Yapay Zeka ve Güvenlik

Hamdi Murat Yıldırım

İD Bilkent Üniversitesi

Bilişim Sistemleri ve Teknolojileri Bölümü

Hukuk ve Yapay Zeka: Konferans & Panel Tartışmaları

Atılım Üniversitesi, 11 Mart 2025 14:30-15:30

Yapay Zeka ve Güvenlik

- **Yapay Zeka (YZ)**:

  "İnsan zekasına benzer şekilde görevleri yerine getirebilen makineler oluşturmayı hedefleyen geniş bir disiplindir; bu, öğrenme, sorun çözme, karar verme ve dil anlama gibi becerileri içerir."

- "İnsan zekasının gerektirdiği işlevleri makinelere kazandırmayı amaçlayan kapsamlı bir alandır; bu, öğrenme, problem çözme, karar verme ve dil anlama gibi yetenekleri kapsar."

Yapay Zeka ve Güvenlik

- **Büyük Dil Modelleri (LLM'ler)**, yapay zekanın doğal dil işleme (NLP) alanındaki en dikkat çekici başarılarından biridir.

- **Popüler LLM’lere örnekler**: GPT-3, GPT-4 (OpenAI), LaMDA, Bard,PaLM 2, (Google), BLOOM (BigScience), Claude (Anthropic), Llama (Meta), DeepSeek LLM, Coder (DeepSeek)

Yapay Zeka ve Güvenlik

- **Makine Öğrenimi (ML)**: bilgisayar sistemlerinin, verilerden anlamlı kalıplar çıkarmak ve bu kalıpları kullanarak, deneyim yoluyla performanslarını artırmak için tasarlanmış algoritmalar ve modeller geliştirdiği yapay zeka alanıdır

- **ML**'nin seçilen çeşitli algoritmalarını kullanan **Derin Öğrenme** teknikleri ile **LLM**'ler geliştirilir.

- **LLM**'lerde **Derin Öğrenme** ML tekniği yanı sıra, **Denetimli Öğrenme** ve **Denetimsiz Öğrenme** kullanılarak karmaşık dil görevlerini tamamlanır. **LLM**'ler, **ML** için bir takım iyileştirmeler sunar.

Yapay Zeka ve Güvenlik

LLM, ML ve YZ konularında **güvenli** uygulama geliştirmek için lisans eğitimi sırasında alınması gereken derslerin genel başlıkları:

- *Temel Bilgisayar Bilimi Dersleri (Programlamaya Giriş, Veri Yapıları ve Algoritmalar, Veritabanı Yönetim Sistemleri, Yazılım Mühendisliği)*
- *Matematik/İstatistik Dersleri*
- *YZ ve ML Dersleri (Teknik ve YZ Etiği dahil)*
- **Bilgisayar/Ağ Güvenliği, Siber Güvenlik, Uygulamalı Kriptografi Dersleri (Şifreleme, Dijital İmzalar, Kimlik ve Bütünlük Doğrulama vb.), Güvenli Kod Yazma Dersleri**
- **Seçmeli Dersler yukarıdaki başlıklarda ileri konular**

Yapay Zeka ve Güvenlik

Yapay zeka algoritmaları, matematiksel formüller ve denklemler üzerine kuruludur.

Makine öğrenmesi, verilerden öğrenen ve tahminler yapan modeller oluşturmak için matematiksel teknikler kullanır.

İstatistik, olasılık ve doğrusal cebir gibi matematiksel alanlar, veri analizinde önemli bir rol oynar.

Yapay zeka algoritmalarında matematiksel yöntemlerin yardımıyla büyük veri kümelerini analiz edilir ve anlamlı bilgiler çıkarılır.

Milyonlarca hatta milyarlarca parametreye sahip olabilen LLM'lerde en iyi parametre değerlerini bulmak için optimizasyon algoritmaları kullanılr.

Yapay Zeka ve Güvenlik

Yapay zeka güvenliği, yapay zeka sistemlerinin (ML, LLM, Bilgisayar Görü, Uzman, Robotik, Öneri, Otonom) güvenilir ve etik bir şekilde geliştirilmesini ve kullanılmasını sağlamayı hedefleyen, izleyen ana başlıklarla ilgili çok yönlü bir disiplindir:

Veri Güvenliği ve Gizliliği
Algoritmik Güvenlik
Siber Güvenlik
Etik ve Sorumluluk
Sağlamlık ve Güvenilirlik

Yapay Zeka ve Güvenlik

Veri Güvenliği ve Gizliliği:

YZ sistemlerinde büyük veri kullanımda, yetkisiz erişim, sızıntı ve kötü niyetle kullanıma karşı korunma yöntemlerine politikalar belirlenmeli, gerçekleştirilmeli ve uygulanmalıdır. Başlıklar:

Veri Şifreleme, Erişim Kontrolü, Veri Bütünlüğü, İnkar edilememezlik, Hesap verilebilirlik

Yapay Zeka ve Güvenlik

Algoritmik Güvenlik:

YZ algoritmaları, çeşitli kötü niyetli yaklaşımlarla bir takım yaygın manipülasyon teknikleri yardımıyla manipüle edilebilir:

YZ modellerini yanıltan özel tasarıma sahip girdiler.

Veri Zehirlenmesi (YZ modellerinin, eğitilirken kötü niyetle değiştirilmiş hatalı verilerle öğrenmesi), **Tersine Mühendislik** ile bir YZ Modelinin nasıl çalıştığını öğrenip, zayıf notları belirlemek.

Yapay Zeka ve Güvenlik

Algoritmik Güvenlik:

YZ algoritmaları, manipülasyon teknikleri **(devam)**:

Derin Sahtecilik (DeepFake)

Algoritmik Önyargı (kötü niyetle ayrıştırıcı önyargı oluşturmak)

Sosyal Mühendislik (bu tip saldırılarda YZ sistemlerinin kullanımı)

Kötü Niyetli Yazılım Güncellemeleri

Yapay Zeka ve Güvenlik

Algoritmik Güvenlik:

YZ algoritmaları, manipülasyon teknikleri **(devam)**:

Sosyal Medya Algoritmaları: kullanıcıların etkileşimlerine göre oluşan yanlış bilgi (toplumsal kutuplaşmayı artmasıyla sosyal uyumu zedelenmesi) veya propaganda (kitleleri yanlış veya bir amaca yönlendirme)

Filtre Balonu Etkisi: Algoritmalar, kullanıcıların ilgi alanlarına ve geçmiş etkileşimlerine göre sunduğu kişileştirilmiş kullanıcıların farklı görüşlere ve bakış açılarına maruz kaldığı içerik

Hasta Ayrımcılığı: demografik gruplara karşı ayrımcılık oluşabilir

Yapay Zeka ve Güvenlik

Algoritmik Güvenlik:

YZ algoritmaları, manipülasyon teknikleri **(devam)**:

Yanlış Teşhis ve Tedavi: Algoritmaların yanlış veya eksik verilerle eğitilmesi sonucunda.

Suç Tahmini: algoritmaların önyargılı verilerle eğitilmesi ile haksız yere suçlamalar olabilir

Reklam Algoritmaları ile Hedefli Manipülasyon: kullanıcı ilgi alanı ve davranışına göre kişileştirilen reklamların istenmeyen ürün veya hizmetleri satın alması sebep olması.

Yapay Zeka ve Güvenlik

Algoritmik Güvenlik:

Anomali Tespiti:

Algoritmaların normal davranış kalıplarından sapmaları tespit edilebilir.

Anormal davranışların işaretlenebilir ve analizi yapılabilir.

Anomali tespiti sayesinde kötü niyetli saldırıları erken aşamada önlenmesinin önü açılabilir.

Yapay Zeka ve Güvenlik

Siber Güvenlik: (YZ sistemlerinin, Siber Saldırılara karşı korunması için alınan ve uygulanan önlemler)

YZ tabanlı

* Ağ trafiği ve sistem günlükleri **sürekli** olarak **izleme** ile siber saldırıları tespit etme ve önleme;

* Saldırı Tespit sistemi (IDS);

* Saldırı Önleme Sistemi (IPS)

Yapay Zeka ve Güvenlik

Etik ve Sorumluluk: Etik ilkelere uyumlu YZ sistemi tasarlama ve geliştirme; YZ sisteminin verdiği kararlara dair sorumluluğun kimlerde olduğu ve hesap verebilmeleri

Sağlamlık ve Güvenirlik: YZ sistemlerinin üretiği sonuçların güvenirliği, testi ve doğrulanması; beklenmedik durum testi. LLM ve Üretken YZ sistemlerinde, modelin gerçek dışı, mantıksız veya doğru olmayan bilgiler üretmesi (**halüsinasyon**)

Yapay Zeka ve Güvenlik

Sağlamlık ve Güvenirlik:

Bazı Halüsinasyon nedenleri:

* Veri eksikliği veya hatalı olması: modelin hatalı öğrenmesi;

* Model karmaşıklığı

* Yaratıcılık ve Rastgelelik

* Eğitim Verilerinde Önyargılar

Yapay Zeka ve Güvenlik

Sağlamlık ve Güvenirlik:

Halüsinasyona karşı bazı önlemler:

Veri kalitesi Artırma ; Model Denetimi ve Doğrulama

Bilgi kaynalarına erişim ile doğrulama ; İnsan Geri Bildirim;

Geri Alma Artırılmış Üretim (RAG); Model Parametrelerini Ayarlama ; İnce ayar

Yapay Zeka ve Güvenlik

LLM uygulamaları Geliştirmek için Kullanılan Yazılım Teknolojileri

*** Temel Altyapı ve Kütüphaneler (Python, TensorFlow, PyTorch vb.)**

*** Bulut Platformları**

*** Veri İşleme ve Yönetimi**

*** Uygulama Geliştirme (LLM’lere erişim sağlayan API'ler, Web Çerçeveleri, Mobil uygulamalara entegrasyon**

Yapay Zeka ve Güvenlik

Güvenli Yazılım Geliştirme Yöntemleri :

Güvenli Yazılım Geliştirme Yaşam Döngüsü (Güvenliğin, yazılım geliştirme sürecinin her aşamasına entegre edilmesi)

Güvenli Kodlama Uygulamaları:

*** Geliştiriciler, güvenli kodlama prensiplerine uygun olarak kod yazmalıdır**

*** OWASP tarafından yayınlanan güvenlik açıklarına karşı önlem alınmalıdır.**

*** Girdi/Çıktı doğrulama ve hata yönetim**

*** Kod analiz araçları ile kodlardaki güvenlik zafiyeti düzenli ve otomatik kontrol edilmelidir. Güvenlik testleri**

Yapay Zeka ve Güvenlik

Güvenli Yazılım Geliştirme Yöntemleri

Erişim Kontrolü ve Kimlik Doğrulama:

*** Çok faktörlü kimlik doğrulama uygulanarak yetkisiz erişimler önlenmeli.**

*** Kullanıcıların sadece gerekli olan kaynaklara erişmeli: Başarılı kimlik doğrulama ardından erişim kontrolü politikasına göre kullanıcının kaynaklara erişim sağlanmalı.**

Yapay Zeka ve Güvenlik

Güvenli Yazılım Geliştirme Yöntemleri

Uygulamalı Kriptografi:

*** Depolanan ve/veya aktarılan hassas veriler güvenli şifreleme algoritmaları ile şifrelenmelidir.**

*** Bu algoritmaların anahtarları yönetim araçları kullanılmalıdır.**

*** Güvenli ağ trafiği için TLS vb. protokoller kullanılmalıdır.**

*** İhtiyaç halinde verileri barındıran dosyaların dijital imzası oluşturulmalı, doğrulanmalı (Veri kaynağına güven ve Kaynağın veriyi oluşturduğu inkar etmesini engellemek için vb.)**

Yapay Zeka ve Güvenlik

LLM'lerde veri mahremiyeti

Veri Anonimleştirme: Kişisel bilgileri veri setlerinden çıkarma veya değiştirme işlemidir. Amaç, verileri analiz edilebilir ve kullanılabilir tutarken bireylerin kimliğini gizlemektir.

LLM'lerde Veri Anonimleştirme:

Gizliliği Koruma; Yasal Uyumluluk (KVVK, GDPR vb.)

Güvenin Sağlanması: LLM'lerde kişilerin kişisel verilerinin kötüye kullanılmasına engel olmak.

Yapay Zeka ve Güvenlik

LLM'lerde veri mahremiyeti

Bazı Veri Anonimleştirme Teknikleri:

Yerine Koyma, Maskeleme, Genelleştirme, Veri Gizleme,

Farklılaştırılmış Gizlilik (Differential Privacy),

K-anonimlik (K-anonymity), L-çeşitliliği (L-diversity),

T-yakınlığı (T-closeness)

Yapay Zeka ve Güvenlik

LLM'lerde veri mahremiyeti

Veri anonimleştirme süreçlerinde kullanılan bazı ileri kriptografik yöntemler:

Homomorfik Şifreleme; Sıfır Bilgi Kanıtları (Zero-Knowledge Proofs);

Güvenli Çok Taraflı Hesaplama (Secure Multi-Party Computation – SMPC);

Veri Karıştırma (Data Shuffling) ve Permütasyon (Permutation): Kriptografik olarak sözdegüvenli rasgele sayı üreteçleri kullanılır (CSPRNG)

Yapay Zeka ve Güvenlik

YZ güvenliği ile ilgili çeşitli konuları ele alan Standartlar ve Düzenlemeler

ISO/IEC 42001: Yapay zeka yönetim sistemleri için uluslararası bir standarttır. Yapay zeka ile ilişkili riskleri ve fırsatları yönetmek için bir çerçeve sunar.

Kişisel Verilerin Korunması Kanunu (KVKK): YZ sistemlerinin kişisel verileri işlerken KVKK'ya uygunluğu

Avrupa Birliği Yapay Zeka Yasası (Yapay zeka uygulamalarını risk seviyelerine göre sınıflandırır)

Yapay Zeka ve Güvenlik

ISO/IEC 42001 yapay zeka güvenliği ile ilgili temel konuları:

1. Risk Yönetimi

2. Veri Yönetimi:

3. Algoritmik Güvenlik ve Şeffaflık:

4. Etik ve Sorumluluk:

5. Sürekli İzleme ve İyileştirme:

Yapay Zeka ve Güvenlik

AB Yapay Zeka Yasası, yapay zeka güvenliği ile ilgili ana başlıklar:

1. Risk Bazlı Yaklaşım:

2. Kabul Edilemez Riskli YZ Sistemleri:

Yasaklananlar: sosyal puanlama sistemleri, insanların bilinçaltını manipüle eden sistemler ve biyometrik tanımlama sistemlerinin bazı türleri

3. Yüksek Riskli Yapay Zeka Sistemleri: sıkı gereklilikler ve denetimler: Siber Güvenlik dahil

Yapay Zeka ve Güvenlik

AB Yapay Zeka Yasası, yapay zeka güvenliği ile ilgili ana başlıklar:

4. Sınırlı Riskli Yapay Zeka Sistemleri: derin sahte (deepfake) üreten sistemlerin bildirilmesi

5. Minimal Riskli Yapay Zeka Sistemleri

6. Genel Amaçlı Yapay Zeka (GPAI) Modelleri

7. Uygulama ve Denetim: (İhlaller durumunda ciddi para cezaları)

Yapay Zeka ve Güvenlik

AB Yapay Zeka Yasası'nda belirtilen risklerin önlenmemesi

Durumda gerçek hayat felaket senaryoları:

Kitlesel Gözetim, Yanlış Tanımlama ve Ayrımcılık

Sosyal Dışlanma ve Ayrımcılık

Otonom Silah Sistemleri: Kontrol Kaybı ve İnsan Kayıpları

Yanlış Bilgi ve Manipülasyon: Demokratik Süreçlerin Zedelenmesi

Kritik Altyapıların Çökmesi: Siber Saldırılar ve Arızalar

Yapay Zeka ve Güvenlik

Kaynaklar

* Google Gemini Sorguları ve Sonuçları

* AB Yapay Zeka Yasası

https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689
https://artificialintelligenceact.eu/ai-act-explorer/

* OWASP YZ Güvenlik ve Gizlilik Kılavuzu

https://owasp.org/www-project-ai-security-and-privacy-guide/