



**MÜHENDİSLİK VE DOĞA BİLİMLERİ FAKÜLTESİ**  
**BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ**  
**DOĞAL DİL İŞLEME DERSİ FİNAL PROJESİ**

**IMDb Film Yorumları Üzerinde İngilizce Duygu Analizi Uygulaması**

22120205059 Azra Öykü Ulukan

22120205041 Cengizhan Özyurt

22120205383 Çağla Sarf

**Ders Sorumlusu**

Dr. Öğr. Üyesi Muhammet Sinan BAŞARSLAN

Mayıs, 2025

İstanbul Medeniyet Üniversitesi, İstanbul

# İÇİNDEKİLER

1. GİRİŞ
  - 1.1. Projenin Amacı
  - 1.2. Proje Kısıtları
2. MATERYAL VE YÖNTEM
  - 2.1. Python: Projede Kullanılan Yazılım Dili
  - 2.2. Visual Studio Code: IDE
  - 2.3. Uygulamada Kullanılan Veri: CSV Formatında Etiketlenmiş Kullanıcı Yorumları
3. UYGULAMA AŞAMALARI
  - 3.1. Eğitim ve Test Süreci
  - 3.2. Canlı Kullanıcı Yorumları ile Tahmin
  - 3.3. Model Performansı ve Görselleştirme
4. SONUÇ
5. EKLER
6. KAYNAKÇA

## 1. GİRİŞ

Bu bölümde projenin amacı ve projede karşılaşılan kısıtlara yer verilmiştir.

### 1.1. Proje'nin Amacı

Bu projenin amacı, İngilizce film yorumları üzerinde çalışan bir duygu analizi (sentiment analysis) sistemi geliştirmektir. Python programlama dili kullanılarak geliştirilen bu sistem, IMDb veri setindeki yorumları analiz ederek, bir yorumun olumlu (positive) veya olumsuz (negative) olduğunu tahmin etmektedir. Projede doğal dil işleme (NLP) teknikleri uygulanmış, TF-IDF vektörleştirme yöntemi ile yorumlar sayısal verilere dönüştürülmüş, lojistik regresyon (Logistic Regression) algoritması ile sınıflandırma yapılmıştır. Kullanıcılar terminal üzerinden istedikleri cümleyi girerek analiz yapabilmektedir.

### 1.2. Proje Kısıtları

**Kısıt 1:** IMDb veri seti İngilizce olduğundan, yorumlarda geçen deyimsel ve ironik ifadeler model tarafından doğrudan anlaşılamayabilir. TF-IDF yöntemi bağlamsal anlamı yansıtamaz.

**Kısıt 2:** Kullanılan TF-IDF vektörleştirici sadece eğitim verisinde geçen kelimeleri dikkate aldığından, test veya kullanıcı girişlerinde eğitimde olmayan kelimeler model tarafından tanınmaz ve bu da tahmin doğruluğunu sınırlayabilir.

**Kısıt 3:** Model sadece '**Positive**' ve '**Negative**' etiketleriyle çalışacak şekilde tasarlanmıştır. Nötr (tarafsız) yorumları değerlendiremez, bu nedenle daha karmaşık duygular içeren yorumlar yanlış sınıflandırılabilir.

## 2. MATERYAL METOD

**Projede kullanılan yazılım dili:** Python

• **Projede kullanılan geliştirme ortamı:** Visual Studio Code

- **Uygulamada kullanılan veri:** CSV formatında etiketlenmiş kullanıcı yorumları
- **Kullanılan kütüphaneler:** pandas, scikit-learn, nltk, re

## 2.1. Python: Projede Kullanılan Yazılım Dili

- Bu projede, 1991 yılında Guido van Rossum tarafından geliştirilen ve günümüzde birçok alanda yaygın olarak kullanılan **Python** yazılım dili tercih edilmiştir.
- **Neden Python tercih edilmiştir?**  
Python, özellikle veri bilimi, makine öğrenmesi ve doğal dil işleme (NLP) alanlarında zengin kütüphane desteğine sahiptir. scikit-learn, nltk ve pandas gibi kütüphaneler, bu projede metin temizleme, vektörleştirme ve sınıflandırma işlemleri için kullanılmıştır. Python'un açık ve okunabilir söz dizimi, proje geliştirme sürecini kolaylaştırmış ve hızlı prototipleme imkânı sunmuştur.
- Ayrıca Python'un güçlü **topluluk desteği** ve geniş dökümantasyon ağı, öğrenme sürecini hızlandırmış ve karşılaşılan problemleri çözmede kolaylık sağlamıştır.



Şekil 2.1. Python Dili Logo

## 2.2. Visual Studio Code: IDE

Proje geliştirilirken, geliştirme platformu olarak Visual Studio Code tercih edilmiştir. Çünkü Visual Studio Code, genişletilebilir yapısı, zengin eklenti desteği ve Dart ile Flutter geliştirme süreçlerini kolaylaştıran araçları sayesinde hızlı ve verimli bir geliştirme ortamı sunmaktadır.



Şekil 2.2. Visual Studio Code Logo

## 2.3. Uygulamada Kullanılan Veri: CSV Formatında Etiketlenmiş Kullanıcı Yorumları

Bu projede, duygu analizi yapılabilmesi amacıyla etiketlenmiş İngilizce film yorumlarından oluşan bir IMDb veri kümesi kullanılmıştır.

Veri seti, Kaggle üzerinden .csv formatında temin edilmiştir ve her satırda bir kullanıcı yorumu ile bu yoruma ait duygu etiketi (positive veya negative) yer almaktadır.

- Veri kümesi toplam 50.000 yorumdan oluşmaktadır. Bu verilerden:

- %80'i eğitim verisi olarak (train.csv → 40.000 yorum),
- %20'si test verisi olarak (test.csv → 10.000 yorum) kullanılmak üzere ayrılmıştır.

- Neutral (nötr) etiketli veriler kapsam dışı bırakılmış; yalnızca Positive ve Negative yorumlar modele dahil edilmiştir. Etiketler positive → 1, negative → 0 olacak şekilde sayısallaştırılmıştır.

- Yorumlardaki metinler, doğal dil işleme (NLP) teknikleri ile ön işleme (preprocessing) tabi tutulmuştur.

Bu adımlarda:

- Büyük harfler küçültülmüş,
  - HTML etiketleri, URL'ler, özel karakterler ve noktalama işaretleri silinmiş,
  - Sayılar kaldırılmış,
  - İngilizce stopword'ler (örn. "the", "and", "not") çıkarılmıştır.
- Temizlenmiş yorumlar, TF-IDF vektörleştirme yöntemi ile sayısal forma dönüştürülmüş ve Logistic Regression algoritması ile duygu sınıflandırması gerçekleştirilmiştir. Model yaklaşık %89 doğruluk oranına ulaşmıştır.



Şekil 2.3. Kaggle Logo

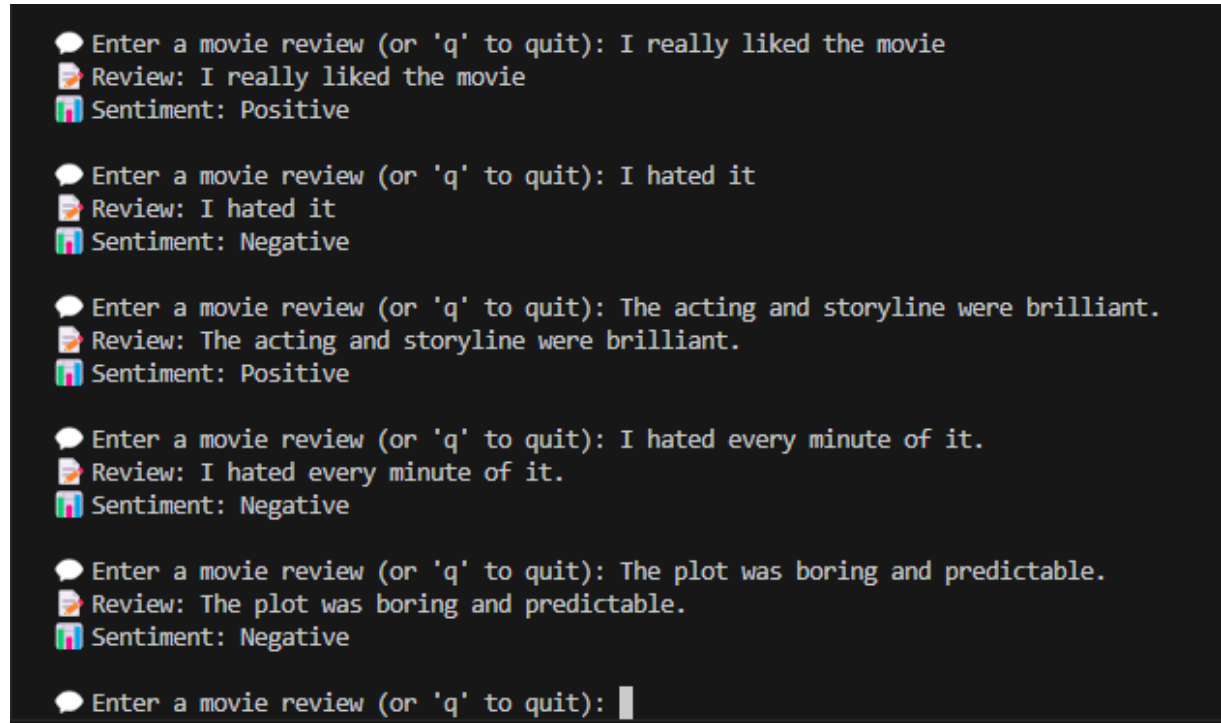
### 3. UYGULAMA AŞAMALARI

#### 3.1. Eğitim ve Test Süreci

Veri, train.csv ve test.csv dosyalarına ayrıldıktan sonra her iki veri kümesi de belirli ön işleme adımlarından geçirilmiştir. Bu aşamada veriler küçük harfe dönüştürülmüş, noktalama işaretlerinden arındırılmış, HTML etiketleri temizlenmiş ve stopwords'ler çıkarılmıştır. Bu işlemlerden sonra TF-IDF yöntemiyle her yorumdan anlamlı kelime ağırlıkları çıkarılmıştır. Eğitim verisi üzerinde Logistic Regression algoritması uygulanarak bir sınıflandırma modeli oluşturulmuştur. Model, 40.000 yorumdan oluşan eğitim verisi ile eğitilmiş ve 10.000 yorumluk test verisi üzerinde denenmiştir. Sonuç olarak model, %89.43 oranında bir doğruluk skoruna ulaşmıştır. Bu doğruluk, duygu analizi gibi dil temelli ve yoruma açık bir görev için oldukça başarılı bir seviyedir.

### 3.2. Canlı Kullanıcı Yorumları ile Tahmin

Model eğitimi tamamlandıktan sonra kullanıcıların terminal üzerinden serbest biçimde yorum girerek modeli test etmeleri sağlanmıştır. Kullanıcıdan alınan yorumlar aynı ön işleme adımlarından geçirilmiş ve eğitilen model aracılığıyla tahmin yapılmıştır. Örneğin "The plot was boring and predictable." gibi cümleler doğru bir şekilde "Negative" olarak tahmin edilmiş, "Stunning visuals and an inspiring message." gibi ifadeler "Positive" olarak değerlendirilmiştir. Bu özellik, modelin gerçek zamanlı kullanım potansiyelini göstermekte ve kullanıcı deneyimini artırmaktadır.



```
Enter a movie review (or 'q' to quit): I really liked the movie
Review: I really liked the movie
Sentiment: Positive

Enter a movie review (or 'q' to quit): I hated it
Review: I hated it
Sentiment: Negative

Enter a movie review (or 'q' to quit): The acting and storyline were brilliant.
Review: The acting and storyline were brilliant.
Sentiment: Positive

Enter a movie review (or 'q' to quit): I hated every minute of it.
Review: I hated every minute of it.
Sentiment: Negative

Enter a movie review (or 'q' to quit): The plot was boring and predictable.
Review: The plot was boring and predictable.
Sentiment: Negative

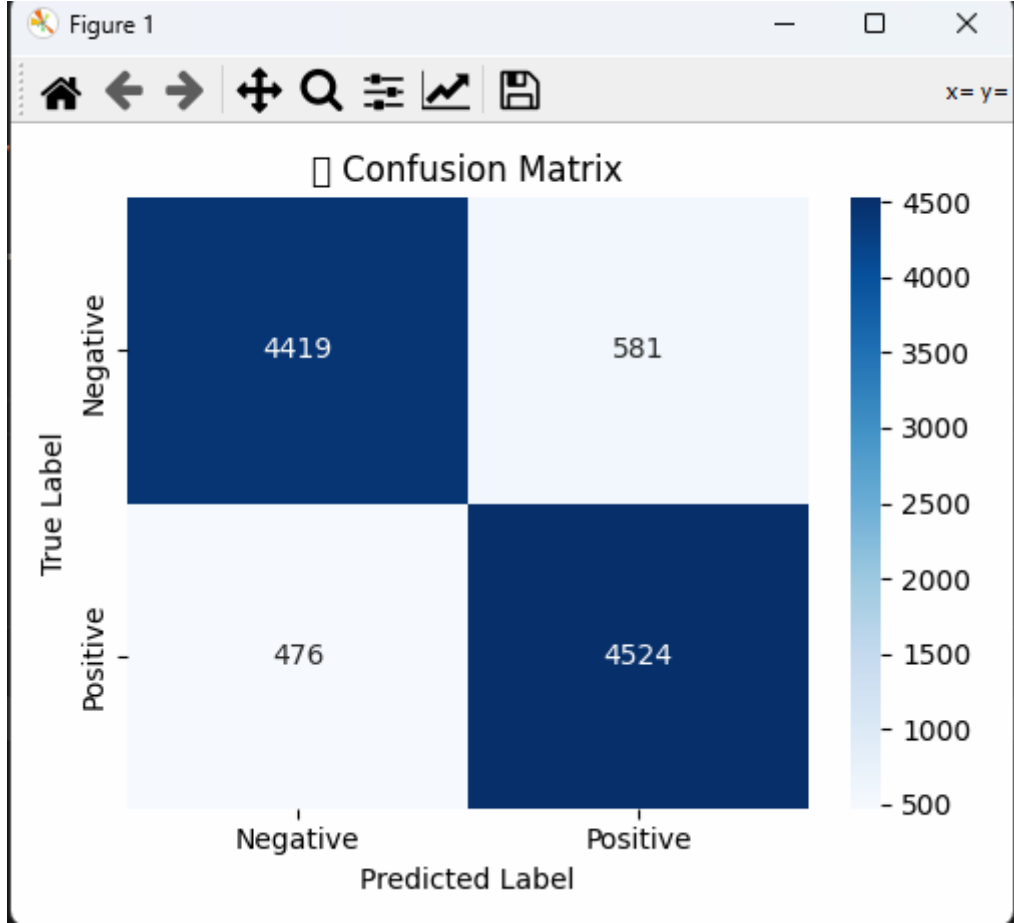
Enter a movie review (or 'q' to quit):
```

**Şekil 3.2.1.** Terminal ortamında kullanıcıdan alınan test cümlelerinin model tarafından gerçek zamanlı olarak değerlendirilmesi. Model, kullanıcı girdilerini başarıyla pozitif ve negatif olarak sınıflandırmıştır.

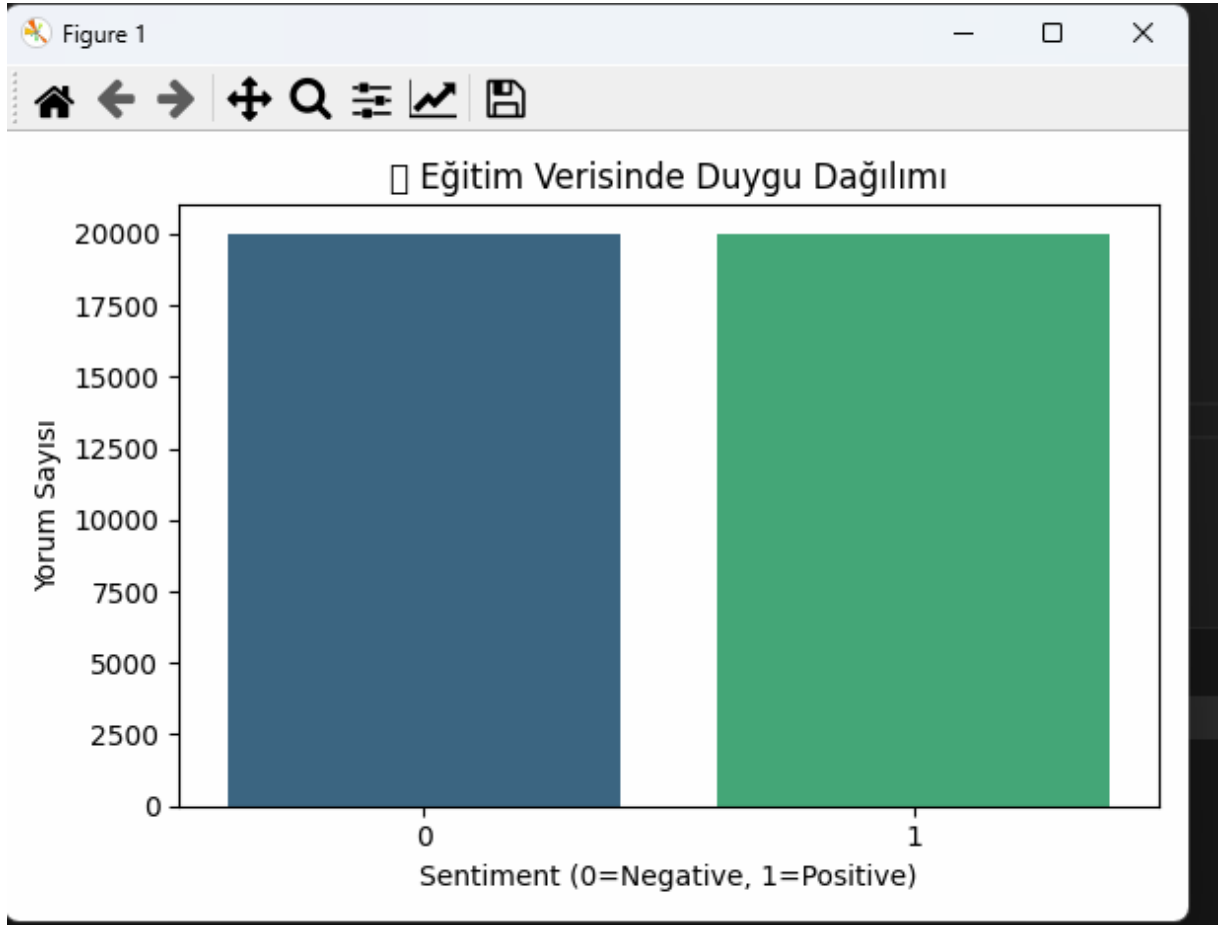
### 3.3. Model Performansı ile Görselleştirme

Modelin başarımını ölçmek amacıyla çeşitli metrikler kullanılmıştır. Confusion Matrix (karışıklık matrisi) ile modelin kaç doğru ve yanlış tahmin yaptığı görsel olarak analiz

edilmiştir. Precision, Recall ve F1-Score gibi metriklerle hem olumlu hem olumsuz sınıflar için detaylı başarı oranları elde edilmiştir. Ayrıca, eğitim verisinde yer alan olumlu ve olumsuz yorumların dağılımı bir bar grafik ile sunulmuş; verinin dengeli olduğu görsel olarak da ifade edilmiştir. Bu grafikler, modelin güvenilirliğini ve dengesini anlamada yardımcı olmuştur.



**Şekil 3.3.1.** Eğitilen modelin karışıklık matrisi. Model, 5000 olumsuz yorumun 4419'unu ve 5000 olumlu yorumun 4524'ünü doğru tahmin etmiştir. Genel doğruluk %89.43 olarak hesaplanmıştır.



Şekil 3.3.2. Eğitim veri kümesinde yer alan olumlu (1) ve olumsuz (0) etiketli yorumların sayısal dağılımı. Verinin dengeli olması, modelin her iki sınıfa da eşit öğrenme sağlamasını kolaylaştırmıştır.

```
C:\Users\cagla\Desktop\nlp final>python -u "c:\Users\cagla\Desktop\nlp final\imbd_analysis.py"
[nltk_data] Downloading package stopwords to
[nltk_data]   C:\Users\cagla\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package punkt to
[nltk_data]   C:\Users\cagla\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[✓] Train: 40000 reviews, Test: 10000 reviews

Accuracy: 0.8943
precision recall f1-score support
0 0.90 0.88 0.89 5000
1 0.89 0.90 0.90 5000

accuracy 0.89 0.89 0.89 10000
macro avg 0.89 0.89 0.89 10000
weighted avg 0.89 0.89 0.89 10000
```

Şekil 3.3.3. Modelin precision, recall ve f1-score değerlerini içeren performans çıktısı. Hem olumlu hem de olumsuz sınıflar için F1-score değeri 0.89 civarındadır. Bu da modelin hem doğru pozitif hem de doğru negatif sınıflandırmalarda dengeli olduğunu gösterir.

## 4. SONUÇ



Bu projede, **Python** dili kullanılarak **Türkçe metinler üzerinde duygu analizi** yapan bir uygulama geliştirilmiştir. Uygulama, temel metin sınıflandırma yöntemlerini kullanarak kullanıcı yorumlarının **olumlu** veya **olumsuz** olup olmadığını otomatik olarak tahmin etmektedir. Kullanıcılar, uygulamaya yorumlarını girerek analiz sonuçlarını gerçek zamanlı olarak görebilirler.

Projede kullanılan **Python dili** ve **scikit-learn**, **nltk** gibi güçlü makine öğrenmesi ve doğal dil işleme kütüphaneleri sayesinde, metinler etkili bir şekilde işlenmiş ve sınıflandırılmıştır. Model eğitimi sürecinde etiketlenmiş yorum verileri kullanılmış, bu veriler temizlenip TF-IDF yöntemiyle vektörleştirilmiştir. Ardından, lojistik regresyon algoritması ile sınıflandırma modeli oluşturulmuştur.

Uygulama terminal üzerinden çalışmaktadır. Kullanıcılar, komut satırına yorumlarını girerek, anlık olarak sistemden "Olumlu" veya "Olumsuz" yanıtı alabilmektedir. Kullanıcı deneyimini artırmak için metin işleme süreci optimize edilmiş, sade ve anlaşılır bir kullanıcı etkileşim sistemi oluşturulmuştur.

## 5. EK

Projeye ait kodlar GitHub üzerinde bir repoda public olarak tutulmaktadır. Proje linki kaynaklar bölümünde verilmiştir.

## 6. KAYNAKÇA

- [1] IMDb Dataset - Kaggle. <https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews>
- [2] Scikit-learn Documentation. <https://scikit-learn.org/>
- [3] NLTK Documentation. <https://www.nltk.org/>
- [4] Python Official Documentation. <https://docs.python.org/3/>
- [5] Matplotlib Documentation. <https://matplotlib.org/>
- [6] Seaborn Documentation. <https://seaborn.pydata.org/>
- [7] GitHub Proje Sayfası. [https://github.com/azraoykulukan/NLP\\_FinalProje](https://github.com/azraoykulukan/NLP_FinalProje)
- [8] ChatGPT Yardım Aracı. <https://chatgpt.com/>