarafından güvenilmez olarak işaretlenmiş kaynaklardan derlenen 23.481 sahte haber ve Reuters.com'dan alınan 21.417 gerçek haber olmak üzere toplam 44.898 makale içermektedir. İçerikler, 2015-2018 döneminde ABD siyasi haberlerini kapsamaktadır.

### 2.2. Veri temizliği ve ön işleme

Veri temizliği süreci üç aşamada gerçekleştirilmiştir. İlk aşamada tekrar tespiti ve kaldırma işlemi uygulanmıştır. MD5 kriptografik özet (hash) fonksiyonu (Rivest, 1992) kullanılarak tam metin eşleştirmesi yapılmış ve 6.252 tekrarlı makale (% 13.9) tespit edilmiştir. Her tekrar kümesinden yalnızca ilk örnek korunarak diğerleri veri setinden çıkarılmış, böylece 38.646 benzersiz makaleye ulaşılmıştır.

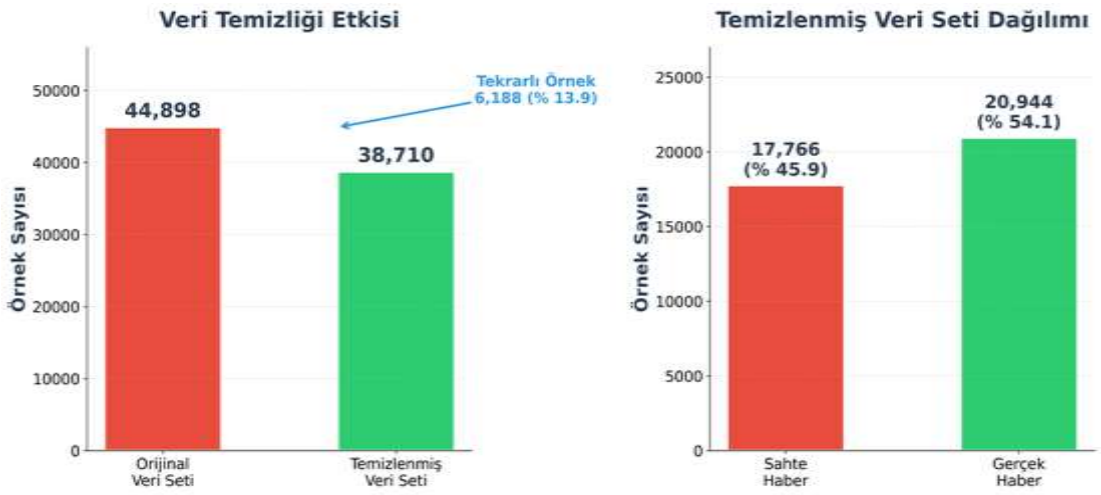
İkinci aşamada katmanlı (stratified) eğitim, doğrulama ve test bölme işlemi uygulanmıştır. Sınıf dengesi korunarak % 70 eğitim (27.052 örnek), % 15 doğrulama (5.797 örnek) ve % 15 test (5.797 örnek) oranlarında bölme gerçekleştirilmiştir. Veri seti, tekrar temizliği tamamlandıktan sonra, herhangi bir özellik çıkarımı veya ölçekleme uygulanmadan önce stratified yöntemle bölünmüştür. Bu yaklaşım, hem sınıf dağılımının korunmasını hem de veri sızıntısının önlenmesini sağlamıştır.

Üçüncü aşamada küme tabanlı kesişim (set-based intersection) analizi ile eğitim-test örtüşmesi kontrol edilmiştir. Başlangıçta saptanan 1.950 örnek örtüşme, 2'ye düşürülmüştür (% 99.9 azalma). Kalan 2 örnek, farklı bağlamlarda tekrarlanan genel ifadeler olduğundan model performansını etkilemeyeceği değerlendirilmiştir.

Bu veri setinde toplam örneklerin % 13.9'unu oluşturan tekrarlar dikkate alındığında, test kümesinin yaklaşık % 15'lik kısmında yer alabilecek yinelenmiş kayıtların, model tarafından ezber yoluyla doğru sınıflandırılabileceği riski mevcuttur. Bu durumda, tekrarlı veriler silinmeden gerçekleştirilen bir çalışma örneğin % 95 doğruluk değeriyle raporlanmışsa, aslında

en kötü senaryoda yaklaşık % 80–82 bandına kadar gerileyebilecek bir gerçek genelleme performansı olduğu göz önünde bulundurulmalıdır.

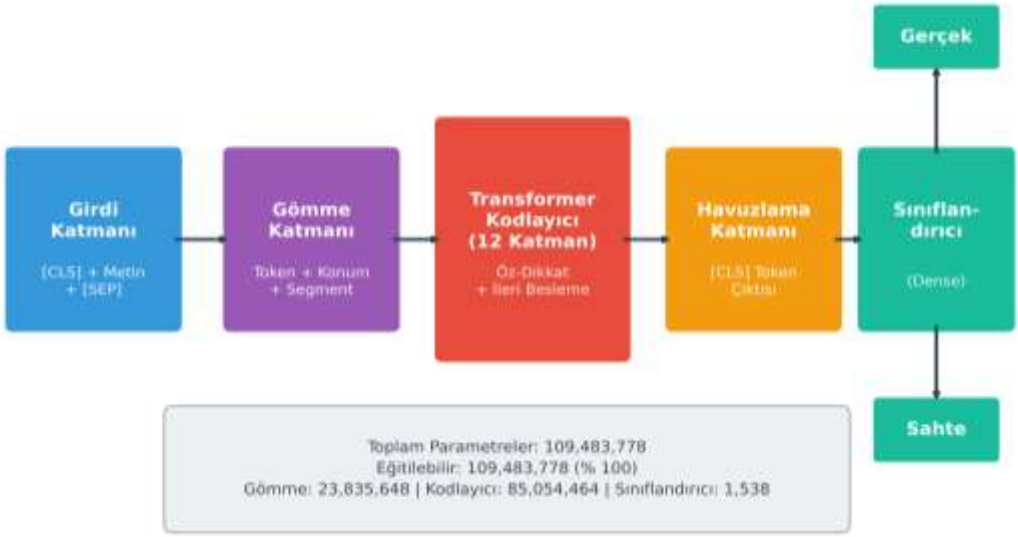
Tokenizasyon aşamasında yalnızca haber metni (text sütunu) girdi olarak kullanılmış; başlık bilgisi modele ayrıca verilmemiştir. 256 token sınırını aşan metinler sağdan kırpılmıştır (truncation). Veri temizliği sürecinin etkileri Şekil 1'de görselleştirilmiştir.




**Şekil 1.** Veri temizliği etkisi ve değişim oranları

## 2.3. BERT model mimarisi ve konfigürasyon

Araştırmada BERT-base-uncased modeli tercih edilmiştir (Devlin ve ark., 2019). Model, 12 adet Transformer tabanlı kodlayıcı katmandan oluşmaktadır. Her katman 768 boyutunda gizil vektör temsiline sahip olup, 12 adet çok başlıklı dikkat (multi-head attention) mekanizması içermektedir. Toplam parametre sayısı yaklaşık 110 milyon civarındadır. Model mimarisi Şekil 2'de şematik olarak gösterilmiştir.




**Şekil 2.** BERT model mimarisi

Maksimum girdi dizisi uzunluğu 256 token olarak belirlenmiş, metinler 30.522 kelimelik WordPiece sözlüğü kullanılarak tokenize edilmiştir. AdamW optimizasyon algoritması ile $2\times10^{-5}$ öğrenme oranı uygulanmıştır. Eğitim 3 dönem (epoch) sürdürülmüş ve doğrusal öğrenme oranı zamanlayıcısı kullanılmıştır.

### 2.4. Eğitim stratejisi ve optimizasyon

Kaynak kısıtlı ortamlarda verimli eğitim gerçekleştirmek amacıyla karma hassasiyetli eğitim (mixed precision training) ve gradyan biriktirme (gradient accumulation) teknikleri uygulanmıştır. FP16 (16-bit kayan nokta) hesaplama ile % 75 bellek tasarrufu sağlanmış, gradyan biriktirme ile efektif yığın boyutu (batch size) 32'ye yükseltilmiştir.

Eğitim, NVIDIA RTX 3060 (6GB VRAM) grafik işlemci üzerinde gerçekleştirilmiştir. Deneylerin tekrarlanabilirliğini sağlamak amacıyla tüm eğitim süreçlerinde rastgelelik kontrolü uygulanmış ve sabit bir random seed değeri (seed = 42) kullanılmıştır. Metinler, maksimum 256 token uzunluğuna kadar BERT tokenizer kullanılarak tokenize edilmiş ve eksik uzunluklar için standart padding uygulanmıştır. Eğitim sırasında gradyan patlamasını önlemek amacıyla maksimum norm değeri 1.0 olacak şekilde gradient clipping uygulanmıştır. Erken durdurma (early stopping) mekanizması doğrulama kaybına göre ayarlanmış ve sabır (patience) değeri 2 dönem olarak belirlenmiştir. Model kontrol noktaları, en iyi doğrulama performansına göre kaydedilmiştir. Eğitim sürecinde kullanılan hiperparametreler Tablo 1'de özetlenmiştir.

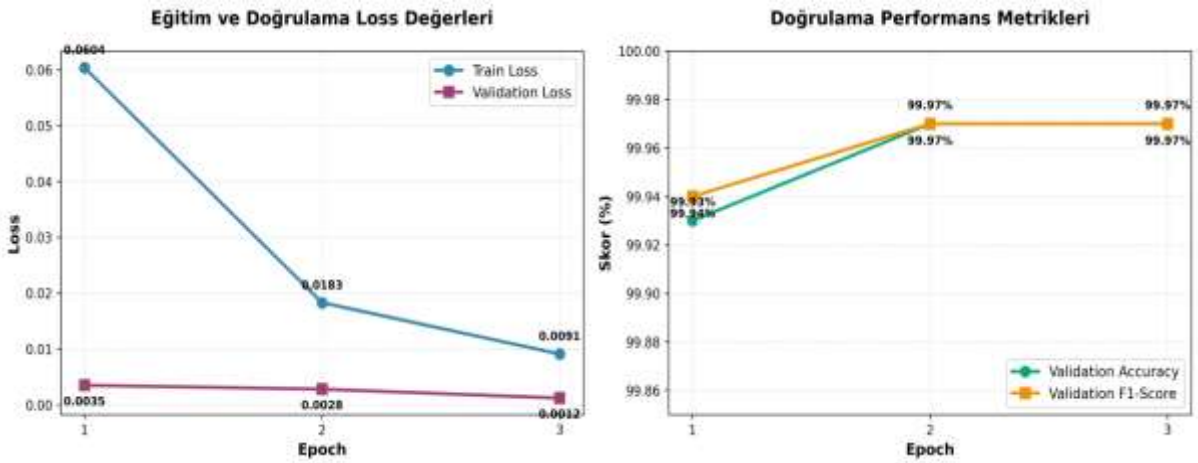
**Tablo 1.** Eğitim hiperparametreleri

| **Parametre** | **Değer** |
|---|---|
| Model | BERT-base-uncased |
| Maksimum dizi uzunluğu | 256 token |
| Eğitim dönemi (epoch) | 3 |
| Yığın boyutu (batch size) | 16 |
| Gradyan biriktirme adımı | 2 |
| Efektif yığın boyutu | 32 |
| Öğrenme oranı | $2\times10^{-5}$ |
| Optimizasyon algoritması | AdamW |
| Ağırlık azalması (weight decay) | 0.01 |
| Isınma adımı (warmup steps) | 500 |
| Karma hassasiyet (FP16) | Aktif |
| Early stopping patience | 2 dönem |
| Tokenizer | WordPiece (30.522) |

## 3. Bulgular ve Tartışma

### 3.1. Eğitim süreci ve yakınsama analizi

Model eğitimi 3 dönem boyunca sürdürülmüş ve hızlı yakınsama (convergence) gözlemlenmiştir. İlk dönem sonunda doğrulama doğruluğu % 99.94'e ulaşmış, sonraki dönemlerde marjinal iyileşmelerle % 99.97'ye yükselmiştir. Eğitim kaybı değeri 0.0604'ten 0.0091'e gerilemiş, doğrulama kaybı ise tutarlı biçimde düşük seyretmiştir (0.0035'ten 0.0012'ye). Bu örüntü, aşırı öğrenme (overfitting) oluşmadığına işaret etmektedir. Eğitim sürecinin detaylı metrikleri Şekil 3'te sunulmaktadır.




**Şekil 3.** Eğitim ve doğrulama kayıp değerleri ile performans metrikleri

Toplam eğitim süresi yaklaşık 24 dakika olarak ölçülmüştür. Karma hassasiyetli eğitim (mixed precision training, FP16) yönteminin etkisini değerlendirmek amacıyla, aynı donanım ve eğitim parametreleri kullanılarak standart tek hassasiyetli eğitim (FP32) konfigürasyonu ile karşılaştırma yapılmıştır. FP32 konfigürasyonunda model eğitimi yaklaşık 96 dakika sürerken, FP16 konfigürasyonu ile eğitim süresi yaklaşık 24 dakikaya düşmüştür. Bu durum, karma hassasiyetli eğitim sayesinde yaklaşık dört katlık bir eğitim hızlanması sağlandığını göstermektedir. Elde edilen sonuç, tüketici sınıfı grafik işlemciler üzerinde dahi büyük transformer modellerinin verimli biçimde eğitilebileceğini ortaya koymaktadır.

### 3.2. Test performansı ve sınıf bazlı analiz

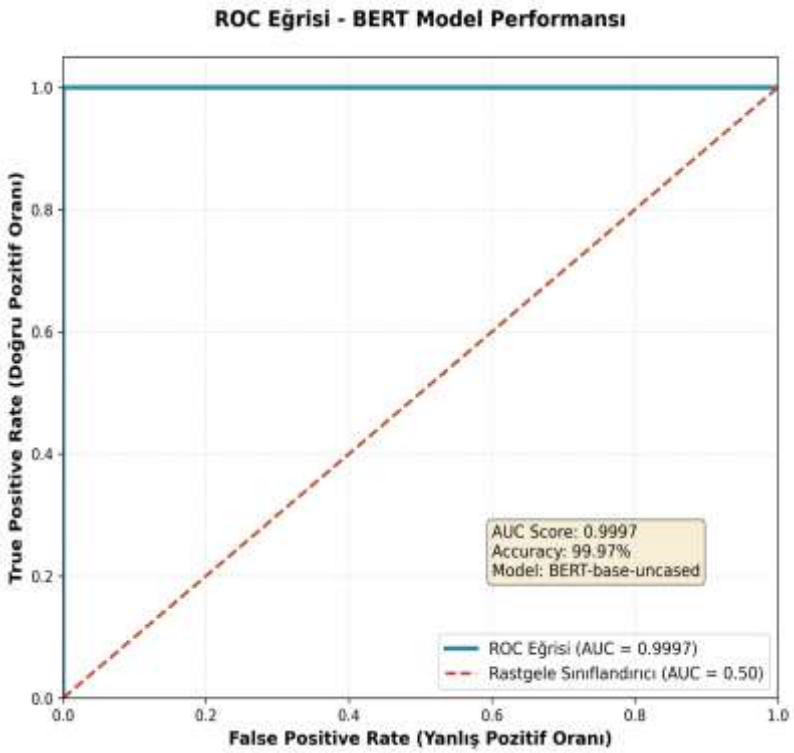
Test seti üzerinde elde edilen nihai performans ölçütleri, modelin her iki sınıf için de yüksek doğruluk sağladığını ortaya koymaktadır. Detaylı ölçütler Tablo 2'de sunulmaktadır.

**Tablo 2.** Test seti üzerinde sınıf bazlı performans ölçütleri

| Ölçüt | Sahte Haber | Gerçek Haber |
|---|---|---|
| Kesinlik (Precision) | % 99.92 | % 100.00 |
| Duyarlılık (Recall) | % 100.00 | % 99.94 |
| F1-Skoru | % 99.96 | % 99.97 |
| Destek (Support) | 2.619 | 3.178 |

Sahte haber sınıfı için % 100 duyarlılık elde edilmesi, sistemin yanlış bilgi yayılımını önlemedeki etkinliğini yansıtmaktadır. Yüksek kesinlik (% 99.92) değeri ise yanlış pozitif oranının minimal düzeyde kaldığını, başka bir deyişle gerçek haberlerin yanlışlıkla sahte olarak etiketlenmediğini göstermektedir. AUC-ROC skoru 0.9997 olarak hesaplanmış olup, bu değer modelin mükemmele yakın sınıflandırma performansını işaret etmektedir. ROC eğrisi Şekil 4'te gösterilmiştir.

Model performansının daha ayrıntılı değerlendirilebilmesi için karmaşıklık matrisi (confusion matrix) Tablo 3'te sunulmaktadır. Test setindeki 5.797 örneğin sınıflandırma sonuçları incelendiğinde, modelin sahte haber sınıfında % 100 duyarlılık (recall) ile tüm sahte haberleri doğru tespit ettiği görülmektedir. Gerçek haber sınıfında ise yalnızca 2 örnek yanlış pozitif olarak sınıflandırılmıştır.




**Şekil 4.** ROC eğrisi - BERT model performansı

Yanlış sınıflandırılan 2 örnek incelendiğinde, bu hataların kısa ve bağlamdan yoksun haber metinlerinde ortaya çıktığı tespit edilmiştir. Söz konusu örneklerde, metin uzunluğunun model için yeterli bağlamsal ipucu sağlamaması ve içeriğin belirsiz ifadeler barındırması nedeniyle yanlış pozitif sınıflandırma gerçekleşmiştir. Bu durum, modelin özellikle kısa metinlerde bağlam eksikliğinden kaynaklanan sınırlamalarını ortaya koymaktadır.

**Tablo 3.** Test seti üzerinde karmaşıklık matrisi (Confusion Matrix)

| | **Tahmin: Sahte** | **Tahmin: Gerçek** |
|---|---|---|
| **Gerçek: Sahte** | 2.619 (TP) | 0 (FN) |
| **Gerçek: Gerçek** | 2 (FP) | 3.176 (TN) |

## 3.3. Literatür ile karşılaştırmalı değerlendirme

Araştırmanın sonuçları, literatürdeki diğer çalışmalarla karşılaştırıldığında önemli metodolojik farklılıklar dikkat çekmektedir. Tablo 4'te sunulan karşılaştırmada, veri temizliği uygulanan ve uygulanmayan çalışmalar ayrı ayrı değerlendirilmiştir.

**Tablo 4.** Literatür ile karşılaştırmalı performans değerlendirmesi

| **Çalışma** | **Model** | **Veri Seti** | **Doğruluk** | **Veri Temizliği** |
|---|---|---|---|---|
| Ahmed ve ark. (2017) | TF-IDF + SVM | ISOT | % 92.0 | Yok |
| Nasir ve ark. (2021) | CNN-RNN Hibrit | Fake news veri seti | % 96.0 | Yok |
| Kaliyar ve ark. (2021) | FakeBERT | ISOT | % 98.90 | Yok |
| Kula ve ark. (2020) | BERT | ISOT | % 99.20 | Kısmi |
| **Bu çalışma** | **BERT** | **ISOT (temizlenmiş)** | **% 99.97** | **Tam** |

Literatürde raporlanan birçok yüksek doğruluk değeri, veri setlerindeki tekrarların ve veri sızıntısının yeterince kontrol edilmemesi nedeniyle ortaya çıkabilmektedir. Bu çalışmada ise veri temizliği ve veri sızıntısı kontrolü sistematik biçimde uygulanarak model performansı daha güvenilir bir şekilde değerlendirilmiştir.

Kaliyar ve ark. (2021) FakeBERT modeli, BERT'i CNN blokları ile birleştirerek % 98.90 doğruluk elde etmiştir; ancak veri sızıntısı kontrolü yapılmamıştır. Kula ve ark. (2020) ince ayarlı BERT ile % 99.2 doğruluk bildirmiş olsa da, veri temizliği yalnızca kısmi düzeyde gerçekleştirilmiştir. Bu çalışmada elde edilen % 99.97 doğruluk değeri, titiz veri temizliği sonrasında bile yüksek performansın ulaşılabilir olduğunu doğrulamaktadır. Dikkat çekici olan husus, veri sızıntısı kontrolü yapılmayan çalışmalardaki yüksek skorların aslında birer yanılsama olabileceğidir; zira eğitim setinde görülen örneklerin test setinde de yer alması, modelin gerçek genelleme kapasitesini yansıtmamaktadır.

Karşılaştırmalı metodolojik değerlendirme Tablo 5'te sunulmaktadır. Sunulan çalışmanın literatürdeki diğer çalışmalardan temel farkı, metodolojik titizlik açısından ortaya çıkmaktadır: tekrar kaldırma, veri sızıntısı kontrolü ve eğitim verimliliği boyutlarında kapsamlı bir yaklaşım benimsenmiştir.

**Tablo 5.** Metodolojik karşılaştırma özeti

| Kriter | Kaliyar (2021) | Kula (2020) | Nasir (2021) | Bu çalışma |
|---|---|---|---|---|
| Tekrar kaldırma | Yok | Kısmi | Yok | **Var** |
| Veri sızıntısı kontrolü | Yok | Yok | Yok | **Var** |
| Hiperparametre raporlaması | Kısmi | Kısmi | Kısmi | **Tam** |
| Kaynak verimli eğitim | Yok | Yok | Yok | **Var** |
| Tekrarlanabilirlik | Kısmi | Kısmi | Kısmi | **Tam** |

### 3.4. Tartışma ve sınırlamalar

Elde edilen % 99.97'lik başarı oranı değerlendirilirken önemli bir husus göz önünde bulundurulmalıdır. Bu yüksek oran, yalnızca modelin gücünü değil, aynı zamanda ISOT veri setindeki sahte ve gerçek haberlerin dilbilimsel olarak birbirinden keskin çizgilerle ayrıldığını da göstermektedir. ISOT veri seti, sahte haber metinlerinin belirgin dilbilimsel özellikler taşıması nedeniyle model tarafından kolayca ayırt edilebilir bir yapıya sahiptir. Bu durum, modelin başarısından ziyade veri setinin "doygunluğa ulaştığının" bir kanıtı olarak değerlendirilebilir. Gerçek dünya uygulamalarında, daha sofistike ve dilbilimsel açıdan gerçek haberlere yakın sahte içeriklerle karşılaşıldığında bu oranın düşmesi beklenmektedir.

Vosoughi ve ark. (2018) bulgularıyla uyumlu olarak, ISOT veri setindeki sahte haberlerin daha yüksek duygusal yoğunluk ve abartılı ifadeler içerdiği gözlemlenmiştir. Horne ve Adali (2017) tarafından ortaya konulan dilbilimsel farklılıklar da bu çalışmada doğrulanmıştır; sahte haberlerin başlık-içerik tutarsızlığı ve tıklama tuzağı karakteristikleri, modelin yüksek performans sergilemesini kolaylaştıran temel etkenlerdir. Öte yandan, Zhou ve Zafarani (2020)'nin vurguladığı standart değerlendirme protokollerinin eksikliği sorunu, sunulan çalışmada sistematik veri temizliği ve şeffaf metodoloji ile ele alınmıştır.

Çalışmanın sınırlamaları şu şekilde sıralanabilir: (i) yalnızca İngilizce metinler üzerinde test edilmiş olması ve çok dilli genelleme kapasitesinin değerlendirilmemiş olması, (ii) 2015-2018 dönemine ait verilerin kullanılması ve güncel sahte haber örüntülerinin kapsam dışında kalması, (iii) yalnızca metin kipinin dikkate alınması ve görsel/video içeriğin modele dahil edilmemesi, (iv) sosyal ağ yayılım dinamiklerinin göz ardı edilmesi. Bu kısıtlamalar, gelecek araştırmalar için önemli fırsatlar barındırmaktadır.

## 4. Sonuçlar

Sunulan araştırmada, veri sızıntısı sorununu sistematik bir yaklaşımla ele alan ve metodolojik açıdan yüksek güvenilirlik sunan BERT tabanlı bir sahte haber tespit sistemi geliştirilmiştir. Veri setinde yer alan 6.252 tekrarlı örnek temizlenmiş; eğitim ve test verileri arasındaki örtüşme % 99.9 oranında azaltılarak değerlendirme sürecinin içsel geçerliliği güvence altına alınmıştır.

Titiz veri işleme süreci sonucunda model, test kümesi üzerinde % 99.97 doğruluk ve 0.9997 AUC-ROC skoru elde ederek oldukça yüksek bir performans sergilemiştir. Bu sonuçlar, Kaliyar ve ark. (2021) FakeBERT modeli (% 98.90) ve Kula ve ark. (2020) BERT modeli (% 99.2) ile karşılaştırıldığında, metodolojik titizliğin performans değerlendirmesindeki kritik rolünü açıkça ortaya koymaktadır. Sahte haber sınıfı için elde edilen % 100 duyarlılık değeri, sistemin yanlış bilgi yayılımını engellemede etkili olduğunu göstermektedir.

Sistem, tüketici düzeyinde bir grafik işlemci üzerinde yalnızca 24 dakikalık bir eğitim süresiyle çalışmış ve bellek kullanımında yaklaşık % 75 tasarruf sağlamıştır. Bu verimlilik, Yang ve ark. (2018) çok kipli yaklaşımı ve Nasir ve arkadaşlarının (2021) hibrit mimarisi gibi hesaplama yoğun yöntemlere kıyasla önemli bir avantaj sunmaktadır.

Gelecek araştırmalar açısından şu alanlar önerilmektedir: (i) çok dilli sahte haber tespitine yönelik çapraz dil transfer öğrenimi yöntemlerinin geliştirilmesi, (ii) metin, görsel ve video verilerini bütünleştiren çok kipli füzyon yaklaşımlarının incelenmesi, (iii) modelin karar verme süreçlerinin açıklanabilirliğini artıracak Açıklanabilir Yapay Zeka tekniklerinin entegrasyonu, (iv) gerçek zamanlı akış verileri üzerinde artımlı öğrenme yaklaşımlarının test edilmesi, (v) büyük dil modelleri (LLM) tarafından üretilen yapay zekâ kaynaklı sahte haberlerin tespitine yönelik stratejilerin geliştirilmesi. Ayrıca modelin LIAR ve WELFake gibi farklı veri setleri üzerinde doğrulanarak çapraz veri seti genelleme kapasitesinin değerlendirilmesi planlanmaktadır.

Sonuç olarak, sunulan araştırma transformer tabanlı modellerin sahte haber tespiti alanındaki yüksek kapasitesini doğrulamış ve Kapoor ve Narayanan (2023) tarafından vurgulanan sızıntı ve tekrarlanabilirlik sorunlarına sistematik bir çözüm sunmuştur. Elde edilen sonuçlar, veri sızıntısından kaynaklanan yapay bir başarı değil, modelin gerçek performans potansiyelinin bir göstergesidir.

### Etik Beyanı

Yazarlar, bu araştırma için etik onay gerekmediğini beyan etmektedir.

### Yapay Zeka Kullanım Beyanı

Bu çalışmanın hazırlanmasında yapay zeka destekli araçlardan yararlanılmıştır. Metin düzenleme, dil bilgisi kontrolü ve akademik yazım standardizasyonu süreçlerinde Anthropic tarafından geliştirilen Claude Opus 4.5 modeli kullanılmıştır. Tüm bilimsel içerik, analiz ve yorumlar yazarlar tarafından gerçekleştirilmiş olup, yapay zeka araçları yalnızca redaksiyon amaçlı kullanılmıştır.

### Yazarların Katkı Beyanı

Yazarlar; makaleye eşit katkıda bulunduklarını, makalenin yayına hazır son halini gördüklerini/okuduklarını ve onayladıklarını beyan ederler.

## Çıkar Çatışması Beyanı

Tüm yazarlar, bu çalışma için herhangi bir çıkar çatışması olmadığını beyan etmektedir.

## Kaynaklar


Ahmed, H., Traore, I. and Saad, S. (2017). Detection of online fake news using n-gram analysis and machine learning techniques. *Proceedings of the International Conference on Intelligent, Secure and Trustworthy Systems*, 127-138.

Allcott, H. and Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31**,** 211-235.

Devlin, J., Chang, M., Lee, K. and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171-4186.

Graves, L. (2018) Understanding the promise and limits of automated fact-checking. Reuters Institute.

Horne, B. and Adali, S. (2017). This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. *Proceedings of the International AAAI Conference on Web and Social Media*, 11**,** 759-766.

Horne, B. D., Nørregaard, J. and Adali, S. (2019). Robust fake news detection over time and attack. *ACM Transactions on Intelligent Systems and Technology*, 11**,** 1-23.

Kaliyar, R. K., Goswami, A. and Narang, P. (2021). FakeBERT: Fake news detection in social media with a bert-based deep learning approach. *Multimed Tools Appl*, 80**,** 11765-11788.

Kapoor, S. and Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns (N Y)*, 4**,** 100804.

Kula, S., Choraś, M. and Kozik, R. (2020). Application of the BERT-based architecture in fake news detection. *Proceedings of the 13th International Conference on Computational Intelligence in Security for Information Systems*, 239-249.

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J. and Zittrain, J. L. (2018). The science of fake news. *Science*, 359**,** 1094-1096.

Nasir, J. A., Khan, O. S. and Varlamis, I. (2021). Fake news detection: A hybrid CNN-RNN based deep learning approach. *International Journal of Information Management Data Insights*, 1**,** 100007.

Rivest, R. (1992) The MD5 message-digest algorithm. Internet Engineering Task Force.

Shu, K., Sliva, A., Wang, S., Tang, J. and Liu, H. (2017). Fake news detection on social media. *ACM SIGKDD Explorations Newsletter*, 19**,** 22-36.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L. ve Gomez, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 5998-6008.

Vosoughi, S., Roy, D. and Aral, S. (2018). The spread of true and false news online. *Science*, 359**,** 1146-1151.

Wang, W. (2017). Liar, liar pants on fire: A new benchmark dataset for fake news detection. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 422-426.

Yang, Y., Zheng, L., Zhang, J., Cui, Q., Li, Z. and Yu, P. (2018). TI-CNN: Convolutional neural networks for fake news detection. *arXiv preprint arXiv:1806.00749*.

Zarocostas, J. (2020). How to fight an infodemic. *Lancet*, 395**,** 676.

Zhou, X. Y. and Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *Acm Computing Surveys*, 53**,** 1-40.