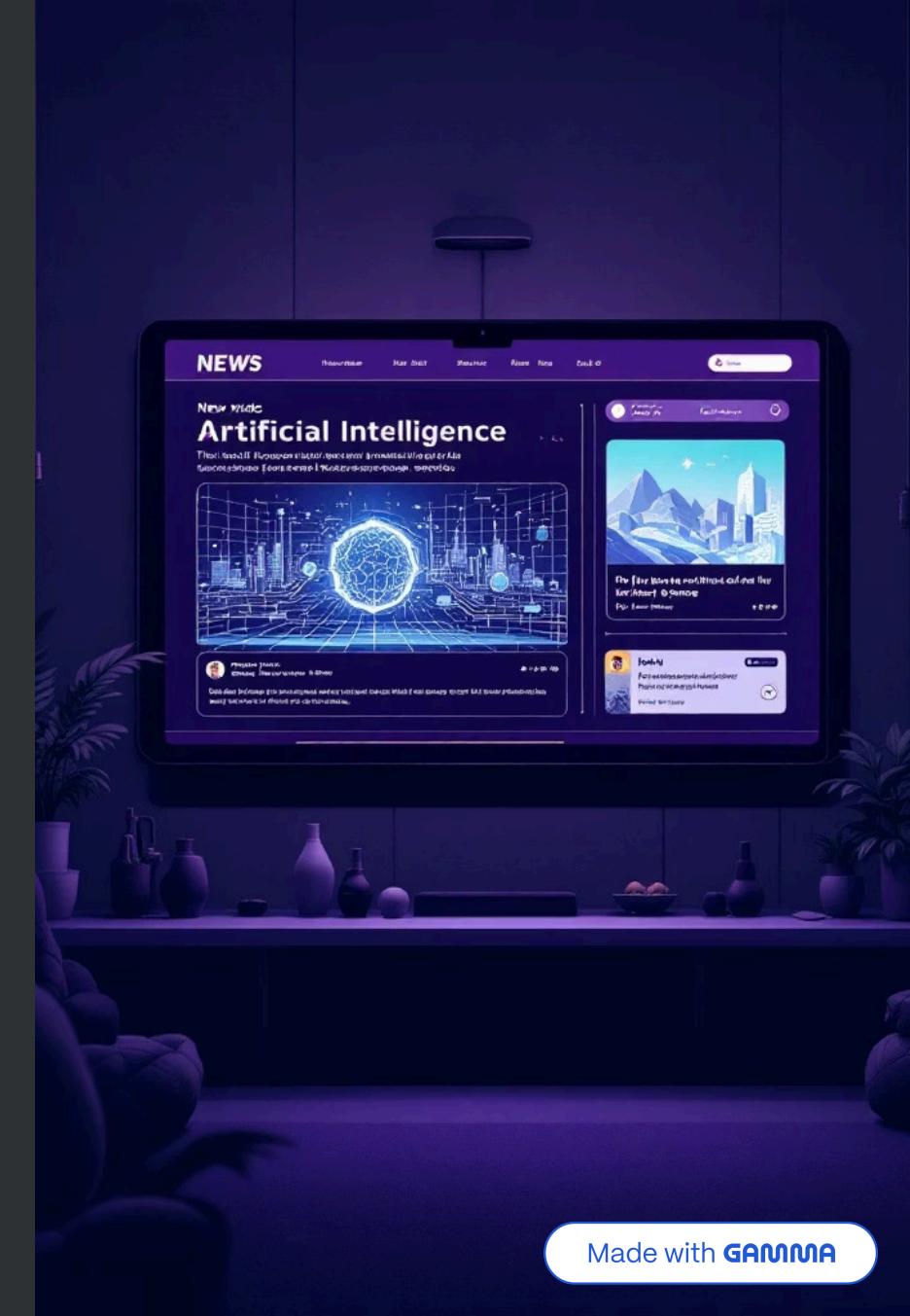


Haber Başlıklarını Sınıflandırma

BERT ile Doğal Dil İşleme

210.000 haber başlığını BERT modeliyle sınıflandırarak doğal dil işleme tekniklerinin pratik uygulamasını gösteriyoruz. Python ile model eğitimi ve interaktif web arayüzü ile gerçek zamanlı test.



Veri Seti: HuffPost Haber Arşivi

Veri Seti Özellikleri

2012-2022 yılları arasında HuffPost'tan toplanan 210.000 haber başlığı. Veri seti, 42 farklı kategoriye ait haberlerin başlıklarını, yazarlarını, bağlantılarını ve özetlerini içeriyor.

2018 öncesi yaklaşık 200.000, sonrası 10.000 başlık bulunuyor.





İnteraktif Web Uygulaması

Eğitilen BERT modelini kullanarak gerçek zamanlı haber başlığı sınıflandırması yapabilen web uygulaması. Kullanıcılar herhangi bir haber başlığını girerek modelin tahminini ve güven oranını görebilirler.

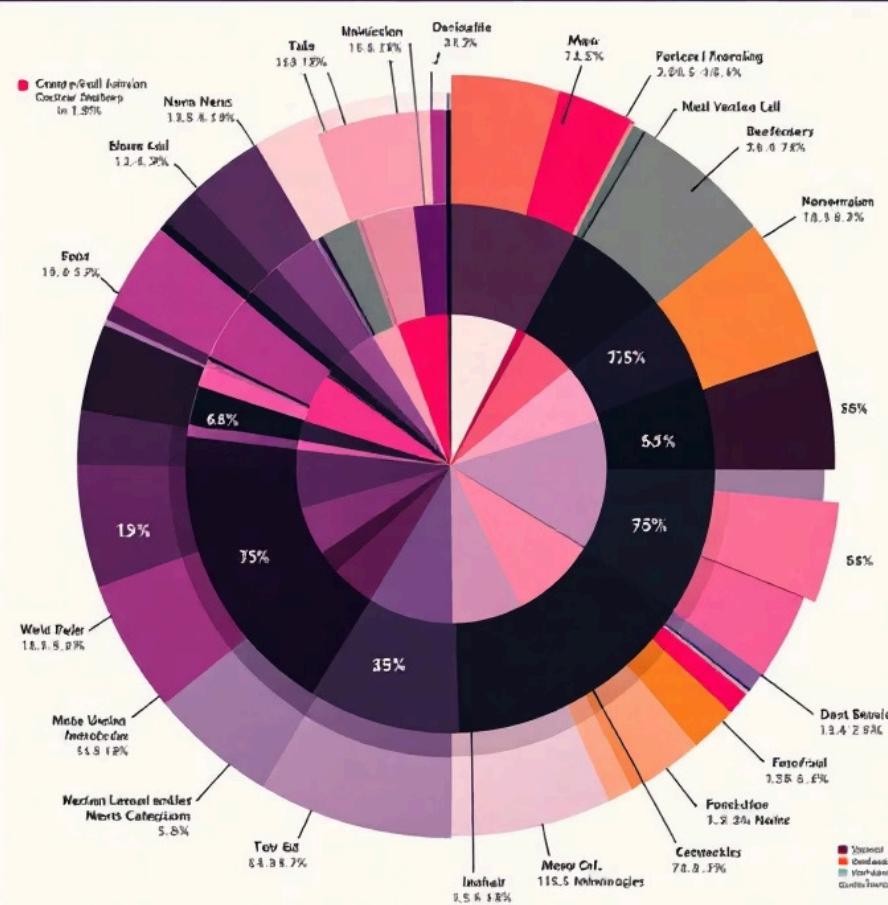
Uygulama özellikleri:

- Anlık sınıflandırma sonuçları
- 42 kategori arasında tahmin
- Güven skoru gösterimi
- Kullanıcı dostu arayüz

News Categories – Pie chart

- 177 -

What is the basis of your view categorization?



Kategorilerin Dağılımı

Toplam 42 kategoriden oluşan veri setinde en baskın kategoriler: POLITICS (35.602), WELLNESS (17.945) ve ENTERTAINMENT (17.362). Dengesiz dağılım model performansını etkileyen önemli bir faktör.

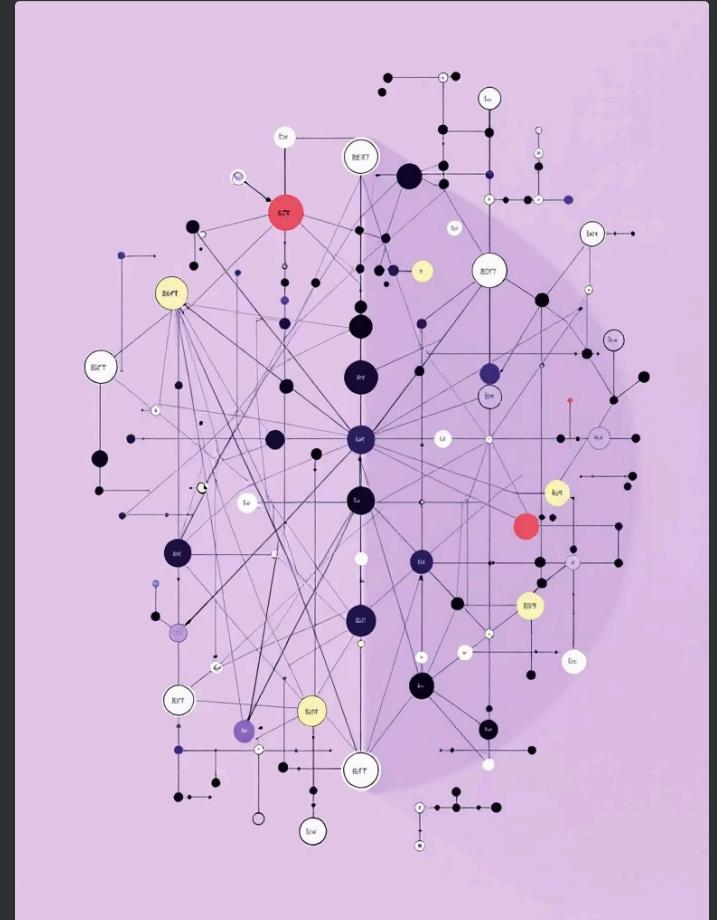


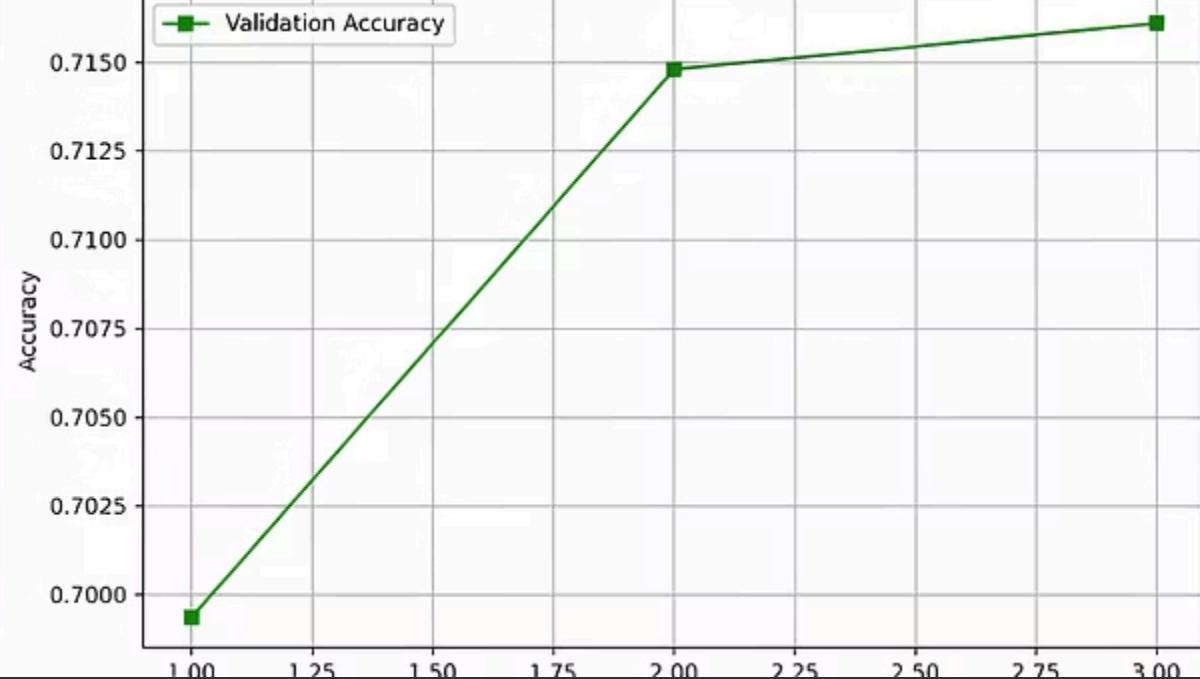
Model Mimarisi: BERT-base-uncased

Önceden Eğitilmiş Dil Modeli

BERT (Bidirectional Encoder Representations from Transformers), metni her iki yönden işleyen devrimci bir NLP modelidir. Case-sensitive olmayan versiyonu kullanılarak haber başlıklarından anlamsal vektörler çıkarıldı.

42 kategori için sınıflandırma katmanı eklenerek fine-tuning yapıldı.





Model Öğrenme Davranışı

Grafik Analizi

Training Loss: Eğitim kaybı ilk epokta hızlı düşüş gösteriyor, model veriyi öğreniyor.

Validation Loss: 1. epoktan sonra doğrulama kaybı sabitleniyor, overfitting riski var.

Model Performansı

Validation Accuracy: Doğruluk oranı sürekli artıyor ve %71-72'de sabitleniyor.

Optimum Nokta: Model ezberlemeden önce ideal performansa ulaşıyor, erken durdurma önerilir.

Genel Model Performansı

72%

Doğruluk

Validation setinde elde edilen accuracy
oranı

42

Kategori

Sınıflandırma yapılan toplam konu sayısı

210K

Haber Başlığı

Modelin eğitildiği veri miktarı

40'tan fazla sınıfı karmaşık bir problem için %71-72 doğruluk orta-iyi başarı seviyesi. F1-score'un accuracy ile paralel ilerlemesi, sınıflar arasında aşırı dengesiz başarım olmadığını gösteriyor.

En İyi Performans Gösteren Kategoriler

01

STYLE & BEAUTY

F1: ~0.89 - Tematik kelimeler açık

02

WEDDINGS

F1: ~0.88 - Anlamsal çakışma düşük

03

TRAVEL

F1: ~0.86 - Özgün içerik yapısı

04

DIVORCE

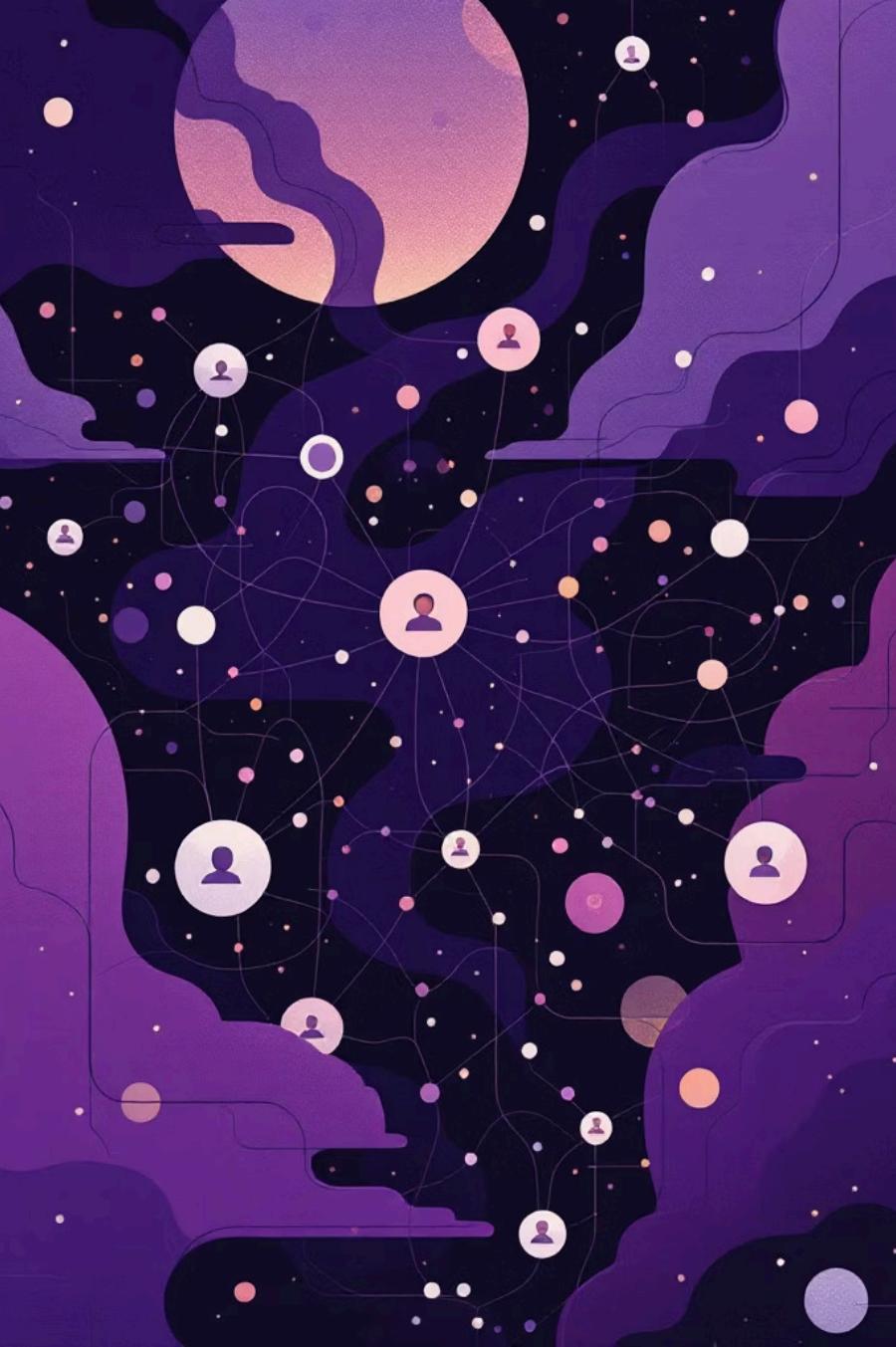
F1: ~0.84 - Ayırt edici terminoloji

05

HOME & LIVING

F1: ~0.83 - Yeterli veri miktarı

Bu kategoriler net tematik kelimelere sahip, diğer sınıflarla az örtüşüyor ve model tarafından yüksek precision ve recall ile öğrenilmiştir.



En Zorlanan Kategoriler

⚠️ U.S. NEWS

F1: ~0.38

Genel haber dili içeriyor

⚠️ GOOD NEWS

F1: ~0.39

Ayırt edici kelimeler zayıf

⚠️ IMPACT

F1: ~0.45

Birden fazla kategori ile örtüşüyor

Bu sınıflar WORLD NEWS, POLITICS ve ENTERTAINMENT gibi kategorilerle sık karışıyor. Genel haber dili ve sınırlı veri miktarı performansı etkiliyor.

Dengesiz Veri Dağılımı ve Çözüm Önerileri

Problem

- POLITICS (~30K) ve WELLNESS (~15K) baskın
- Eğitim, Sanat gibi sınıflar 1K civarında
- Model büyük sınıflara bias geliştirdi
- Küçük sınıflarda recall düşüşü

Çözüm Önerileri

- Class-weight kullanımı
- Oversampling / augmentation
- Daha fazla veri toplama
- Regularization artırma

Dengesiz veri dağılımı, modelin performansını ciddi şekilde etkiliyor. Bu sorunu azaltmak için sınıf ağırlıkları kullanımı, veri artırma teknikleri veya daha fazla veri toplama stratejileri uygulanabilir.

