Proje Ana Alanı : Yazılım

Proje Tematik Alanı : Yapay Zeka

Proje Adı (Başlığı) : Müzik Eserlerinin Nesnel Skorlandırılması: Ses ve Şarkı Sözü Verilerinin Bütünleşik Yapay Zeka Analizi

Özet

Günümüzde film ve oyun endüstrilerinde yaygın olan standart puanlama sistemlerinin müzik alanındaki eksikliğinden yola çıkan bu proje, müzik eserlerini yapay zeka destekli bütünleşik bir analizle nesnel olarak skorlamayı amaçlamaktadır. Geliştirilen sistem, Python tabanlı bir iş akışı (pipeline) üzerinde ses ve şarkı sözü verilerini modüler olarak işlemektedir. Metodoloji kapsamında öncelikle Demucs modeli ile ayrıştırılan vokal verisi, Whisper modeli aracılığıyla metne dönüştürülmüştür. Elde edilen metin üzerinde fonksiyonel yöntemlerle kafiye ve redif tespiti yapılırken; DistilBERT ile anlamsal derinlik, Lojistik Regresyon ile içerik toksisitesi analiz edilmiştir. Ses analizi tarafında ise MFCC ve spektral özellikler çıkarılarak Gradient Boosting algoritmasıyla akustiklik, enerji ve dans edilebilirlik parametreleri tahmin edilmiştir.

Elde edilen tüm sayısal veriler; müzik türüne (Pop, Rock, Klasik vb.) özgü belirlenen katsayılarla ağırlıklandırılarak, 0-100 aralığında nihai bir kalite skoruna dönüştürülmüştür. Eğitilen modellerin sonuçları MAE ve F1 skoru metrikleri ile ölçülmüş olup modeller yüksek başarı oranları sergilemiştir. Bu çalışma; müziği yalnızca işitsel bir sinyal olarak değil, anlamsal ve teknik içeriğiyle birlikte ele alan, insan algısıyla uyumlu ve tekrarlanabilir bir skorlama modeli sunmaktadır.

Anahtar kelimeler: Müzik, Yapay Zeka, Dil İşleme

Amaç

Günümüzde üretilen medya içerikleri üretildiği andan itibaren incelemeye ve değerlendirmeye alınır. Filmler, diziler ve bilgisayar oyunları hem eleştirmenler hem de kullanıcılar tarafından estetik, içerik, teknik yeterlilik ve toplumsal etki açısından sürekli olarak analiz edilmektedir. Bu sektörlerde değerlendirilen eserler belli bir aralıktaki skorlar üzerinden ifade edilebilmektedir. Örneğin, filmler IMDb gibi platformlarda 1–10 aralığında puanlanırken, bilgisayar oyunları Metacritic üzerinden 0–100 aralığında sayısal skorlarla değerlendirilmektedir. Ancak bu platformlardaki puanlar ve kullanıcı yorumları, büyük ölçüde kişisel beğenilere dayandığından nesnel olmaktan uzaktır, çarpıtmaya açıktır. Buna karşın müzik alanında, filmler ve oyunlarda olduğu gibi herkes tarafından kabul edilmiş, standart bir skorlandırma sistemi bulunmamaktadır. Bu bağlamda projenin amacı, müzik eserlerini yapay zeka tekniklerinden faydalanarak nesnel ölçütler doğrultusunda analiz etmek ve elde edilen analiz sonuçlarını sayısal bir skor haline dönüştürmektir.

Giriş

Müzik, havadaki titreşimlerin zaman içinde örgütlenmiş bir biçimde algılanmasıyla oluşan, yalnızca zaman içinde var olabilen geçici bir işitsel olgudur. Bununla birlikte müzik, ritim, melodi, armoni ve tını gibi bileşenlerin etkileşimi sayesinde dinleyici tarafından anlamlandırılan çok katmanlı bir algısal deneyimdir (Samama, 2016, pp. 27–31). Müziklerin skorlandırılması ve sınıflandırılmasına yönelik ilk sistematik yaklaşımlar 20. yüzyılın ortalarında, özellikle 1950–1970 yılları arasında müzikoloji ve psikolojinin kesiştiği

çalışmalarda ortaya çıkmıştır (Cuddy, 2009); bu dönemde temel sorun, müziğin estetik ve duygusal etkilerinin nesnel ölçütlerle ifade edilememesi idi. Buna karşın, sinema sektöründe müzikler görsel anlatı ve sahne bağlamı ile birlikte değerlendirildiği için skorlandırılabilmiş; benzer biçimde oyun sektöründe de müzikler oynanış ve etkileşim unsurları ile ilişkilendirilerek ölçülebilir hale gelmiştir (Austin, Moore, Gupta, & Chordia, 2010). Bu bağlamsal yapı, müziğin tek başına değil, ait olduğu medya içeriği üzerinden değerlendirilmesini mümkün kılmıştır.

Proje başlangıcından önce konu ile alakalı literatür taraması yapılmış, geçmiş çalışmalar incelenmiştir; özellikle *"A Survey on Evaluation Metrics for Music Generation" çalışması*, müzik üretimi değerlendirme metriklerinin sınıflandırılması, insan algısıyla nicel metrikler arasındaki uyum sorunları ve daha kapsayıcı değerlendirme çerçeveleri önerileri nedeniyle proje kapsamında temel bir referans olarak alınmıştır (Kader & Karmaker, 2025). Aynı zamanda çalışma içerisinde bulunan eksiklikler tespit edilmiştir. Çalışma şarkı sözlerini anlam derinliği, kafiye, uyak, bağlamından değerlendirilmemiştir. Bu durum vokal müziğin kısmen devre dışı kalmasını sağlar.

Bu doğrultuda çalışmanın ana hipotezi, müzik kalite skorlarının yapay zekâ teknikleri kullanılarak nesnel ve tekrarlanabilir biçimde elde edilebileceğidir. Ana hipotezi destekleyen alt hipotez ise, müzik kalite skoru oluşturma sözlerin anlamsal içerik, biçimsel yapı ve bağlamsal tutarlılık açısından da analiz edilmesinin, elde edilen skorların insan algısıyla benzer şekilde elde edileceği yönündedir.

Sözlerin anlamlandırılması üzerine yapılan araştırmalar, NLP (Doğal Dil İşleme) yöntemleriyle şarkı sözlerinin anlamsal ve duygusal yapısının nicel olarak analiz edilebileceğini göstermiştir (Chen, Wang, Yu, & Zhou, 2024). Metin verisinin redif ve kafiye açısından incelenmesinde ise yapay zeka teknikleri yerine doğrudan fonksiyon kullanımının daha uygun olduğu ortaya çıkmıştır (Can, Duygulu, Can, & Kalpaklı, 2010). Ses verisi üzerinde yapılan çalışmalar, ham ses verisinin verimli bir analiz sağlamadığını, bunun yerine ses verisinden sayısal özellikler çıkarmanın hem verimlilik hem de doğruluk açısından daha etkili olduğunu ortaya koymuştur (Abdusalomov et al., 2022). Son olarak, ses ve metin üzerinden elde edilen özellikler ve model çıktıları, doğrudan bir skor oluşturmak üzere birleştirilecektir (Garcia‑Bernabeu & Hilario‑Caballero, 2021).

Yöntem

Bu proje kapsamında geliştirilen sistem, şarkıları ardışık ve modüler adımlar halinde işleyen bir pipeline mimarisi üzerine kuruludur. Pipeline, bir girdinin ardışık ve bağımlı aşamalardan geçirilerek çıktıya dönüştürüldüğü yapılandırılmış süreçtir (Xin, Miao, Parameswaran & Polyzotis, 2021). Proje içerisinde tüm yazılım süreci için Python yazılım dili kullanılmıştır. Aynı zamanda iş verimini artırmak ve projeyi yapılabilir hale getirmek için çeşitli kütüphanelerden faydalanılmıştır.

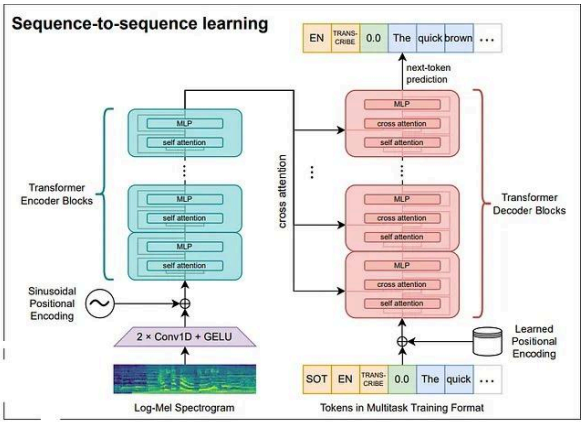
Projenin ileri aşamalarında kullanılmak üzere ilk önce ses verisi enstrüman ve vokal olarak iki parça haline bölünmelidir. Projede özellikle şarkı sözü ve şarkıcı sesi üzerinde durulmak istendiği için bu işlem uygulanacaktır. İşlem için Demucs isimli derin öğrenme modeli kullanılmıştır. Demucs, zaman alanında çalışan, U-Net benzeri evrişimsel sinir ağları ve

bidirectional LSTM (Çift Yönlü Uzun Kısa Süreli Bellek) katmanları içeren, uçtan uca eğitilmiş bir modeldir; ses dalga formunu doğrudan işleyerek vokal ve eşlik (enstrümantal) bileşenlerini ayırır (facebookresearch, 2025). Tablo 1'de Demucs modelinin çıktı aşamasında kullanılan parametreler verilmiştir.

Parametre Adı	Parametre Değeri	Açıklaması
--two-stems	vocals	Çıkışı iki kaynağa ayırır: vokal ve vokal dışı (enstrümantal)
-n	htdemucs_ft	Kullanılan model; fine-tuned HTDemucs mimarisi
-d	cuda	İşlemlerin GPU (CUDA) üzerinde çalıştırılması
--shifts	10	10 zaman kaydırmalı çıkarım yaparak ayrım kalitesini artırır
--overlap	0.5	Segmentlerin %50 örtüşerek işlenmesini sağlar
--segment	7	Sesin 7 saniyelik parçalara bölünerek işlenmesi
--float32	aktif	32-bit kayan nokta hesaplama kullanılır

Tablo 1: Demucs modeli içinde kullanılmış parametreler

Sistem içerisine şarkı sözleri doğrudan verilmediğinden bu işlem kod içerisinde yapılmıştır. İlk olarak MP3 formatında alınan veri konuşma tanıma modellerine uyumlu WAV veri tipine çevrilmiştir. Ardından, OpenAI tarafından geliştirilen ve büyük ölçekli sinir ağlarına dayanan Whisper modelinin large-v3 sürümü yüklenmekte ve bu model, dönüştürülen ses dosyası üzerinde çalıştırılmaktadır. Transkripsiyon işlemi sırasında dil parametresi açıkça Türkçe olarak belirlenerek modelin dilsel bağlamı doğru biçimde ele alması sağlanmaktadır. Sonuç olarak sistem, ham ses verisini girdi olarak alıp, dil bilgisel ve anlamsal açıdan tutarlı bir metin çıktısı üreten, uçtan uca bir otomatik konuşma tanıma (Automatic Speech Recognition – ASR) hattı sunmaktadır. Whisper modelinin yapısı Görsel 1 içerisinde verilmiştir.



Görsel 1: Whisper modelinin veri işleme hattı ve token üretimini gösteren mimari şema. Okoye (2024), Figure 21'den alınmıştır.

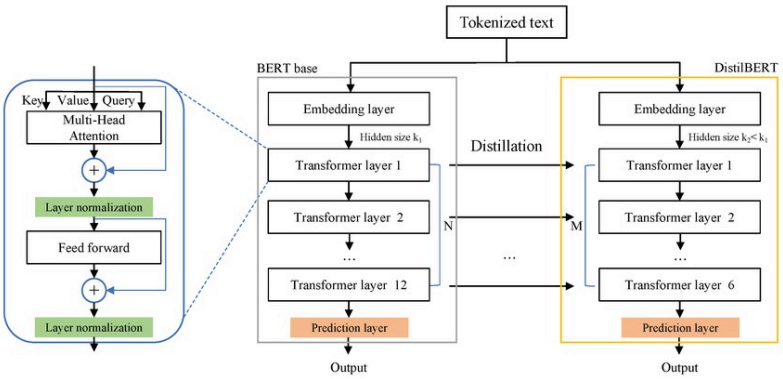
Whisper large-v3 modeli çıktıyı her bir sekans ayrı bir eleman olacak şekilde liste formatında

vermektedir. Yazdığımız fonksiyon bu listeyi tek bir metin haline getirir. Fonksiyon sonrasında redif ve kafiye unsurlarının tespitine geçer; bu süreçte öncelikle birleştirilen metin satır-temelli olarak yeniden bölütlenir (sentence/line segmentation) ve her bir satırın dize sonu sözcüğü hedef öge (target token) olarak belirlenir, ardından bu sözcükler küçük harfe dönüştürme, noktalama işaretlerinin giderilmesi ve Türkçeye özgü karakterlerin korunması gibi normalizasyon adımlarından geçirilerek biçimsel gürültü azaltılır. Devamında, dize sonu sözcükler üzerinde olası son eklerin sistematik olarak üretilmesiyle alt-dizi (suffix) uzayı oluşturulur ve bu uzayda tekrar sıklığı yüksek olan birimler istatistiksel eşikleme yöntemiyle redif adayı olarak etiketlenir. Redifin ayrıştırılmasının ardından geriye kalan sözcük gövdesi üzerinde sesli harf merkezli fonetik çözümleme uygulanarak kafiye çekirdeği (rhyme nucleus) tespit edilir; bu çekirdek yapıdaki sesli harf sayısı ve uzunluğu temel alınarak kafiye türleri sınıflandırılır. Metin uzunluğunu kafiye ve redif miktarını suni biçimde artırdığından cümle sayısı başına düşen ortalama kafiye alınmıştır. Bu şekilde metnin teknik bağlamdan kalitesi ölçülmesi sağlanmıştır.

Metinden tespit edilen bir diğer unsur söz anlam derinliğidir. Yapılan araştırma sonucunda internette açık kaynaklı olarak doğrudan etiketli veri bulunamamıştır. Bu sebepten ötürü verilen sentetik olarak etiklendirilmiştir. Bu aşamada veri olarak Evabot (2021) tarafından sağlanan Spotify Lyrics Dataset kullanılmıştır. Bu veri seti Spotify platformundan API anahtarı tekniği ile çekilmiştir, CC0: Public Domain lisansına sahiptir. Bu veri setinden alınmış 6760 adet şarkı sözü öncesinde bir liste haline getirilmiştir. Sonrasında bu sözler lokal ortama indirilmiş bir LLM (Büyük dil modeli) tarafından puanlandırılmıştır. LLM olarak kuantize edilmiş Mistral-7B modeli kullanılmıştır. Modele verilen sistem promptu (modelin davranışını ve rolünü belirleyen talimatlar) ile 0-10 arası bir puanlandırılma gerçekleştirilmiştir. Sonrasında etiketlendirilmesi yapılan şarkı sözleri CSV formatında model eğitimi için kaydedilmiştir. Veri setinde ayrıca bulunan explicit değeri bahsedilen CSV içinde bulunmaktadır. Bu sütun şarkı sözlerinin şiddet, hakaret vb. unsurların ne kadar barındığını tespit eder. Bu iki etiket üzerine iki ayrı model eğitimi gerçekleşmiştir. CSV verisinde bulunan şarkı sözleri, metin tabanlı makine öğrenmesi ve doğal dil işleme (NLP) uygulamalarında doğrudan modele verilemez. Bu nedenle, öncelikle metinlerin sayısal temsillere dönüştürülmesi ve dilsel olarak işlenmesi gerekir. Bu aşamada NLTK ve TF-IDF vektörleştirici birlikte kullanılmıştır.

NLTK (Doğal Dil Araç Takımı), şarkı sözlerinin ön işleme sürecinde temel rol oynar. Bu kapsamda metinler küçük harfe dönüştürülür, noktalama işaretleri ve anlamsal katkısı olmayan durak kelimeler (stopwords) temizlenir. Ayrıca kelimeler token'lara ayrılarak (tokenization) metnin dilsel yapısı sadeleştirilir. Gerekli durumlarda kök bulma (stemming) veya gövdeleme (lemmatization) uygulanarak kelime varyasyonları ortak bir temsilde birleştirilir (Loper & Bird, 2002). Bu ön işleme adımlarının ardından, temizlenmiş ve normalize edilmiş şarkı sözleri TF-IDF (Terim Sıklığı – Ters Doküman Sıklığı) yöntemi kullanılarak sayısal vektörlere dönüştürülür. TF-IDF, her bir kelimenin ilgili şarkı sözü içerisindeki kullanım sıklığını (TF) ve bu kelimenin tüm veri kümesindeki ayırt ediciliğini (IDF) birlikte değerlendirerek ağırlıklandırma yapar. Bu sayede bir şarkı sözü içinde çok sık geçen ancak veri kümesinin genelinde yaygın olan kelimelerin etkisi azaltılırken, belirli şarkılara özgü ve ayırt edici kelimeler daha yüksek ağırlıklarla temsil edilir. Sonuç olarak her bir şarkı sözü, kelime uzayında yüksek boyutlu fakat sayısal bir vektör haline getirilir (Aizawa, 2003). Bahsedilen iki yöntem ile şarkı sözleri model eğitimine uygun biçime getirilmiştir.

Projede şarkı sözlerinin anlam derinliği modeli için anlamlandırma kapasitesi yüksek, gelişmiş bir model kullanılması gerekmiştir. Bu sebepten ötürü DistilBERT mimarisi tercih edilmiştir. BERT (Dönüştürücülerden Çift Yönlü Kodlayıcı Temsilleri) derin öğrenme tabanlı bir dil modelidir. Yaptığı iş bağlamında NLP (Doğal dil işleme) modelidir. Çalışma prensibi ise diğer NLP modellerden farklıdır. Kelimenin anlamını iki yönlü olarak anlamlandırır; yani bir kelimenin anlamını hem solundaki hem de sağındaki bağlama göre öğrenir (Devlin *et al*., 2019). DistilBERT ise BERT'in daha hızlı ve hafif bir versiyonu olup, boyutu küçültülmüş ve eğitim süresi kısaltılmıştır (Sanh *et al*., 2019). DistilBERT mimarisi yapı olarak Görsel 2'de belirtilen gibidir.



Görsel 2: DistilBERT mimarisinin çalışma mekanizmasını gösteren şema. (Adel et al., 2022), Figure 2'den alınmıştır.

DistilBERT modeli , proje kapsamında şarkı sözlerinin anlam derinliğini sayısal bir değer olarak tahmin etmek üzere, sınıflandırma yerine bir regresyon problemi çatısı altında "distilbert-base-uncased" ön eğitimli ağırlıkları kullanılarak yapılandırılmıştır. Modelin eğitim sürecinde (fine-tuning - Devlin et al., 2018), öğrenme performansını optimize etmek adına ağırlık güncellemelerinin büyüklüğünü belirleyen öğrenme oranı (learning rate) 2e-5, modelin tüm veri setini kaç kez işlediğini ifade eden epok (epoch) sayısı 30 ve işlem birimi başına düşen veri adedini gösteren yığın boyutu (batch size) 8 olarak belirlenmiştir. Model başarısının değerlendirilmesinde ise tahmin edilen değerler ile gerçek değerler arasındaki farkı ölçen Ortalama Mutlak Hata (MAE) esas alınmıştır. Elde edilen nihai performans değerleri çalışmanın Bulgular bölümünde sunulmuştur.

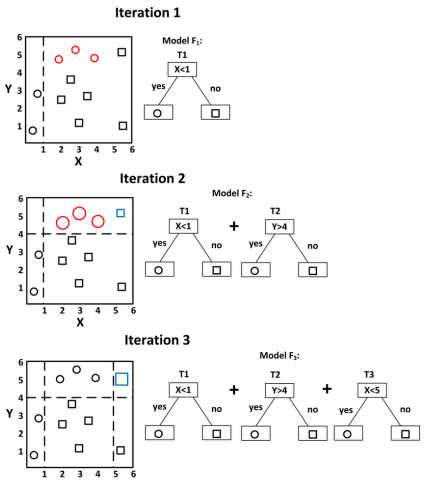
Explicit modeli aşamasında, öznitelik vektörleri ile hedef sınıflar arasındaki doğrusal ilişkiyi modelleyen Logistic Regression (Lojistik Regresyon) algoritması kullanılmıştır (Genkin, Lewis, & Madigan, 2007). Eğitim sürecinde model, "explicit" içeriği temsil eden belirleyici ifadelere pozitif coefficients (katsayılar) atayarak öğrenme işlemini gerçekleştirmiş (Nobata, Tetreault, Thomas, Mehdad, & Chang, 2016) ve global minimuma ulaşmak için 3000 iterasyon yakınsama sağlamıştır. Modelin genelleştirme kapasitesini korurken nadir öznitelikleri baskılamamak adına inverse regularization strength (ters düzenlileştirme gücü) parametresi 2 olarak ayarlanmış; veri setindeki yapısal dengesizliği yönetmek için ise class weight (sınıf ağırlığı) parametresiyle sakıncalı içeriğin hata maliyeti artırılmıştır (King & Zeng, 2001). Eğitim sonucunda üretilen probability scores (olasılık skorları), modelin sakıncalı içeriği yakalama hassasiyetini maksimize etmek amacıyla standart değerden saptırılarak 0.38 olarak belirlenen decision threshold (karar eşiği) üzerinden nihai sınıf etiketlerine dönüştürülmüş olup (Zou, Xie, Lin, Wu, & Ju, 2016), modelin performansı Bulgular bölümünde detaylandırılmıştır.

Şarkı üzerinden feature çıkararak eğitilecek model proje pipelinesi içerisindeki son aşamadır. Model eğitimini için kullanılan şarkı isimleri ve etiketleri, *Hit song science – 34740 songs (+spotify features)* adlı veri setinden alınmıştır (multispiros, n.d.). Veri seti içerisinde bulunan akustiklik, dans edilebilirliği, enerjiklik durumu parametreleri tahmin etmesi üzerine eğitilmiştir. Veri seti içerisinde ismi geçen şarkılar web scraping tekniği ile lokal ortama aktarılmıştır . İndirilen şarkılar yalnızca özellik çıkarımı amacı ile bellek ortamında alınmış, özellik çıkarımı tamamlandıktan sonra silinmiştir . Şarkılarda özellik çıkarımı sürecinde, verilerin zamansal ve spektral karakteristiğini yüksek çözünürlükle modellemek amacıyla, literatürde etkinliği kanıtlanmış segmentasyon tabanlı bir öznitelik çıkarım metodolojisi uygulanmıştır (Alías, Socoró, & Sevillano, 2016). Ses sinyalleri doğası gereği durağan olmayan (non-stationary) bir yapı sergilediğinden, sinyalin istatistiksel özelliklerinin zamanla değişimi göz önüne alınmalıdır. Bu bağlamda, ham ses dosyaları öncelikle 120 saniyelik standart bir süreye normalize edilmiş ve akabinde 10 saniyelik on iki ardışık pencereye (segment) ayrıştırılarak sinyalin yerel düzeyde incelenmesi sağlanmıştır (Giannakopoulos & Pikrakis, 2014).

Her bir segment üzerinden, insan işitsel algısını modellemedeki başarısı nedeniyle Mel-frekans kepstral katsayıları (MFCC) hesaplanmıştır (Logan, 2000). Buna ek olarak, sinyalin fiziksel ve tınısal dokusunu tanımlamak adına spektral merkez (centroid), spektral yuvarlanma (rolloff), RMS enerjisi ve sıfır geçiş oranı (ZCR) gibi psikoakustik tanımlayıcılar kullanılmıştır. Elde edilen bu özniteliklerin zaman içerisindeki değişimini tek bir vektörde temsil edebilmek için, literatürdeki "doku penceresi" yaklaşımına uygun olarak her bir özniteliğin ortalama ve standart sapma değerleri alınmış (Tzanetakis & Cook, 2002); böylece her örneklem için toplam 132 boyuttan oluşan, zamansal değişimi kapsayan hibrit bir öznitelik vektörü elde edilmiştir.

Öznitelik çıkarım aşamasını takiben, elde edilen yüksek boyutlu vektörlerin sınıflandırılması ve modelin tahmin gücünün artırılması amacıyla, topluluk öğrenme (ensemble learning) temelli Gradient Boosting algoritması tercih edilmiştir (Friedman, 2001). Bu yöntem, zayıf öğrenicileri (karar ağaçlarını) ardışık bir şekilde eğiterek ve her adımda önceki modelin hata payını minimize ederek (gradient descent) güçlü bir tahminleyici oluşturur (Hastie, Tibshirani, & Friedman, 2009).

Modelin karmaşıklığını dengelemek ve aşırı öğrenme (overfitting) riskini kontrol altında tutmak adına hiperparametreler deneysel süreçler sonucunda optimize edilmiştir. Bu doğrultuda; modelin iterasyon kapasitesini belirleyen tahminci sayısı (*n_estimators*) 300, modelin öğrenme hızını ve genelleme yeteneğini kontrol eden öğrenme oranı (*learning_rate*) 0.05 ve ağaçların karmaşıklığını sınırlayan maksimum derinlik (*max_depth*) 4 olarak belirlenmiştir. Deneylerin tekrarlanabilirliğini (reproducibility) ve sonuçların tutarlılığını sağlamak amacıyla ise rastgelelik çekirdeği (*random_state*) 42 olarak sabitlenmiştir. Eğitilen modelin her bir öznitelik grubu özelindeki performansı, ayırt edicilik düzeyleri ve elde edilen nihai doğruluk skorları, Bulgular bölümünde detaylı olarak sunulmuştur. Gradient Boosting algoritmasının örneği Görsel 3'te verilmiştir.



Görsel 3. Gradient Boosting modelini gösteren mimari şema (Del Río-Villegas, 2018, Figure 5).

Kullanılan tüm yöntemlerin sonuç olarak doğrudan tek bir skoru işaret etmesi gerekmektedir. Elde edilen çıktılar bir metrik olarak birleştirilmiştir. Elde edilen veriler farklı aralıklarda (0-1, 0-10, 0-60, 0-100) olduğu için, sağlıklı bir toplama işlemi yapabilmek adına tüm parametreler 0-100 ölçeğine normalize edilmiştir. Özellikle toksisite analizi gibi negatif yönlü parametreler, "kalite skoru" mantığına uygun olarak ters çevrilmiş (inverse normalization); böylece düşük toksisite oranının yüksek puana karşılık gelmesi sağlanmıştır. Ayrıca, şarkıların türüne göre şarkının kalitesini belirleyen unsurlar değişebilmektedir. Bu süreçte her bir bileşenin genel skor üzerindeki etkisi, müzik öğretmenlerine sorularak belirlenmiştir. Tablo 2'de skorlama sisteminde kullanılan bileşenlerin ham ölçeklerini, uygulanan normalizasyon işlemlerini ve nihai ağırlık katsayılarını göstermektedir.

Bileşen	Ham Ölçek	Normalizasyon Formülü	Ağırlık	Rap	Klasik müzik	Türkü	Pop	Rock
Toksisite olmama durumu	$0 - 1$	$(1 - x) * 100$	0.10	0.15	0.05	0.05	0.08	0.10
Anlamsal Analiz	$0 - 10$	$(x / 10) * 100$	0.20	0.30	0.35	0.40	0.20	0.25
Kafiye + Redif	$0 - 60$	$(x / 60) * 100$	0.17	0.30	0.03	0.25	0.10	0.10
Akustiklik	$0 - 1$	$x * 100$	0.17	0.05	0.45	0.20	0.12	0.20
Enerji	$0 - 1$	$x * 100$	0.17	0.10	0.05	0.05	0.20	0.25
Dans Edilebilirlik	$0 - 1$	$x * 100$	0.19	0.10	0.07	0.05	0.30	0.10

Tablo 2: Skorlama sisteminde yer alan unsurlar hakkında bilgiler

Proje İş-Zaman Çizelgesi

AYLAR							
İşin Tanımı	**Haziran**	**Temmuz**	**Ağustos**	**Eylül**	**Ekim**	**Kasım**	**Aralık**
Literatür Taraması	X	X	X	X	X		
Veri Toplanması		X	X	X	X		
Sentetik Veri Etiketlenmesi			X				
Kafiye ve Redif Fonksiyonu Yazımı			X				
Metin Modellerinin Eğitimi			X	X			
Ses Modellerinin Eğitimi				X	X		
Rapor yazımı						X	X

Bulgular

Proje sonucunda Tablo 2'de bahsedilen unsurları çıkaracak pipeline oluşturulmuştur. Pipeline içerisinde 2 adet NLP, 1 adet müzik feature modeli bulunmaktadır. Modellerin eğitim sonucunda alınan test skorları Tablo 3'teki gibidir.

Model Adı	Model Türü	Test Örnek Sayısı	Test Skor Türü	Alınan Skor
Explicit modeli	Logistic Regression	500	Makro-F1	0.9521
Söz anlam modeli	DistilBERT	600	MAE	0.8470
Şarkı feature modeli - Acousticness	Gradient Boosting	550	MAE	0.1520
Şarkı feature modeli - Danceability	Gradient Boosting	550	MAE	0.0928
Şarkı feature	Gradient	550	MAE	0.1243

modeli - Energy	Boosting			

Tablo 3: Eğitilen modeller ile ilgili bilgiler

Sistemin örnek kullanımını göstermek için iki adet şarkı ile test yürütülmüştür. Bu şarkılar Hammâmizâde İsmail Dede Efendi tarafından bestelenmiş "Yine Bir Gülnihal" ile söz bestesi anonim olan "Zühtü" türküsüdür. Tablo 4'te şarkıların 100 üzerinden alınmış toplam puanı ve alan bazlı puanları verilmiştir.

Şarkı Adı	Toksisite olmama durumu	Anlamsal Analiz	Kafiye + Redif	Akustiklik	Enerji	Dans Edilebilirlik	Türe Göre Ağırlıklandırılmış Toplam Skor
Yine Bir Gülnihal	98	83	65.15	64.5	51.3	57.9	71.5475
Zühtü	97	74.4	68.33	57.1	57.9	68.6	69.4375

Tablo 4: "Yine Bir Gülnihal" ile "Zühtü" parçalarının projeye göre skor durumları

Sonuç ve Tartışma

Tablo 3'te sunulan veriler incelendiğinde, geliştirilen modellerin yüksek doğruluk oranlarına ulaştığı görülmektedir. Explicit İçerik Modeli, 0.9521 gibi yüksek bir Makro-F1 skoruyla sakıncalı içeriği tespit etmede üstün bir başarı sergilemiştir. Bu durum, Lojistik Regresyon'un dengelenmiş sınıf ağırlıkları ve optimize edilmiş karar eşiği (0.38) ile birleştiğinde metin sınıflandırmada ne kadar efektif olabileceğini kanıtlamaktadır. Anlam Derinliği Modeli (DistilBERT) tarafında elde edilen 0.8470 MAE değeri, 0-10 ölçeğindeki bir değerlendirme için hata payının oldukça düşük olduğunu ve modelin şarkı sözlerindeki bağlamsal derinliği insan değerlendirmesine yakın bir seviyede tahmin edebildiğini göstermektedir. Şarkı Feature Modeli kapsamında eğitilen Gradient Boosting modelleri ise özellikle "Danceability" (0.0928 MAE) ve "Energy" (0.1243 MAE) tahminlerinde oldukça hassas sonuçlar üretmiştir. Bu başarı, MFCC ve spektral özniteliklerin (centroid, rolloff vb.) birleştirilerek oluşturulan 132 boyutlu hibrit vektörlerin, sesin karakteristiğini başarılı bir şekilde temsil ettiğini doğrulamaktadır.

Proje geçmiş çalışmalardan farklı olarak şarkıların puanlandırılmasını yalnızca ses dalgaları üzerinden çözülebilecek bir problem olarak görülmemiştir. Şarkı sözleri geçmiş çalışmaların aksine teknik ve derinlik bakımından incelenmiş, şarkı üzerinden özellik çıkarımı yapılmıştır. Şarkı puanlandırmaya çeşitli ve yenilikçi bir çözüm önerisi sunan çalışma yürütülmüştür.

Öneriler

Gelecekteki çalışmalarda, mevcut ses modelinde kullanılan akustiklik, dans edilebilirlik ve enerji parametrelerinin ötesine geçilerek; tempo, valans (duygusal pozitiflik) ve

enstrümantallik gibi ek özniteliklerin sürece dahil edilmesi, modelin tahmin tutarlılığını ve ayırt edicilik kapasitesini doğrudan artıracaktır. El ile belirlenen özniteliklerin (hand-crafted features) yanı sıra ham ses verisini spektrogramlar üzerinden işleyen Evrişimsel Sinir Ağları (CNN) veya Audio Spectrogram Transformer gibi uçtan uca öğrenme mimarilerine geçiş yapılması, sesin zamansal ve tınısal derinliğinin mevcut 132 boyutlu vektörden çok daha yüksek bir çözünürlükle temsil edilmesine olanak sağlayacaktır.

Modellerin genelleştirme kabiliyetini ve farklı kültürel dokulara adaptasyonunu güçlendirmek adına, eğitimde kullanılan veri setlerinin hem niceliksel hacmi hem de niteliksel çeşitliliği genişletilmelidir. Özellikle yerel makam yapılarını, farklı dillerdeki edebi sanatları ve çeşitli kayıt kalitelerini içeren daha kapsamlı ve dengeli veri kümelerinin sisteme entegre edilmesi, modelin belirli türlere olan yanlılığını (bias) minimize edecektir; bu doğrultuda veri setlerindeki yapısal dengesizlikleri gidermek için sentetik veri üretimi veya veri artırma (data augmentation) tekniklerinden yararlanılması, sistemin çok daha geniş bir müzikal yelpazede yüksek doğrulukla çalışmasını mümkün kılacaktır.

Kaynaklar

Samama, L. (2016). *Meaning of music*. Amsterdam University Press. https://doi.org/10.1017/9789048528929.003

Austin, A., Moore, E., Gupta, U., & Chordia, P. (2010). *Characterization of movie genre based on music score*. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 421–424). IEEE.

Kader, F. B., & Karmaker, S. (2025). *A survey on evaluation metrics for music generation*. *arXiv*. https://arxiv.org/abs/2509.00051

Chen, Y., Wang, H., Yu, K., & Zhou, R. (2024). *Artificial intelligence methods in natural language processing: A comprehensive review*. *Highlights in Science, Engineering and Technology*. https://doi.org/10.54097/vfwgas09

Abdusalomov, A. B., Safarov, F., Rakhimov, M., Turaev, B., & Whangbo, T. K. (2022). *Improved feature parameter extraction from speech signals using machine learning algorithm*. Sensors, 22(21), 8122. https://doi.org/10.3390/s22218122

Garcia‐Bernabeu, A., & Hilario‐Caballero, A. (2021). *Monitoring multidimensional phenomena with a multicriteria composite performance interval approach* (arXiv:2107.08393). arXiv. https://doi.org/10.48550/arXiv.2107.08393

Can, E. F., Duygulu, P., Can, F., & Kalpaklı, M. (2010). Redif extraction in handwritten Ottoman literary texts. In *2010 20th International Conference on Pattern Recognition* (pp. 1–4). IEEE. https://doi.org/10.1109/ICPR.2010.478

Cuddy, L. L. (2009). Development of music perception and cognition research: An autobiographical account from a Canadian perspective. *Psychomusicology: Music, Mind & Brain, 20*(1 & 2). https://doi.org/10.5084/pmmb2009/20/43

Xin, D., Miao, H., Parameswaran, A., & Polyzotis, N. (2021). *Production machine learning pipelines*. In *Proceedings of the 2021 International Conference on Management of Data* (pp.

2639–2652). ACM. https://doi.org/10.1145/3448016.3457566

facebookresearch. (2025). *demucs* [Computer software]. GitHub. https://github.com/facebookresearch/demucs

Okoye, O. (2024, November 2). Whisper: Functionality and Finetuning. Medium. https://medium.com/@okezieowen/whisper-functionality-and-finetuning-ba7f9444f55a

Eva Bot. (2022). *Spotify lyrics dataset* [Dataset]. Kaggle. https://www.kaggle.com/datasets/evabot/spotify-lyrics-dataset

Loper, E., & Bird, S. (2002). *NLTK: The Natural Language Toolkit*. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics* (pp. 63–70). Association for Computational Linguistics. https://aclanthology.org/W02-0109/

Aizawa, A. (2003). *An information-theoretic perspective of tf–idf measures*. *Information Processing & Management, 39*(1), 45–65. https://doi.org/10.1016/S0306-4573(02)00021-3

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. *Proceedings of NAACL*, 4171–4186. doi:10.18653/v1/N19-1423

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. arXiv preprint arXiv:1910.01108.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development, 7*(3), 1247-1250. https://doi.org/10.5194/gmd-7-1247-2014

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science, 7*, e623. https://doi.org/10.7717/peerj-cs.623

Adel, H., Dahou, A., Mabrouk, A., Abd Elaziz, M., Kayed, M., El-Henawy, I. M., Alshathri, S., & Amin Ali, A. (2022). Improving crisis events detection using DistilBERT with Hunger Games Search Algorithm. Mathematics, 10(3), 447. https://doi.org/10.3390/math10030447

Genkin, A., Lewis, D. D., & Madigan, D. (2007). Large-scale Bayesian logistic regression for text categorization. *Technometrics, 49*(3), 291-304. https://doi.org/10.1198/004017007000000245

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis, 6*(5), 429-449. https://doi.org/10.3233/IDA-2002-6504

Zou, Q., Xie, S., Lin, Z., Wu, M., & Ju, Y. (2016). Finding the best classification threshold in imbalanced classification. *Big Data Research, 5*, 2-8. https://doi.org/10.1016/j.bdr.2015.12.001

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive language detection in online user content. *Proceedings of the 25th International Conference on World Wide Web*, 145-153. https://doi.org/10.1145/2872427.2883062

multispiros. (n.d.). *Hit song science – 34740 songs (+spotify features)* [Dataset]. Kaggle. https://www.kaggle.com/datasets/multispiros/34740-hit-and-nonhit-songs-spotify-features

Pérez Molano, G. (2023). *An overview of web scraping: Technical aspects and exercises*. Graduate School, Polytechnic University of Puerto Rico. ,

Alías, F., Socoró, J. C., & Sevillano, X. (2016). A review of physical and perceptual feature extraction for music audio classification. *EURASIP Journal on Audio, Speech, and Music Processing*, *2016*(1), 1-27.

Giannakopoulos, T., & Pikrakis, A. (2014). *Introduction to audio analysis: A MATLAB® approach*. Academic Press.

Logan, B. (2000, October). Mel Frequency Cepstral Coefficients for Music Modeling. In *ISMIR* (Vol. 270, pp. 1-11).

Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, *10*(5), 293-302.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189-1232.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

Del Río-Villegas, R. (2018). *A simple example of visualizing gradient boosting* [Figure]. ResearchGate. https://www.researchgate.net/figure/A-simple-example-of-visualizing-gradient-boosting_fig5_326379229