

# İSTANBUL TOPKAPI ÜNİVERSİTESİ

## MÜHENDİSLİK FAKÜLTESİ

### BİLGİSAYAR MÜHENDİSLİĞİ

**Proje Başlığı:** IMDB Film Yorumları Üzerine Duygu Analizi: Geleneksel Derin Öğrenme ve Transformer Mimarilerinin Karşılaştırılmalı Analizi

**Ad Soyad:** Umut Torun

**Öğrenci Numara:** 23040101063

**Öğrenci E-Posta:** umutturun@stu.topkapi.edu.tr

**GitHub:** [https://github.com/umutturun63/FET312\\_ImdbSentiment\\_Proje](https://github.com/umutturun63/FET312_ImdbSentiment_Proje)

**YouTube:** <https://www.youtube.com/watch?v=j3nQIM03Xfs>



# Problem Tanımı

Günümüzde internet kullanıcıları, ürün ve hizmetler hakkında görüşlerini çevrimiçi platformlarda paylaşmaktadır. Film endüstrisi özelinde, kullanıcı yorumları filmlerin başarısını tahmin etmede ve pazarlama stratejilerini belirlemede kritik rol oynamaktadır. Bu proje, IMDB platformundaki film yorumlarını analiz ederek, kullanıcı duygularını otomatik olarak pozitif veya negatif olarak sınıflandırmayı amaçlamaktadır.

**Temel Araştırma Sorusu:** "Doğal dil işleme teknikleri kullanılarak, film yorumlarından kullanıcı duyguları ne ölçüde doğru tespit edilebilir?"

**Görev Türü:** Bu proje, ikili sınıflandırma problemidir. Metin tabanlı veri üzerinde Doğal Dil İşleme (NLP) ve Duygu Analizi (Sentiment Analysis) uygulanmaktadır.



# Hedef Değişkenler

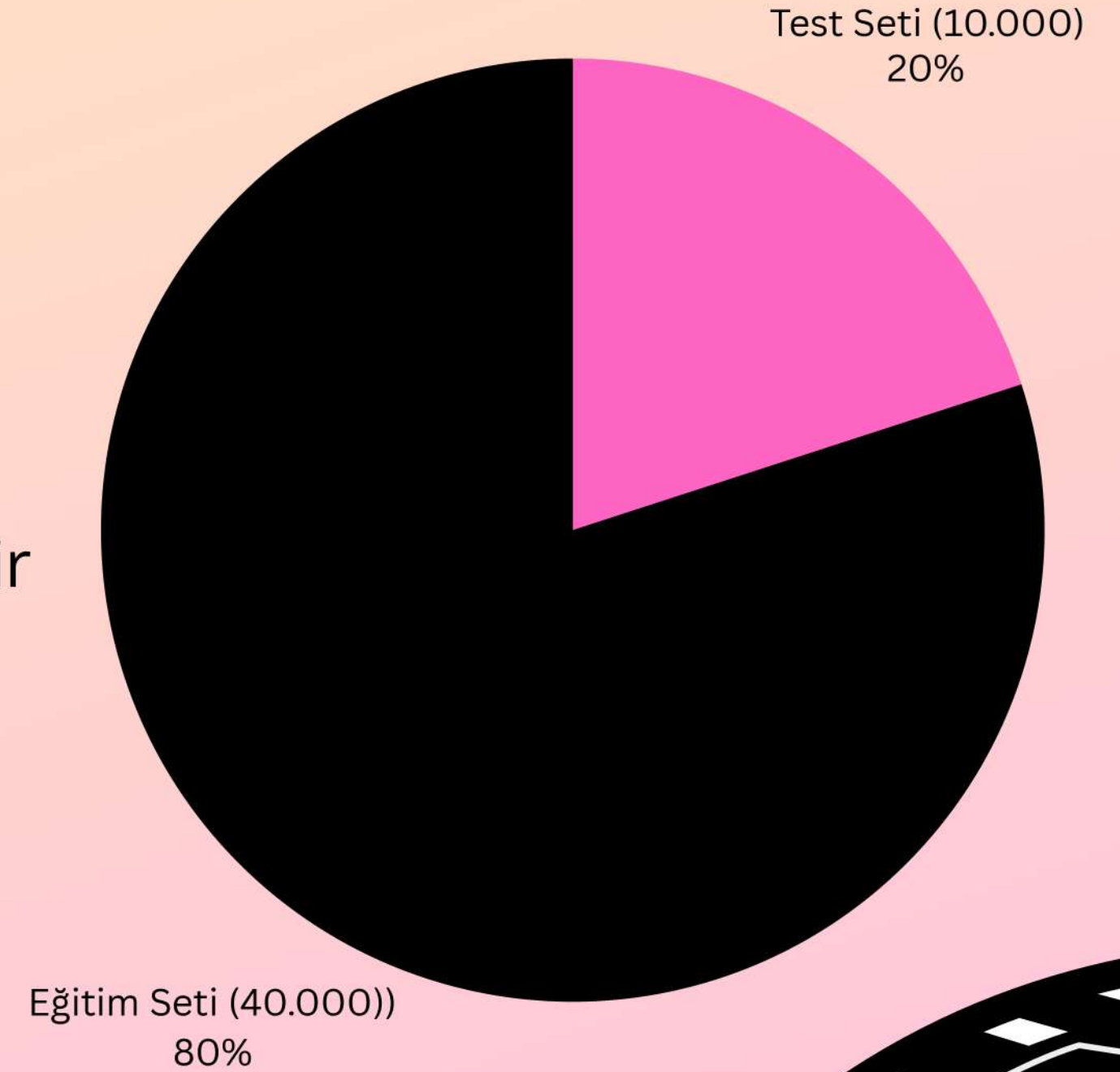
- Hedef Değişken: Sentiment (duygu etiketi)
- Değer Aralığı: Binary (0: Negatif, 1: Pozitif)
- Pozitif Sınıf: 1 (Pozitif yorumlar)
- Negatif Sınıf: 0 (Negatif yorumlar)
- Veri Dengesi: 50% pozitif, 50% negatif

## Başarı Kriterleri

Projenin başarısı aşağıdaki metriklerle değerlendirilmektedir

- **Doğruluk (Accuracy):**  $\geq 0.87$  (Hedef: %85 ve üzeri)
- **F1 Skoru:**  $\geq 0.87$  (Kesinlik ve duyarlılık dengesi)
- **Precision (Kesinlik):**  $\geq 0.88$
- **Recall (Duyarlılık):**  $\geq 0.88$

## Veri Seti Dağılımı





# Veri Ön İşleme

## 1) Klasik Modeller LG, LSTM, CNN

- **Temizlik:** HTML etiketleri, noktalama işaretleri ve sayılar temizlendi.
- **Dönüşüm:** Tüm metinler küçük harf çevrildi.
- **Stopwords:** İngilizce etkisiz kelimeler çıkarıldı.
- **Tokenization:** En sık geçen 20.000 kelime belirlendi, metinler sayısal dizilere dönüştürüldü ve Padding uygulandı.

## 2) Transformer Modeller BERT, XLNet

- **Raw Input:** Stopwords silinmedi.
- **Tokenizer:** Hugging Face BertTokenizer ve XLNetTokenizer kullanıldı.
- **Attention Mask:** Modelin padding alanlarını görmezden gelmesi için maskeleme yapıldı.



# Kullanılan Modeller

**Lojistik Regresyon:** Hızlı, yorumlanabilir, referans model.

**LSTM:** Metni sırayla işleyerek kelimeler arasındaki bağlamsal ilişkiyi hafızasında tutar

**CNN:** Filtreler kullanarak metindeki yan yana gelen önemli kelime gruplarını yakalar.

**BERT:** Cümleyi çift yönlü okuyarak kelimenin tam anlamını ve ironiyi çözer.

**XLNet:** Permütasyon tekniğiyle kelimeler arası en uzak ilişkileri bile yakalar.

# Model Karşılaştırılması



**XLNET: %90.56**



**BERT: %89.85**

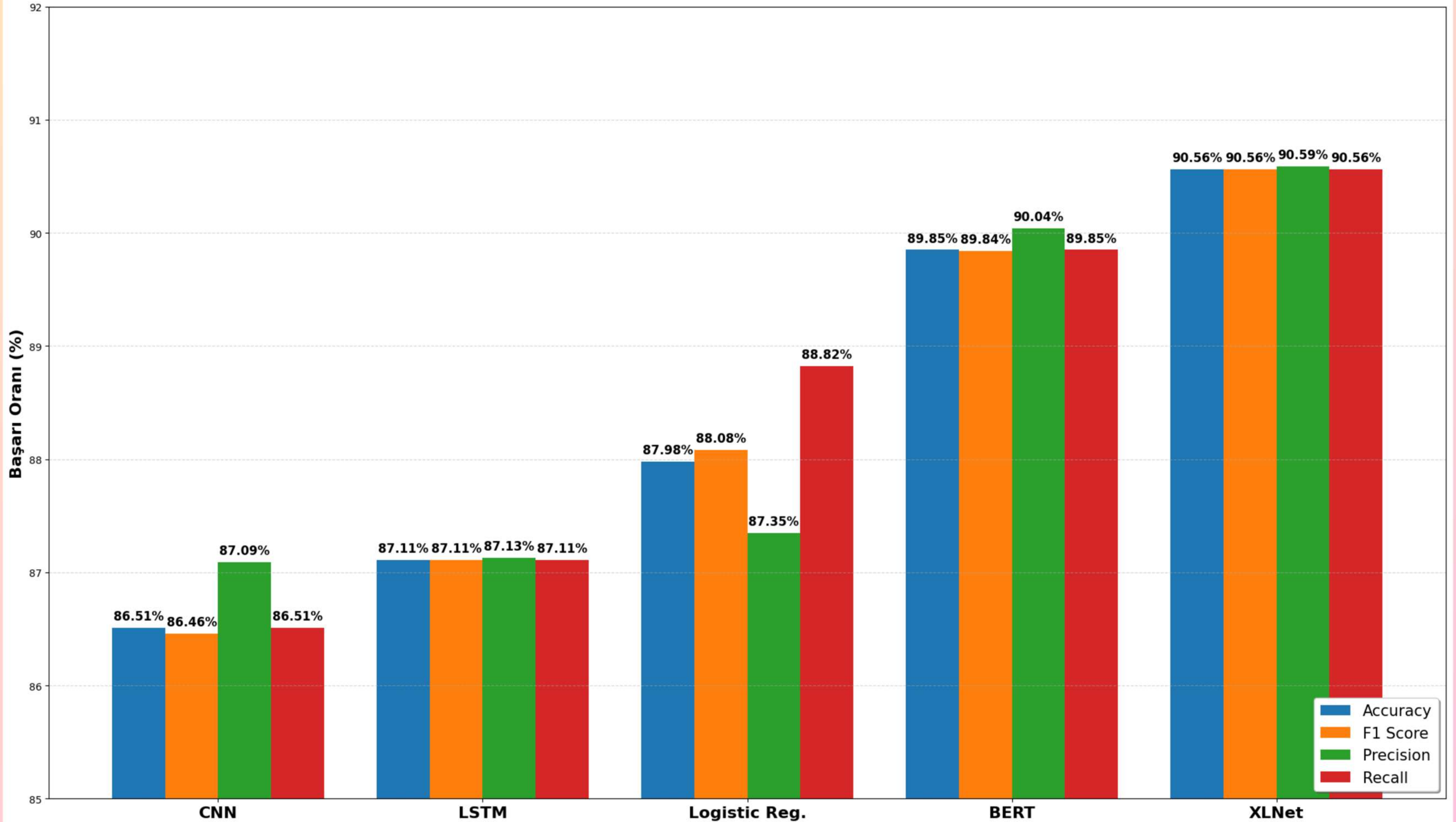


**LR: %87.98**

MODEL	ACCURACY	F1 SCORE	PRECISION	RECALL
XLNET	%90.5600	%90.5584	%90.5874	%90.5600
BERT	%89.8500	%89.8379	%90.0406	%89.8500
LR	%87.9800	%88.0800	%87.3500	%88.8200
LSTM	%87.1100	%87.1086	%87.1258	%87.1100
CNN	%86.5100	%86.4568	%87.0933	%86.5100



## Tüm Modellerin Performans Karşılaştırması

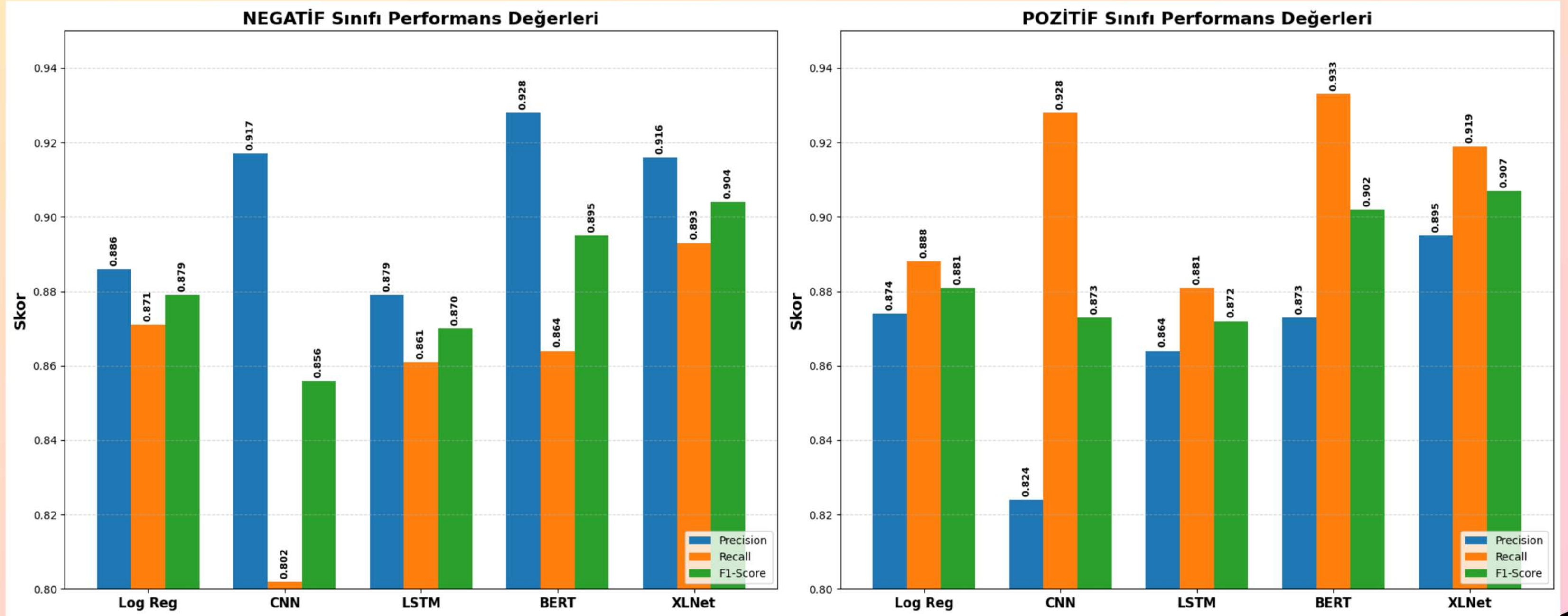


# Sınıf Bazlı Model Karşılaştırılması

SINIF	MODEL	PRECISION	RECALL	F1-SCORE
NEGATIVE	CNN	0.9174	0.8024	0.8561
NEGATIVE	LSTM	0.8789	0.8608	0.8698
NEGATIVE	BERT	0.9280	0.8640	0.8949
NEGATIVE	XLNet	0.9164	0.8926	0.9044
NEGATIVE	Log Reg	0.8863	0.8714	0.8788
POSITIVE	CNN	0.8244	0.9278	0.8731
POSITIVE	LSTM	0.8636	0.8814	0.8724
POSITIVE	BERT	0.8728	0.9330	0.9019
POSITIVE	XLNet	0.8953	0.9186	0.9068
POSITIVE	Log Reg	0.8735	0.8882	0.8808



# Sınıf Bazlı Model Karşılaştırması (Sütun Grafiği)

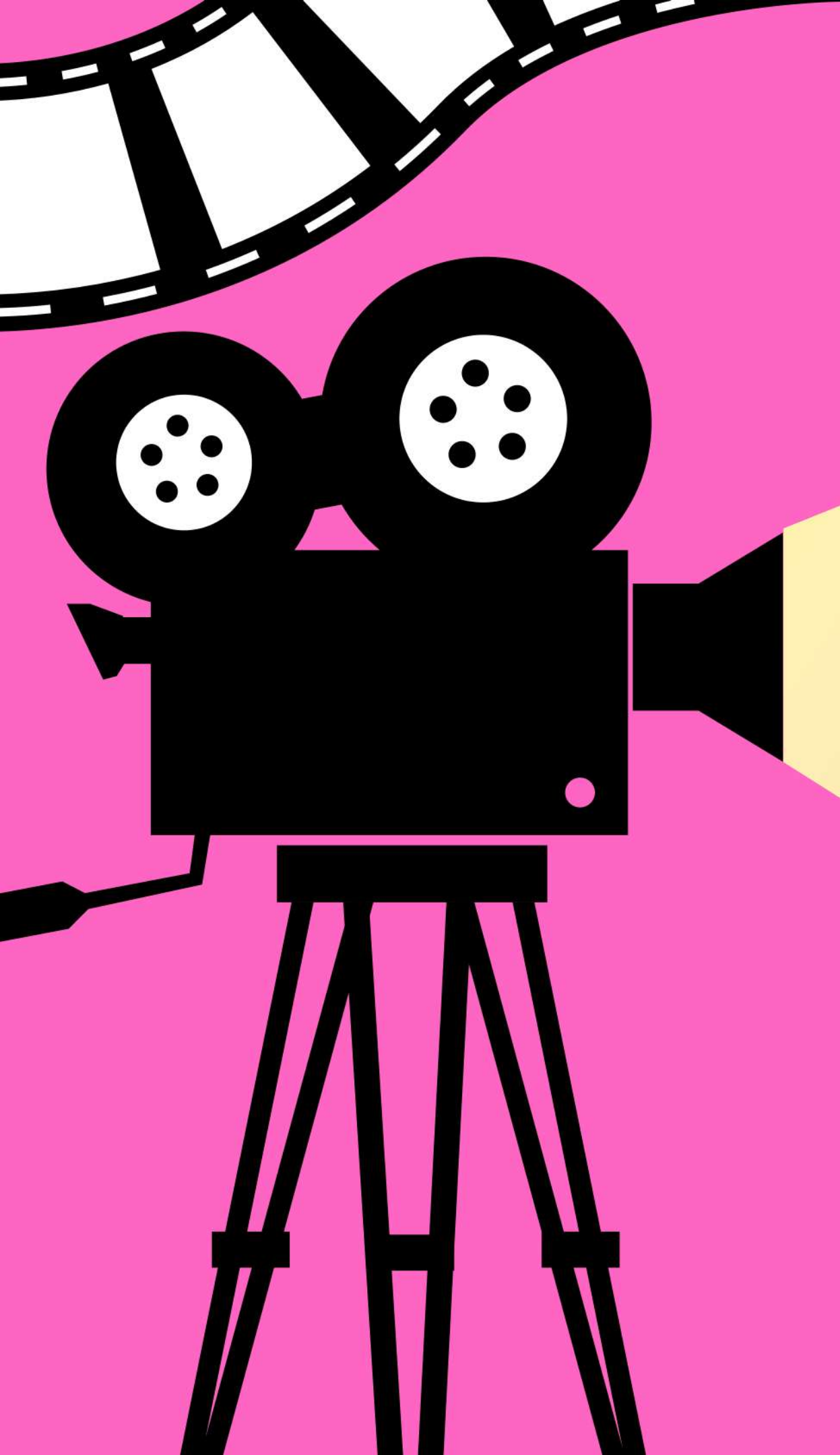




# Sonuç ve Değerlendirme

- **Hedef Aşıldı:** %87 olan başarı hedefi, XLNet ile %90.56 seviyesine çıkarıldı.
- **Transformer Üstünlüğü:** Ön eğitilmiş modellerin, sıfırdan eğitilen modellere göre dilin bağlamını ve ironiyi çok daha iyi anladığı kanıtlandı.
- **Baseline Başarısı:** Lojistik Regresyon, basit yapısına rağmen klasik derin öğrenme modelleriyle (LSTM/CNN) yarışan bir performans sergiledi.
- **Öneri:** En yüksek doğruluk için XLNet, en düşük kaynak kullanımı için Lojistik Regresyon önerilmektedir.





***Teşekkürler!***