

Yapay Zeka Moderasyon

Yapay zeka moderasyon sistemleri ve işleyiş prensipleri

Şirket Bilgileri

-  Şirket: Devexa Yazılım LTD. ŞTi.
 -  Merkez: Kadıköy, İstanbul
 -  İletişim: ai@niva.media
 -  Web: niva.media | devexa.tech
 -  Versiyon: v1.1
 -  Yürürlük Tarihi: 19 Temmuz 2025
 -  Son Güncelleme: 19 Temmuz 2025
-

Giriş

Niva, kullanıcı güvenliğini sağlamak ve topluluk kurallarını korumak amacıyla gelişmiş yapay zeka destekli moderasyon sistemleri kullanır. Bu sistemler metin, görsel, ses ve davranış analiziyle çalışır.

Otomatik İçerik Analizi

Metin Analizi

-  Küfür ve hakaret tespiti
-  Spam içerik algılama
-  Tehdit ve zorbalık analizi
-  Uygunsuz dil ve ifadelerin engellenmesi

Görsel Analizi

-  Uygunsuz görsel içerik tespiti
-  Cinsel içerik tanıma
-  Şiddet ve zararlı görsel analizleri
-  Sahte profil görsellerinin tespiti

Ses Analizi

-  Ses tonu ve duygu analizi
 -  Saldırgan, tehditkâr konuşma tespiti
 -  Ses kalitesi ve uygunluk kontrolü
 -  Arka plan gürültüsü filtreleme
-

Gerçek Zamanlı Moderasyon

Anlık Müdahale

-  Milisaniye düzeyinde içerik analizi

- 🛑 Gerçek zamanlı içerik engelleme
- 🚫 Geçici susturma/uyarı sistemi
- ⚠️ Kullanıcıya otomatik bildirim gönderimi

📊 Risk Değerlendirmesi

- 🎯 Davranış örüntüsü analizi
 - 📈 Risk puanı belirleme
 - 🔍 Şüpheli hareket takibi
 - 🚨 Acil durum senaryolarında otomatik devreye girme
-

🧠 Makine Öğrenmesi

📚 Sürekli Öğrenme

- 🔄 Model güncellemeleri
- 📊 Yeni tehdit türlerini öğrenme
- 🎯 Doğruluk oranlarının artırılması
- 🚀 Performans optimizasyonu

🌐 Dil ve Anlam Analizi

- 🇹🇷 Türkçe metin desteği
 - 🌍 Çok dilli analiz kabiliyeti
 - 🗣️ Argo, kısaltma ve yerel dil desteği
 - 📄 Anlam tabanlı filtreleme teknolojisi
-

👥 İnsan Moderasyonu ile Entegrasyon

🤝 Hibrit Yapı

- 🤖 AI ön analiz yapar
- 👥 İnsan moderatör nihai kararı verir
- ⚖️ Karmaşık durumlar uzmanlarca incelenir
- 🔄 Geri bildirimle AI kendini geliştirir

📞 Eskalasyon Seviyeleri

- ⚠️ Düşük risk: Tamamen otomatik işlem
 - 💎 Orta risk: İnsan onayı gerekir
 - 🔴 Yüksek risk: Moderasyon ekibi devrede
 - 🚨 Kritik durum: Anında uzman müdahalesi
-

🛡️ Özel Koruma Alanları

🛡️ Çocuk Güvenliği

- 🧒 Yaş tespiti ve filtreleme

- 🚫 Çocuklara yönelik içeriklerin engellenmesi
- 🔒 Grooming davranışlarının algılanması
- 📞 Yasal bildirimlerin otomatik yapılması

🛡️ Kadın Güvenliği

- 💜 Taciz davranışı tespiti
- 🚫 Stalking ve ısrarlı takip analizi
- ⚠️ Tehdit ve psikolojik baskı algısı
- 🔒 Güvenli mod aktivasyon sistemi

📊 Şeffaflık ve Raporlama

📈 Performans Verileri

- 🎯 Doğruluk oranı: %99.2
- ⚡ Ortalama analiz süresi: <50ms
- 🔍 Yanlış pozitif oranı: <%0.5
- 📊 Günlük analiz edilen içerik: 10 milyon+

📋 Kullanıcı Hakları

- 🗣️ AI kararlarına itiraz hakkı
- 📞 İnsan moderatör talep etme imkânı
- 📄 Detaylı açıklama isteme hakkı
- ⚖️ Şeffaf ve adil değerlendirme süreci

🔒 Veri Güvenliği

🛡️ Gizlilik Politikası

- 🔒 Uçtan uca veri şifreleme
- 🗑️ Tüm analiz verileri geçici olarak işlenir
- 🚫 Kişisel veriler kalıcı olarak saklanmaz
- ⚖️ KVKK ve GDPR'ye tam uyum

🔒 Güvenlik Sertifikaları

- 🏢 ISO 27001
- 🔒 SOC 2 Type II
- 🇪🇺 EU GDPR uygunluğu
- ☁️ Güvenli bulut altyapısı (endüstri standartlarına uygun)

🚀 Gelecekteki Geliştirmeler

- 🧠 GPT-4 Turbo'nun daha derin entegrasyonu
- 👁️ Gelişmiş görsel tanıma teknolojileri

- 🗣️ Audio AI ile konuşma analizinde derinleşme
 - 🌐 Metaverse içi içerik denetimi
 - 🖥️ Edge computing ile cihaz üstü anlık moderasyon
-

🧐 Sıkça Sorulan Sorular (SSS)

- ❓ AI sistemimle ilgili yanlış karar verdiyse ne yapmalıyım?
💬 İnsan moderatöre itiraz formu göndererek süreci başlatabilirsiniz.
 - ❓ Kişisel verilerim sistem tarafından saklanıyor mu?
❌ Hayır. Yalnızca geçici analiz yapılır, veriler kalıcı olarak saklanmaz.
-

📅 Son Güncelleme: 19 Temmuz 2025

📄 Versiyon: v1.1

🤖 Kullanılan AI Sistemi: GPT-4 Turbo + Özel Moderasyon Modelleri

✉️ İletişim: ai@niva.media

📌 Bu Güncellemede Neler Değişti? (19 Temmuz 2025 - v1.1)

- 📅 Tüm tarih ve versiyonlar güncellendi
- 📝 Metin yapısı sadeleştirildi, başlıklar hizalandı
- 📄 Metin, görsel ve ses analizleri netleştirildi
- 🧠 Hibrit yapı daha detaylı anlatıldı
- 🛡️ Çocuk ve kadın güvenliği alt başlıkları yeniden yapılandırıldı
- 📊 Performans metrikleri ve kullanıcı hakları kısmı güncellendi
- 🛡️ Güvenlik ve gizlilik bölümü GDPR/ISO/SOC sertifikalarıyla genişletildi
- 🚀 Gelecek planları ve teknoloji roadmap'i eklendi