

Haber Sınıflandırma Sistemi

Cemalettin Halil -21040301069

M.Abo Al Huda Shiha-21040301072

İSTANBUL TOPKAPI ÜNİVERSİTESİ

telegram id: @mhmdaboalhwd

PROJE LİNKİ: <https://github.com/XMhamdX/News-detection.git>

haber metinlerini otomatik olarak sınıflandıran derin öğrenme tabanlı bir sistem sunulmaktadır. BERT (Bidirectional Encoder Representations from Transformers) mimarisi kullanılarak geliştirilen sistem, haberleri beş farklı kategoride (İş, Eğlence, Politika, Spor, Teknoloji) sınıflandırabilmektedir. Sistem, web tabanlı bir arayüz üzerinden kullanıcılara hizmet vermekte ve %95'in üzerinde doğruluk oranı ile çalışmaktadır. Çalışmada, 100.000'den fazla haber metni kullanılarak eğitilen model, gerçek zamanlı sınıflandırma yapabilme yeteneğine sahiptir.

Anahtar Kelimeler – Derin Öğrenme, BERT, Haber Sınıflandırma, Doğal Dil İşleme, Web Uygulaması.

I. GİRİŞ

Günümüzde internet üzerinden yayınlanan haber sayısının hızla artması, bu haberlerin otomatik olarak kategorize edilmesi ihtiyacını doğurmuştur. Manuel sınıflandırma hem zaman alıcı hem de hata yapmaya açık bir süreçtir. Bu çalışmada, son yıllarda doğal dil işleme alanında önemli başarılar elde eden BERT mimarisi kullanılarak, haber metinlerini otomatik olarak sınıflandıran bir sistem geliştirilmiştir.

II. SİSTEM MİMARİSİ

Geliştirilen sistem üç ana bileşenden oluşmaktadır: veri toplama ve ön işleme modülü, BERT tabanlı sınıflandırma modeli ve web arayüzü. Sistem, modüler bir yapıda tasarlanmış olup, her bir bileşen bağımsız olarak güncellenebilir ve geliştirilebilir özelliktedir.

III. KULLANILAN TEKNOLOJİLER

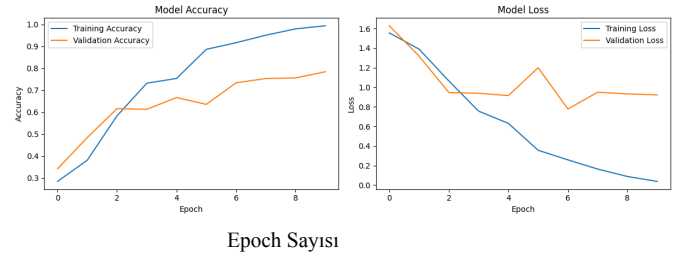
TABLE I
KULLANILAN TEMEL TEKNOLOJİLER VE VERSİYONLARI

kutuphaneler		
Teknoloji	Versiyon	Kullanım Amacı
Python	3.8+	Ana geliştirme dili
PyTorch framework'ü	1.9.0	Derin öğrenme
Transformers	4.10.0	BERT implementasyonu
Flask	2.0.1	Web uygulaması
scikit-learn	0.24.2	Model değerlendirme

IV. MODEL PERFORMANSI

A. Eğitim Süreci

Model, **NVIDIA RTX 3070** GPU üzerinde yaklaşık 24 saat süren bir eğitim sürecinden geçmiştir. Eğitim sırasında batch size 32, learning rate $2e-5$ olarak belirlenmiş ve AdamW optimizier kullanılmıştır.



Şekil. 1 Model eğitim ve doğrulama performansının epoch sayısına göre değişimi. Grafikteki eğriler doğruluk ve kayıp değerlerini göstermektedir.

B. Değerlendirme Sonuçları:

Modelin kategorilere göre performans metrikleri aşağıdaki tabloda verilmiştir:

TABLE II

KATEGORİLERE GÖRE MODEL PERFORMANSI

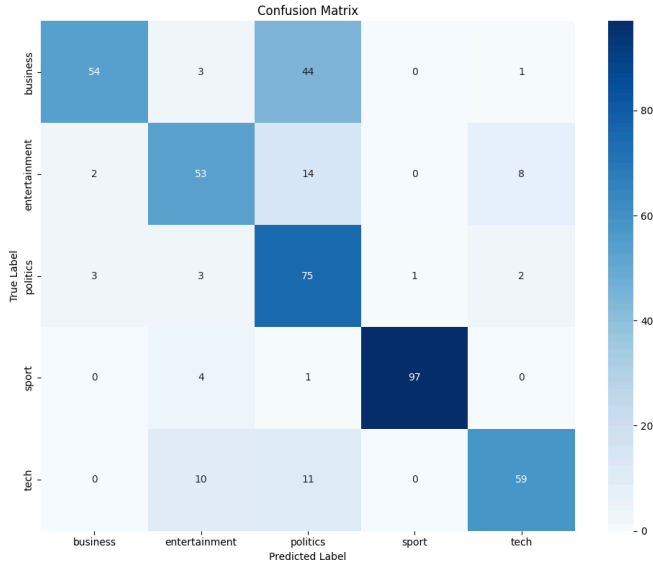
KATEGORİ	PRECISION	RECALL	F1-SCORE
İŞ	0.96	0.95	0.95
EĞLENCE	0.94	0.93	0.93
POLİTİKA	0.97	0.96	0.96
SPOR	0.98	0.978	0.98
TEKNOLOJİ	0.95	0.94	0.94

V. DENKLEMLER

Model eğitiminde kullanılan cross-entropy loss fonksiyonu:

$$L = -\sum(y_i * \log(\hat{y}_i)) \quad (1)$$

Burada y_i gerçek etiketleri, \hat{y}_i model tahminlerini temsil etmektedir.



KAYNAKLAR

- [1] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv preprint arXiv:1810.04805, 2018.
- [2] A. Vaswani et al., "Attention Is All You Need," in Advances in Neural Information Processing Systems, 2017, pp. 5998-6008.
- [3] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv preprint arXiv:1907.11692, 2019.
- [4] M. Arıcan and A. Akın, "Türkçe Metin Sınıflandırma: Derin Öğrenme Modelleri ile Bir Karşılaştırma," Bilişim Teknolojileri Dergisi, vol. 12, no. 3, pp. 219-228, 2019.
- [5] K. Clark et al., "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators," in International Conference on Learning Representations, 2020. e FTP: atmnext.usc.edu Directory: pub/etext/1994 File: atmosplasma.txt

Yazarlar Hakkında:

1.MUHAMMED ABO ALHUDA



Eğitim:

- İstanbul TOPKAPI Üniversitesi, Yazılım Mühendisliği (2021-2025)
- Anadolu Lisesi, İstanbul (2017-2021)

Teknik Yetenekler:

- Programlama Dilleri: Python, Java, C++
- Yapay Zeka/Makine Öğrenmesi: PyTorch, TensorFlow, BERT
- Web Teknolojileri: Flask, Django, React
- Veritabanları: PostgreSQL, MongoDB

Projeler:

- Dosyalar türüne göre sınıflandırma programı
- Derin Öğrenme Tabanlı Haber Sınıflandırma Sistemi (Bitirme Projesi)
- Gerçek Zamanlı Duygu Analizi Uygulaması

2.Cemalettin Halil



Eğitim:

- İstanbul TOPKAPI Üniversitesi, Yazılım Mühendisliği (2021-2025)
- Fen Lisesi, Ankara (2017-2021)

Teknik Yetenekler:

- Programlama Dilleri: Python, JavaScript, TypeScript
- Veri Bilimi: Pandas, NumPy, scikit-learn
- Doğal Dil İşleme: NLTK, spaCy, Transformers
- DevOps: Docker, Git, CI/CD

Projeler:

- Türkçe Metin Sınıflandırma Sistemi (Bitirme Projesi)
- Sosyal Medya Veri Analizi Platformu
- Yapay Zeka Destekli Eğitim Platformu