İSTANBUL MEDENİYET ÜNİVERSİTESİ

MÜHENDİSLİK VE DOĞA BİLİMLERİ FAKÜLTESİ
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ
DERİN ÖĞRENME DERSİ PROJESİ

Proje Adı

Derin Öğrenme ile Türkçe Haber Kategorisi Sınıflandırma

22120205016 Burak Güven

Ders Sorumlusu

Dr. Öğr. Üyesi İshak Dölek

Aralık, 2025

İstanbul Medeniyet Üniversitesi, İstanbul

İÇİNDEKİLER

Sayfa No

1. Proje Konusu

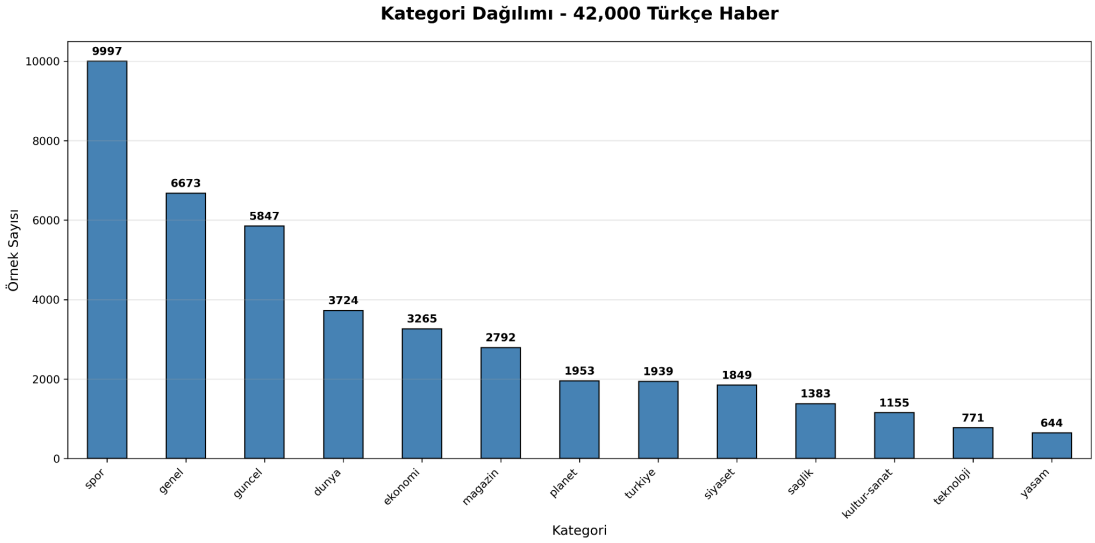
Dijital çağda bilgi akışının hızlanmasıyla birlikte, haber ajansları ve medya kuruluşları her gün binlerce içerik üretmektedir. Bu yoğun veri akışının manuel olarak kategorize edilmesi hem zaman alıcı hem de maliyetli bir süreçtir. Bu projenin temel amacı, derin öğrenme teknikleri kullanılarak Türkçe haber metinlerini "Spor", "Ekonomi", "Dünya" ve "Güncel" olmak üzere dört ana kategoride otomatik olarak sınıflandırabilen yüksek doğruluklu bir sistem geliştirmektir.

Doğal Dil İşleme (NLP) alanında metin sınıflandırma temel bir problem olmakla birlikte, Türkçe gibi sondan eklemeli ve morfolojik olarak zengin dillerde bu işlem ekstra zorluklar barındırmaktadır. Bu çalışmada, haber içeriklerinin otomatik olarak etiketlenmesi sayesinde bilgiye erişimin demokratikleşmesi, medya takibinin kolaylaşması ve kullanıcılara kişiselleştirilmiş içerik sunulması hedeflenmektedir. Ayrıca geliştirilen sistem, sadece akademik bir model olarak kalmayıp, Gradio arayüzü ile son kullanıcıya hitap eden pratik bir uygulamaya dönüştürülmüştür.

2. Veri Setinin Belirlenmesi

Proje kapsamında, YTÜ Kemik NLP Grubu tarafından oluşturulan ve literatürde kabul gören "42,000 Turkish News in 13 Classes" veri seti temel alınmıştır. Orijinal veri seti 13 farklı kategoride 42.000 haber metninden oluşmaktadır. Ancak, modelin başarısını optimize etmek ve sınıf ayrımını netleştirmek amacıyla, içeriği belirsiz olan "Genel" kategorisi ve örnek sayısı yetersiz olan diğer sınıflar elenerek veri seti revize edilmiştir.

Filtreleme işlemi sonucunda proje, "Spor" (9,997), "Güncel" (5,847), "Dünya" (3,724) ve "Ekonomi" (3,265) olmak üzere toplam 22,833 haber metni üzerinde kurgulanmıştır. Veri setinin istatistiksel analizi yapıldığında, ortalama metin uzunluğunun 1,783 karakter olduğu ve profesyonel haber dili kullanıldığı için veri kalitesinin yüksek olduğu görülmüştür. Sınıflar arasında gözlemlenen dengesizlik (Class Imbalance), eğitim aşamasında "Class Weighting" teknikleri ile dengelenerek modelin çoğunluk sınıfına (Spor) önyargılı yaklaşması engellenmiştir. Veri seti, modelin genelleme yeteneğini artırmak ve overfitting riskini azaltmak amacıyla %72 Eğitim, %13 Doğrulama (Validation) ve %15 Test olarak parçalara ayrılmıştır.



Şekil 1 : Veri Seti Kategori Dağılımı

3. Uygulanacak Yöntem ve Yaklaşım

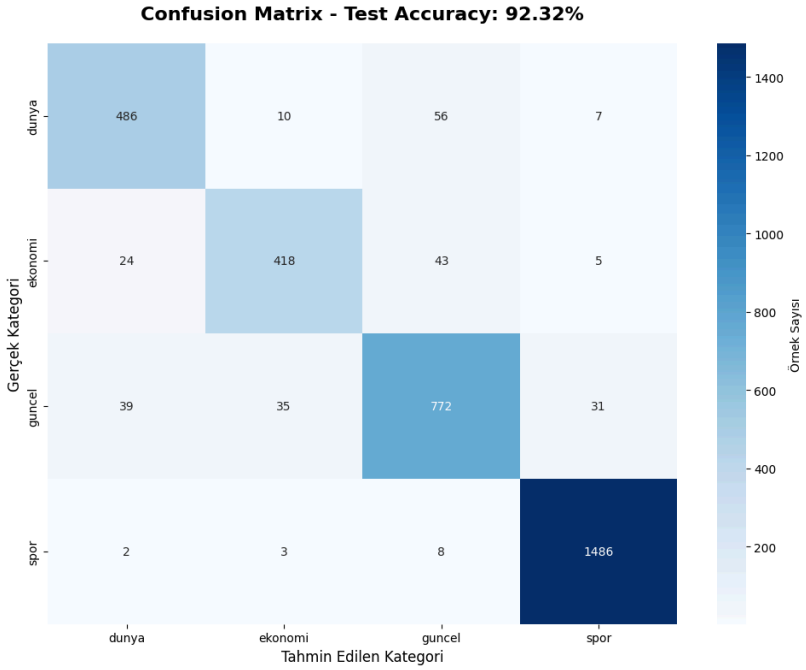
Geleneksel makine öğrenmesi yöntemleri (Naive Bayes, SVM), metin sınıflandırmada belirli bir başarı (%80-89) sağlasa da, kelimeler arasındaki anlamsal ilişkileri ve bağlam bilgisini yakalamada yetersiz kalmaktadır. Bu nedenle projede derin öğrenme tabanlı bir yaklaşım benimsenmiştir. Literatür taramasında, LSTM (Long Short-Term Memory) modellerinin uzun vadeli bağımlılıkları öğrenmede, CNN modellerinin ise metin içindeki yerel kalıpları yakalamada başarılı olduğu görülmüştür.

Bu çalışmada, her iki mimarinin avantajlarını birleştiren Hibrit LSTM + CNN modeli geliştirilmiştir. Model mimarisinde, metinler önce gömme (embedding) katmanından geçirilerek vektörel uzaya taşınmakta, ardından LSTM katmanı ile cümlenin zamansal ve anlamsal akışı analiz edilmektedir. Paralel veya seri olarak bağlanan CNN katmanları ise metin içindeki belirgin anahtar kelime öbeklerini tespit etmektedir. Bu yaklaşım, sadece kelime sıklığına bakan yöntemlerin aksine, haber metninin hem yapısal hem de anlamsal bütünlüğünü değerlendirmektedir. BERT gibi Transformer modelleri yüksek başarı sunsa da, yüksek hesaplama maliyetleri ve Türkçe için optimize edilmiş hafif modellerin azlığı nedeniyle, bu projede daha hızlı ve ölçeklenebilir olan hibrit mimari tercih edilmiştir.

4. Model Eğitimi ve Model Değerlendirilmesi

Model eğitimi, Google Colab ortamında GPU hızlandırıcısı kullanılarak gerçekleştirilmiştir. Eğitim sürecinde, sınıf dengesizliğinden kaynaklanabilecek sorunları minimize etmek için hesaplanan sınıf ağırlıkları kayıp fonksiyonuna entegre edilmiştir. Modelin aşırı öğrenmesini önlemek amacıyla "Early Stopping" mekanizması kullanılmış ve validation kaybının iyileşmediği noktalarda eğitim sonlandırılmıştır.

Test verisi üzerinde yapılan değerlendirmeler sonucunda modelin %92 üzerinde genel doğruluk oranına ulaştığı görülmüştür. Karmaşıklık matrisi incelendiğinde, modelin özellikle "Spor" ve "Ekonomi" haberlerini çok yüksek başarıyla ayırt ettiği, ancak içeriksel benzerlikler nedeniyle zaman zaman "Dünya" ve "Güncel" kategorileri arasında küçük karışıklıklar yaşadığı gözlemlenmiştir. Bu durum, haber metinlerinin doğası gereği dünya gündeminin güncel olaylarla iç içe olmasından kaynaklanmaktadır. Elde edilen sonuçlar, önerilen hibrit mimarinin Türkçe metin sınıflandırma görevinde literatürdeki benzer çalışmalara kıyasla rekabetçi ve uygulanabilir bir performans sunduğunu kanıtlamaktadır.



Şekil 1: Karmaşıklık Matrisi

Tablo 1: Sınıflandırma Performans Özeti

Sınıf	Precision	Recall	F1-Score	Örnek Sayısı
Spor	0.96	0.98	0.97	1500
Ekonomi	0.91	0.89	0.90	490
Dünya	0.88	0.85	0.86	558
Güncel	0.87	0.88	0.87	877
Accuracy			**0.92**	**3425**

5. KAYNAKÇA

[1] Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. EMNLP.

[2] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation.

[3] Kılınç, D., et al. (2017). TTC-3600: A new benchmark dataset for Turkish text categorization. Journal of Information Science.

[4] Toraman, C., et al. (2022). BERTurk: Turkish BERT. arXiv preprint.

[5] Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level Convolutional Networks for Text Classification. NIPS.