İSTANBUL TOPKAPI ÜNİVERSİTESİ
MÜHENDİSLİK FAKÜLTESİ
BİLGİSAYAR MÜHENDİSLİĞİ

Ders/Dönem: FET312 - Derin Öğrenme / 2025-2026 Güz Dönemi

Proje Başlığı: IMDB Film Yorumları Üzerine Duygu Analizi: Makine Öğrenmesi ve Derin Öğrenme Yaklaşımlarının Karşılaştırılmalı Analizi

Ekip Adı: Dev312

Ekip Üyeleri;

- **Ad Soyad:** Umut Torun
- **Öğrenci Numarası:** 23040101063
- **Öğrenci E-Posta:** umuttorun@stu.topkapi.edu.tr
- **Öğrenci İmza:**

GitHub Repo Bağlantısı:

https://github.com/umuttorun63/FET312_ImdbSentiment_Proje

PROBLEM TANIMI & MOTİVASYON

İş/Bilimsel Soru

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): ≥ 0.87 (Hedef: %85 ve üzeri)
- F1 Skoru: ≥ 0.87 (Kesinlik ve duyarlılık dengesi)
- Precision (Kesinlik): ≥ 0.88
- Recall (Duyarlılık): ≥ 0.88

PROJE YÖNETİMİ

Proje Zaman Çizelgesi:

1-4. Hafta: Derin Öğrenme hakkında derste derin öğrenmenin temel kavramlarını öğrenme.

5-6. Hafta: Projede kullanılacak olan Python dili ve temel derin öğrenme

kütüphaneleri hakkında bilgi edinildi ve gerekli yazılım ortamları kuruldu.

7. Hafta: Proje konusu araştırıldı ve seçildi, Projeye uygun veri seti araştırıldı ve incelendi. Projenin base model tasarımı ve geliştirilmesi.

8. Hafta: Proje raporunun son hali.

11. Hafta: Projede kullanılacak olan yeni modellerin ve kütüphanelerin araştırılması.

12. Hafta: Proje de kullanılacak yeni modellerin tasarımı ve geliştirilmesi.

13. Hafta: Proje kod, rapor ve sunumu son hali.

İlgili Çalışmalar

A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis,"in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011, pp. 142-150.

J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proceedings of NAACL- HLT, 2019, pp. 4171-4186.

Projenin Katkısı ve Farklılaşma

Bu proje, aşağıdaki noktalarda farklılaşmaktadır:

Karşılaştırmalı Yaklaşım: Klasik makine öğrenmesi (Logistic Regression) ile derin öğrenme LSTM ve CNN yöntemlerinin sistematik karşılaştırması yapılmaktadır.

Hafif ve Yorumlanabilir Baseline: Logistic Regression modeli, derin öğrenme modellerine kıyasla daha az hesaplama gücü gerektirmekte ve feature importance analizi yapılabilmektedir.

Detaylı Ön İşleme Pipeline'ı: HTML tag temizleme, stopword filtreleme, custom vocabulary oluşturma gibi adımlar manuel olarak implementasyonu gerçekleştirilmiştir.

Eğitim Amaçlı: Proje, derin öğrenme kavramlarını anlamak ve uygulamak için adım adım yaklaşım sunmaktadır.

VERİ AÇIKLAMASI VE YÖNETİMİ

Veri Kümesi Açıklaması

Veri Seti Adı: IMDB Dataset of 50K Movie Reviews

Kaynak: Stanford University - Andrew Maas et al. (2011)

Bağlantı: https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews

Lisans: Public Domain - Akademik ve ticari kullanım için ücretsiz

Açıklama: IMDB veri seti, 50,000 film yorumundan oluşan, duygu analizi için en yaygın kullanılan benchmark veri setlerinden biridir. Veri seti, dengeli bir dağılıma sahiptir (25,000 pozitif, 25,000 negatif yorum).

Veri Şeması

Özellik	Tür	Açıklama	Örnek Değer
Review	String (Text)	Film yorumu metni (HTML içerebilir)	"This movie was fantastic..."
Sentiment	Categorical (Binary)	Pozitif veya negatif etiket	"positive" / "negative"
Label	Integer (0/1)	Sayısal etiket (hedef değişken)	0 (negatif), 1 (pozitif)
cleaned_review	String (Text)	Temizlenmiş metin (ön işleme sonrası)	"this movie was fantastic"

Boyut

Toplam Örnek Sayısı: 50,000 yorum

Sütun Sayısı: 2 (review, sentiment)

Etik, Gizlilik ve Önyargı

Veri seti, kamuya açık IMDB platformundan toplanmış ve kullanıcıların halka açık yorumlarını içermektedir. Kullanıcı isimleri ve kişisel bilgiler veri setinde yer almamaktadır.Veri setinde kişisel tanımlayıcı bilgi (PII) bulunmamaktadır. Yorumlar, kullanıcı kimlikleri olmadan toplanmıştır.

YÖNTEMLER VE MİMARİ

Bu proje de klasik makine öğrenmesi ile baseline model oluşturulmuştur: Logistic Regression algoritması kullanılmış ve hızlı ve yorumlanabilir baseline performans elde etme amaçlanmıştır. Veri setine clean_text fonksiyonu uygulanarak html etiketleri, noktalama işaretleri, sayılar çıkarıldı. Fazla boşluklar tek boşluğa düşürüldü ve tüm metinler küçük harfe dönüştürüldü. Vocabulary fonksiyonu ile stopwords listesindeki kelimeler harici en sık kullanılan 20000 kelime listelendi. Tokenization fonksiyonu ile metinler sayı dizilerine çevrildi. Padding fonksiyonu ile yorum sayı dizileri aynı uzunluğa getirildi. Vectorization fonksiyonu her bir yorumu sabit uzunlukta bir vektöre çevirir ve bir kelime o yorumda geçiyorsa 1 veya geçmiyorsa 0 olarak işaretler. Hyper parametreler denendi en hızlı ve verimli run edecek şekilde ayarlandı. Projenin devamında ise PyTorch kütüphanesi kullanılarak iki temel mimari tasarlanmıştır: metindeki sıralı ve bağlamsal ilişkileri öğrenmek için 100 boyutlu Embedding katmanı, %30 Dropout ve 256 gizli birime sahip 2 katmanlı LSTM modeli ve metindeki yerel n-gram özelliklerini yakalamak için 3, 4 ve 5 boyutlu kernel filtrelerine sahip 1 boyutlu konvolüsyon katmanları içeren CNN modeli. Her iki model de ikili sınıflandırma problemine uygun olarak Sigmoid aktivasyon fonksiyonu, Binary Cross Entropy kayıp fonksiyonu ve Adam optimizasyon algoritması ile eğitilmiş olup, modellerin performansı Lojistik Regresyon tabanlı temel baseline model ile Accuracy, F1 Skoru, Precision ve Recall metrikleri üzerinden karşılaştırmalı olarak değerlendirilmiştir.

DENEY TASARIMI

Ana Amaç

Bu çalışmanın temel amacı, doğal dil işleme (NLP) alanında metin sınıflandırma problemi için geliştirilen farklı mimarilerin performanslarını karşılaştırmalı olarak analiz etmektir. Deneyin merkezinde, ardışık veri işleme yeteneğine sahip LSTM ve yerel özellikleri çıkarma yeteneğine sahip CNN derin öğrenme modellerinin, geleneksel bir makine öğrenmesi yöntemi olan Lojistik Regresyon ile kıyaslanması yer almaktadır. Deneyin ana odağı, kelime frekanslarına dayalı basit yöntemler ile kelime sırasını ve bağlamsal ilişkileri öğrenebilen karmaşık sinir ağı yapılarının, pozitif ve negatif yorumları ayırt etme başarısını sistematik bir şekilde kıyaslamaktır.

Değerlendirme kriterleri

Proje kapsamında toplam 3 farklı model Lojistik Regresyon, LSTM ve CNN geliştirilmiştir. Seçilen temel değerlendirme kriteri olan Accuracy oranına göre yapılan başarı sıralamasında; LSTM modeli %87.98 doğruluk oranı ve metindeki bağlamsal ilişkileri öğrenme yeteneğiyle birinci sırada yer almış, baseline model olarak kullanılan Lojistik Regresyon aynı doğruluk oranıyla %87.98 ikinci sırada yer almış, CNN modeli ise %87.41 doğruluk oranıyla üçüncü sırada yer almıştır.

KULLANILAN ARAÇLAR VE FRAMEWORKERLAR

Projenin geliştirilmesinde Python 3.8+ programlama dili temel alınmıştır. Veri setinin yüklenmesi ve manipülasyonu için Pandas (1.3+), sayısal matris işlemleri ve vektörizasyon için NumPy (1.20+) kütüphaneleri kullanılmıştır. Çalışma kapsamına eklenen derin öğrenme mimarilerini desteklemek amacıyla, LSTM ve CNN modellerinin tasarımı ve eğitimi için PyTorch kütüphanesi (torch, torch.nn) projeye entegre edilmiştir. Scikit-learn (1.0+), hem temel baseline model olan Lojistik Regresyonun eğitimi hem de veri setinin eğitim-test olarak ayrılması ve performans metriklerinin hesaplanması süreçlerinde kullanılmıştır. Metin ön işleme aşamasında NLTK (3.6+) ile stopwords filtrelemesi sağlanırken, elde edilen deneysel sonuçların ve karışıklık matrislerinin görselleştirilmesi Matplotlib ve Seaborn kütüphaneleri ile gerçekleştirilmiştir.

KAYNAKLAR

[1] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proc. 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, USA, Jun. 2011, pp. 142–150. [Online]. Available: http://ai.stanford.edu/~amaas/data/sentiment/

[2] "IMDB Dataset of 50K Movie Reviews," *Kaggle*. [Online]. Available: https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews. [Erişim: 17 Kasım 2025].

[3] GitHub, "IMDb Sentiment Analysis," *GitHub Topics*. [Online]. Available: https://github.com/topics/imdb-sentiment-analysis. [Erişim: 17 Kasım 2025].

[4] A. Mohan, "Sentiment Analysis using LSTM Pytorch," *Kaggle*, [Online]. Available: https://www.kaggle.com/code/arunmohan003/sentiment-analysis-using-lstm-pytorch. [Erişim Tarihi: 21 Aralık 2025].

[5] Ducanger, "IMDB BERT - CNN - LSTM 0.93 ACC," *Kaggle*, [Online]. Available: https://www.kaggle.com/code/ducanger/imdb-bert-cnn-lstm-0-93-acc. [Erişim Tarihi: 21 Aralık 2025].

[6] Google, *Gemini* (Büyük dil modeli). [Yazılım]. [Online]. Available: https://gemini.google.com/ [Erişim: 17 Kasım 2025].

[7] OpenAI, *ChatGPT* (Büyük dil modeli). [Yazılım]. [Online]. Available: https://chat.openai.com/ [Erişim: 17 Kasım 2025].