T.C.
ERZİNCAN BİNALİ YILDIRIM ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
YAPAY ZEKA VE ROBOTİK ANABİLİM DALI

DERİN ÖĞRENME YÖNTEMLERİ KULLANILARAK ÇOK MODLU DUYGU ANALİZİ

Ayşe TEKİN

Danışman: Dr. Öğr. Üyesi Mehmet Kürşat ÖKSÜZ

TEZ JÜRİ ÜYELERİ
Dr. Öğr. Üyesi Mehmet Kürşat ÖKSÜZ
Doç. Dr. Volkan KAYA
Doç. Dr. Bilal ERVURAL

YÜKSEK LİSANS TEZİ
ERZİNCAN, 2026

Kabul ve Onay Sayfası

Dr. Öğr. Üyesi M. Kürşat ÖKSÜZ danışmanlığında, Ayşe TEKİN tarafından hazırlanan bu çalışma ……. tarihinde aşağıdaki jüri tarafından Yapay Zeka ve Robotik Anabilim Dalı'nda Yüksek Lisans Tezi olarak kabul oybirliği/oy çokluğu (…/…) ile kabul edilmiştir.

Başkan :	Dr. Öğr. Üyesi Mehmet Kürşat ÖKSÜZ	İmza:
Üye :	Doç. Dr. Volkan KAYA	İmza:
Üye :	Doç. Dr. Bilal ERVURAL	İmza:
Üye :	Doç. Dr. İsmail AKGÜL	İmza:
Üye :	Dr. Öğr. Üyesi Ömer Faruk GÜRCAN	İmza:

Bu tez Enstitü Yönetim Kurulunun …. / …. / 20.... tarih ve …../…….. sayılı kararı ile onaylanmıştır.

Prof. Dr. Kemal Volkan ÖZDOKUR

Enstitü Müdür V.

Not: Bu tezde kullanılan özgün ve başka kaynaklardan yapılan bildirişlerin, şekil ve tabloların kaynak olarak kullanımı, 5846 sayılı Fikir ve Sanat Eserleri Kanunundaki hükümlere tabidir.

Bilimsel Etiğe Uygunluk Sayfası

“Derin Öğrenme Yöntemleri Kullanılarak Çok Modlu Duygu Analizi” isimli “Yüksek Lisans” tezim tarafımca intihal tespit programı ile incelenmiştir. Buna göre tezimde bilimsel etik ihlali ve intihal olarak nitelendirilebilecek herhangi bir durum olmadığını taahhüt ederim.

Bu çalışmadaki tüm bilgilerin, akademik ve etik kurallara uygun bir biçimde elde edildiğini; aynı zamanda bu kural ve davranışların gerektirdiği gibi, bu çalışmanın özünde olmayan tüm materyal ve sonuçları tam olarak aktardığımı ve referans gösterdiğimi beyan ederim. …./…./20…

(İmza)

Ayşe TEKİN

ÖZET

DERİN ÖĞRENME YÖNTEMLERİ KULLANILARAK ÇOK MODLU DUYGU ANALİZİ

Ayşe TEKİN

Yüksek Lisans Tezi
Erzincan Binali Yıldırım Üniversitesi, Fen Bilimleri Enstitüsü,
Yapay Zeka ve Robotik Anabilim Dalı
Danışman: Dr. Öğr. Üyesi Mehmet Kürşat ÖKSÜZ
2026, 64 sayfa

Çok modlu duygu analizi; metin, ses ve görüntü gibi farklı veri türlerinin birlikte değerlendirilmesiyle duyguların daha kapsamlı biçimde incelenmesini sağlayan bir yaklaşımdır. Metin verilerinde dil yapısı ve kelime seçimi; ses verilerinde tonlama, vurgu ve ses perdesi; görsel verilerde ise yüz ifadeleri ve mimikler duygu durumunun anlaşılmasında önemli ipuçları sunmaktadır. Bu yöntem günümüzde sosyal medya analizinden sağlık, eğitim ve pazarlama gibi birçok alanda kullanılmaktadır.

Bu çalışmada metin ve ses verileri birlikte ele alınarak çok modlu bir duygu analizi modeli geliştirilmiştir. Çalışmada MELD veri seti kullanılmış ve deneyler üç farklı etiketleme senaryosu altında gerçekleştirilmiştir: 3 duygu (negative, neutral, positive), 5 duygu (anger, sadness, joy, neutral, surprise) ve 7 duygu (anger, sadness, joy, neutral, surprise, fear, disgust). Metin modalitesi için önceden eğitilmiş RoBERTa modeli kullanılarak bağlamsal metin temsilleri elde edilmiş; ses modalitesi için ise kendi kendine denetimli öğrenme yaklaşımıyla temsil üreten HuBERT modeli tercih edilmiştir. HuBERT temsilleri MLP tabanlı bir sınıflandırıcı ile duygu sınıflandırma sürecine dâhil edilmiştir.

Metin ve ses modaliteleri geç füzyon yaklaşımıyla karar düzeyinde birleştirilmiştir. Model performansı Weighted F1 (WF1) başta olmak üzere doğruluk, ROC eğrileri ve karışıklık matrisleri ile değerlendirilmiştir. Sonuçlar, geç füzyon modelinin tekil modellere kıyasla daha tutarlı performans sağladığını göstermiştir. WF1 değerleri 3, 5 ve 7 duygu senaryoları için sırasıyla 72, 66 ve 62 olarak elde edilmiştir.

Anahtar Kelimeler: Çok modlu duygu analizi, geç füzyon, roberta, hubert, derin öğrenme

ABSTRACT

MULTIMODAL SENTIMENT ANALYSIS USING DEEP LEARNING METHODS

Ayşe TEKİN

Master's Thesis
Erzincan Binali Yıldırım University, Institute of Science and Technology,
Department of Artificial Intelligence and Robotics
Advisor: Assist. Prof. Mehmet Kürşat ÖKSÜZ
2026, 64 pages

Multimodal sentiment analysis enables a more comprehensive examination of emotions by jointly evaluating different data modalities such as text, audio, and visual cues. Linguistic structures and word choices in text, prosodic features such as intonation and pitch in audio, and facial expressions in visual data provide complementary signals for understanding emotional states. This approach has been widely applied in areas including social media analysis, healthcare, education, and marketing.

In this study, a multimodal emotion recognition model was developed by jointly utilizing text and audio modalities. Experiments were conducted on the MELD (Multimodal EmotionLines Dataset) under three labeling scenarios: 3 emotions (negative, neutral, positive), 5 emotions (anger, sadness, joy, neutral, surprise), and 7 emotions (anger, sadness, joy, neutral, surprise, fear, disgust). For the text modality, contextual representations were obtained using a pre-trained RoBERTa model, while HuBERT was used for audio representation learning through a self-supervised approach. The extracted HuBERT features were classified using a lightweight MLP-based architecture.

Text and audio modalities were integrated using a late fusion strategy, where class probabilities from separately trained models were combined at the decision level. Model performance was evaluated using Weighted F1 alongside accuracy and additional analyses such as ROC curves and confusion matrices. Results indicate that the late fusion model provides more consistent performance compared to single-modality approaches, achieving WF1 scores of 72, 66, and 62 for the 3-, 5-, and 7-class scenarios, respectively.

Keywords: Multimodal sentiment analysis, late fusion, roberta, hubert, deep learning

TEŞEKKÜR

Bu tez çalışmasının hazırlanma sürecinde, bilgi ve deneyimleriyle yol gösteren, her aşamada desteğini esirgemeyen değerli danışman hocam Dr. Öğr. Üyesi Mehmet Kürşat ÖKSÜZ'e içten teşekkürlerimi sunarım.

Maddi ve manevi destekleriyle hayatım boyunca yanımda olan, yüksek lisans sürecinde desteklerini her zaman hissettiren, sabır ve anlayışlarıyla bana güç veren anneme, babama ve kardeşime teşekkür ederim.

Üzerimde emeği bulunan ve artık aramızda olmayan babaanneme.

Ayşe TEKİN

Ocak, 2026

İÇİNDEKİLER

TABLOLAR DİZİNİ

ŞEKİLLER DİZİNİ

SİMGELER VE KISALTMALAR DİZİNİ

1-α	Füzyonda ses modalitesinin ağırlık katsayısı
AI	Artificial Intelligence (Yapay Zekâ)
ANN	Artificial Neural Network (Yapay Sinir Ağı)
AUC	Area Under Curve
BERT	Bidirectional Encoder Representations from Transformers
BiLSTM	Bidirectional Long Short-Term Memory
C	Sınıf Sayısı
CNN	Convolutional Neural Network (Evrişimsel Sinir Ağı)
CUDA	Compute Unified Device Architecture
DNN	Deep Neural Network (Derin Sinir Ağı)
DVM	Destek Vektör Makinesi
ELM	Extreme Learning Machine
FCN	Fully Connected Network
FN	False Negative
FN	Yanlış Negatif
FP	False Positive
FP	Yanlış Pozitif
GCN	Graph Convolutional Network
GPU	Graphics Processing Unit
GRU	Gated Recurrent Unit
HuBERT	Hidden-Unit BERT
KNN	K-Nearest Neighbors (K-En Yakın Komşu)
L	Kayıp (Loss) fonksiyonu
LSTM	Long Short-Term Memory (Uzun Kısa Süreli Bellek)
M	Maskelenmiş zaman adımları kümesi (HuBERT)
MELD	Multimodal EmotionLines Dataset
MFCC	Mel-Frequency Cepstral Coefficients (Mel-Frekans Kepstrum Katsayıları)
MFSC	Mel-Frequency Spectral Coefficients
N_i	i. Sınıfa Ait Örnek Sayısı (Support)
N	Toplam Örnek Sayısı

NLP	Natural Language Processing (Doğal Dil İşleme)
NLP	Natural Language Processing
P(x/y)	Koşullu olasılık
$P_{audio}(y/x_{audio})$	Ses modalitesine ait çıktı olasılığı
$P_{text}(y/x_{text})$	Metin modalitesine ait çıktı olasılığı
RCNN	Recurrent Convolutional Neural Network
ReLU	Rectified Linear Unit
RNN	Recurrent Neural Network (Tekrarlayan Sinir Ağı)
RoBERTa	Robustly Optimized BERT Approach
ROC	Receiver Operating Characteristic
SER	Speech Emotion Recognition (Konuşma Duygu Tanıma)
STFT	Short-Time Fourier Transform
SVM /	Support Vector Machine
TN	True Negative
TN	Doğru Negatif
TP	True Positive
TP	Doğru Pozitif
x	Girdi verisi
y	Hedef sınıf / duygu etiketi
z_t	t anındaki gizli akustik birim
α (Alpha)	Füzyonda metin modalitesinin ağırlık katsayısı

1. GİRİŞ

Duygular, insan iletişiminin temel bir parçasıdır ve bireylerin birbirlerini anlamalarında önemli rol oynamaktadır. Konuşma sırasında kullanılan kelimeler, ses tonu ve vurgu gibi unsurlar, duyguların doğru bir şekilde karşı tarafa aktarılmasını sağlamaktadır. Ancak yapay zekâ ve bilgisayar sistemleri için insan duygularını anlamlandırmak hâlâ karmaşık bir süreçtir. Bu zorluk, duygu analizi kavramının ortaya çıkmasına ve bu konuda pek çok araştırma yapılmasına neden olmuştur.

Duygu analizi; ses, metin veya görsel gibi farklı veri türlerinden elde edilen bilgileri inceleyerek bir kişinin duygu durumunu belirlemeyi amaçlayan bir yöntemdir. Doğal dil işleme (NLP), makine öğrenmesi ve derin öğrenme gibi ileri teknolojiler sayesinde bu alandaki çalışmalar hızla gelişmektedir. Farklı veri türlerinin bir arada kullanılması ise çok modlu duygu analizi kavramını gündeme getirmiştir. Çok modlu yaklaşımlarda metin verilerinde kelime seçimleri ve dil yapısı, ses verilerinde ise tonlama, vurgu ve ritim gibi özellikler birleştirilerek daha doğru ve güvenilir tahminler yapılmaktadır. Görsel verilerde ise yüz ifadeleri, mimikler, göz hareketleri ve baş pozisyonu gibi ipuçları kullanılarak bireyin duygusal durumu analiz edilir.

Duyguların bilgisayar sistemleri tarafından modellenmesi ve işlenmesine yönelik ilk kapsamlı çalışmalar, duygusal hesaplama (affective computing) alanı kapsamında Rosalind W. Picard tarafından ele alınmıştır. Metin tabanlı duygu analizi (sentiment analysis) ise, doğal dil işleme alanındaki gelişmelerle birlikte 2000'li yılların başından itibaren ayrı bir araştırma alanı olarak şekillenmiştir (Picard, 1997). İlk dönem duygu analizi çalışmalarında, metinlerde yer alan olumlu ve olumsuz ifadelerin belirlenmesi amacıyla kelime tabanlı (lexicon-based) ve istatistiksel yöntemlerden yararlanılmıştır (Pang vd., 2002). Araştırmacılar, sözcüklerin duygusal değerini ölçerek bir metnin genel duygu yönelimini sınıflandırmaya çalışmışlardır (Liu B. , 2022). 2000'li yıllarda internetin yaygınlaşması ve sosyal medya platformlarının hızla gelişmesiyle birlikte, kullanıcı yorumları ve çevrim içi içeriklerden duygusal bilgi çıkarımı yapılması bu alanın kapsamını genişletmiştir (Go vd., 2009). Böylece duygu analizi, metin temelli sınırlı bir yaklaşım olmaktan çıkıp, insan duygularını farklı veri türlerinden öğrenebilen çok boyutlu bir araştırma alanına dönüşmüştür (Qaisar vd., 2020). Günümüzde ise NLP, derin öğrenme modelleriyle birleşerek daha karmaşık duygusal örüntüleri tespit edebilmekte ve insan-bilgisayar etkileşimi sistemlerinin duygusal farkındalığını artırmaktadır (Kim, 2014).

İşitsel veriler, insan duygularının anlaşılmasında önemli bir bilgi kaynağıdır. Konuşma sırasında sesin tonu, yüksekliği, ritmi, vurgusu ve hızındaki değişiklikler, bireyin duygusal durumuna dair güçlü ipuçları taşımaktadır. (Schuller vd., 2009) Bu nedenle ses tabanlı duygu analizi, kelimelerin içeriğinden bağımsız olarak duygusal ifadelerin tespit edilmesinde etkili bir yöntem olarak görülmektedir (Zhang vd., 2020). Bu alandaki erken dönem çalışmalar, temel akustik özelliklerin (örneğin temel frekans, enerji, formant yapıları) istatistiksel olarak incelenmesine dayanırken, son yıllarda mel-frekans spektrum katsayıları (MFCC), spektrogram ve mel-spektrogram gibi özniteliklerin kullanımıyla daha yüksek doğruluk oranlarına ulaşılmıştır (Nwe vd., 2003). Derin öğrenme yöntemlerinin gelişmesiyle birlikte Evrişimsel Sinir Ağları (CNN) ve Tekrarlayan Sinir Ağları (RNN/LSTM) ses verilerinden duygusal örüntülerin otomatik olarak çıkarılmasında yaygın biçimde kullanılmaya başlanmıştır. (Neumann vd., 2017) Ses tabanlı yaklaşımlar, metin veya görsel verilerle birleştirildiğinde duyguların hem anlamsal hem akustik boyutlarının birlikte değerlendirilmesini sağlayarak daha kapsamlı bir analiz imkânı sunmaktadır.

1.1. Çalışmanın Amacı

Bu çalışmada ses ve metin tabanlı çok modlu duygu analizi ele alınmıştır. Çalışmanın temel amacı, ses verilerinden önceden eğitilmiş HuBERT modeli aracılığıyla elde edilen akustik temsiller ile metin verilerinden önceden eğitilmiş RoBERTa modeli kullanılarak çıkarılan bağlamsal metin temsillerini, ayrı ayrı derin öğrenme tabanlı sınıflandırma modelleriyle işleyerek, bu modalitelere ait çıktıların geç füzyon (late fusion) yaklaşımıyla birleştirildiği bir duygu sınıflandırma sistemi geliştirmektir.

Bu kapsamda ses ve metin modaliteleri bağımsız olarak modellenmiş, elde edilen sınıflandırma çıktıları karar düzeyinde birleştirilmiş ve tek modlu yaklaşımlara kıyasla daha yüksek doğruluk oranı elde edilmesi hedeflenmiştir. Bu sayede çok modlu yöntemlerin duygu sınıflandırmadaki katkıları ortaya konulmuştur.

1.2. Çalışmanın Önemi ve Katkıları

Duygu analizi, müşteri memnuniyeti ölçümlerinden sosyal medya içeriklerinin değerlendirilmesine, psikolojik danışmanlıktan insan-robot etkileşimine kadar birçok alanda önemli bir yere sahiptir. Bu çalışmada önerilen ses ve metin tabanlı çok modlu yaklaşım, tek

bir veri türüne dayalı analizlerin eksikliklerini gidermeyi ve duyguların daha doğru biçimde sınıflandırılmasını sağlamayı hedeflemektedir.
Çalışma, ses ve metin modalitelerinin bağımsız biçimde modellenerek, bu modalitelere ait sınıflandırma çıktılarının geç füzyon yöntemiyle karar düzeyinde birleştirildiği yapısıyla literatürdeki benzer çalışmalardan ayrılmakta; hem akademik anlamda çok modlu duygu analizi alanına katkı sağlamakta hem de gerçek dünyada uygulanabilir bir duygu analiz sistemi için temel oluşturmaktadır.

Yapay zekâ sistemlerinde insan benzeri etkileşimlerin geliştirilebilmesi, duygusal zekâ bileşenlerinin modellenmesini gerekli kılmaktadır (Ekman, 1992). Bu durum, çok modlu duygu analizi çalışmalarına olan ilgiyi artırmakta ve gelecekte daha gerçekçi insan-bilgisayar etkileşimleri tasarlanmasına zemin hazırlamaktadır. (Zhuang vd., 2025)

1.3. Çalışmanın Özgün Yanı

Bu çalışmanın özgünlüğü ve gerekliliği, kullanılan veri seti, modelleme yaklaşımı ve çok modlu birleştirme stratejisinin birlikte ele alınış biçiminde yatmaktadır. Literatürdeki çok modlu duygu analizi çalışmalarının önemli bir bölümü, tek cümlelik ve bağlamdan bağımsız metinler veya tekil konuşma segmentleri üzerinden yürütülmekte; karşılıklı diyalog yapısını ve anlamsal bağlam sürekliliğini yeterince yansıtmamaktadır. Bu tez çalışmasında kullanılan MELD veri seti, çok konuşmacılı, bağlama duyarlı ve karşılıklı diyalog yapısı içermesi nedeniyle, duyguların bağlamsal olarak nasıl evrildiğini modellemeye olanak tanımakta ve bu yönüyle birçok mevcut çalışmadan ayrışmaktadır.
Ayrıca literatürde Transformer tabanlı ön-eğitimli modeller yaygın biçimde kullanılmakla birlikte, bu modellerin tek modlu ve çok modlu yapılardaki katkıları çoğunlukla farklı çalışmalar ve farklı deneysel kurulumlar altında raporlanmaktadır. Bu çalışma kapsamında, RoBERTa (metin) ve HuBERT (ses) modelleri aynı veri seti ve aynı deneysel kurgu altında hem tek modlu hem de çok modlu senaryolarda sistematik olarak karşılaştırılmış; böylece çok modlu yapının katkısı doğrudan ve tutarlı biçimde analiz edilmiştir.

Çalışmanın bir diğer özgün yönü, geç füzyon aşamasında kullanılan ağırlık katsayısının (α) sabit bir değer olarak belirlenmemesi ve literatürde yaygın olan tek seferlik birleştirme yaklaşımlarının ötesine geçilmesidir. Bu tezde, seed tabanlı çoklu çalıştırmalar gerçekleştirilmiş; her çalıştırmada ortaya çıkabilecek rastlantısal farklılıkların etkisini azaltmak amacıyla α katsayısı deneysel olarak optimize edilmiş ve sonuçlar ortalama performans

üzerinden raporlanmıştır. Bu yaklaşım, geç füzyonun daha kararlı ve genellenebilir biçimde değerlendirilmesini sağlamaktadır.

Bu yönleriyle çalışma, bağlamsal diyalog yapısına sahip bir veri seti üzerinde Transformer tabanlı metin ve ses temsillerini, optimize edilmiş ve seed tabanlı geç füzyon stratejisiyle bütüncül bir deneysel çerçeve içinde ele alarak, çok modlu duygu analizi literatüründeki önemli bir metodolojik boşluğu doldurmayı hedeflemektedir.

Bu çalışmanın dayandığı bilimsel arka planı ve araştırma boşluğunu ortaya koymak amacıyla bir sonraki bölümde, duygu analizi ve çok modlu duygu tanıma alanında gerçekleştirilen çalışmalar sistematik biçimde incelenmektedir. Literatürde kullanılan veri setleri, modalite türleri (metin/ses/görüntü), temsil çıkarım yaklaşımları ve füzyon stratejileri özetlenerek, mevcut yöntemlerin güçlü yönleri ve sınırlılıkları tartışılmakta; böylece bu tez kapsamında önerilen yaklaşımın konumlandığı çerçeve netleştirilmektedir.

Bu tez beş bölümden oluşmaktadır. Birinci bölümde çalışmanın konusu, amacı, önemi ve özgün yönleri ele alınarak araştırmanın genel kapsamı ortaya konulmuştur. İkinci bölümde, duygu analizi alanına ilişkin kavramsal çerçeve ve ilgili çalışmalar birlikte ele alınmıştır. Bu kapsamda, alanın kuramsal altyapısını oluşturan temel kavramlara yer verilmiş; makine öğrenmesi, derin öğrenme, doğal dil işleme, transformer tabanlı modeller ve çok modlu füzyon yaklaşımları açıklanmıştır. Ayrıca literatürde gerçekleştirilen tek modlu (ses, metin ve görüntü) ve çok modlu duygu analizi çalışmaları incelenerek mevcut yöntemlerin güçlü yönleri ve sınırlılıkları değerlendirilmiştir. Üçüncü bölümde, çalışmada izlenen yöntem ayrıntılı olarak açıklanmış; kullanılan veri seti, veri hazırlama ve ön işleme adımları, tercih edilen derin öğrenme mimarileri ve çoklu modalite birleştirme yaklaşımı sunulmuştur. Dördüncü bölümde gerçekleştirilen deneyler sonucunda elde edilen bulgular analiz edilmiştir. Son olarak beşinci bölümde, elde edilen sonuçlar literatür bağlamında tartışılarak çalışmanın genel değerlendirmesi yapılmış ve gelecekte yapılabilecek çalışmalara yönelik önerilere yer verilmiştir.

Bu bölümde çalışmanın amacı, kapsamı ve araştırma probleminin genel çerçevesi ortaya konulmuştur. Araştırmanın teorik altyapısını oluşturmak ve çalışmanın konumlandığı bilimsel bağlamı açıklamak amacıyla bir sonraki bölümde, duygu analizi alanına ilişkin temel kavramlar ile ilgili çalışmalar birlikte ele alınmaktadır.

2.KAVRAMSAL ÇERÇEVE VE İLGİLİ ÇALIŞMALAR

1.1. Makine öğrenmesi

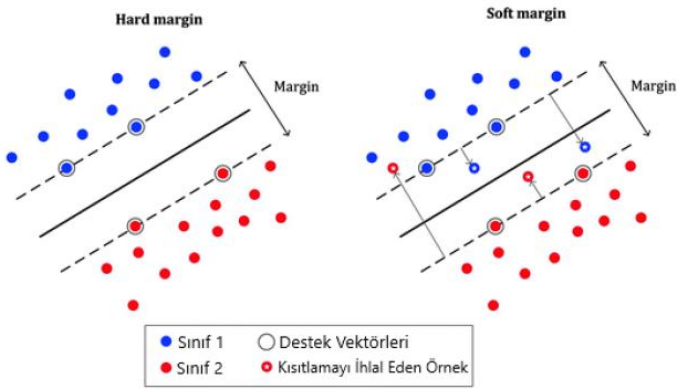
Makine öğrenmesi, bilgisayar sistemlerinin açıkça programlanmadan verilerden öğrenerek karar verme ve tahmin yapabilmesini sağlayan bir yapay zekâ yaklaşımıdır. Bu yöntem, verilerdeki örüntüleri öğrenerek yeni ve daha önce görülmemiş veriler üzerinde genelleme yapmayı amaçlamaktadır. (Goodfellow, 2016)

Makine öğrenmesi yaklaşımları genel olarak denetimli, denetimsiz ve pekiştirmeli öğrenme olmak üzere üç ana başlık altında incelenmektedir. Denetimli öğrenmede model, etiketli veriler kullanılarak eğitilmekte ve sınıflandırma veya regresyon gibi görevleri yerine getirmektedir. Denetimsiz öğrenmede ise etiketli veri bulunmamakta; model, verideki gizli yapıları ve benzerlikleri keşfetmeye çalışmaktadır. Pekiştirmeli öğrenmede ise bir ajan, çevresiyle etkileşim kurarak aldığı ödül ve cezalara bağlı olarak davranışlarını zamanla iyileştirmektedir. (Sutton, 1992)

2.1.1. Destek vektör makineleri (support vector machines- SVM)

1963 yılında Vladimir Vapnik ve Alexey Chervonenkis tarafından temelleri atılan Destek Vektör Makineleri (DVM), istatistiksel öğrenme kuramına dayalı bir gözetimli öğrenme yaklaşımıdır. DVM, iki farklı sınıfa ait örnekleri birbirinden en uygun biçimde ayıracak bir karar sınırının belirlenmesine dayanmaktadır. Bu ayrım, verileri temsil eden noktalara en uzak mesafeyi bırakan bir hiper düzlem aracılığıyla gerçekleştirmektedir.

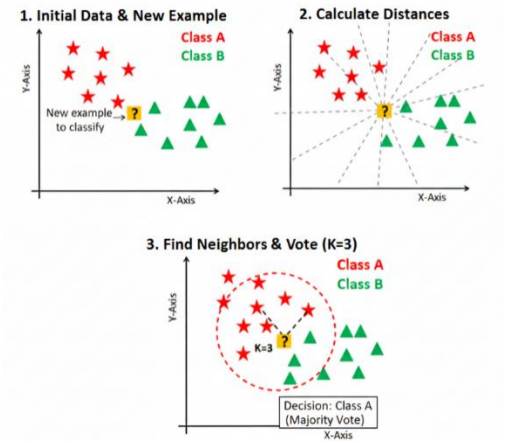
DVM'nin karar sınırını belirleyen temel bileşenler, hiper düzleme en yakın konumlanan ve sınıflandırma açısından en yüksek etkiye sahip olan destek vektörleridir. Bu yaklaşım, modelin yalnızca bu kritik örnekler üzerinden tanımlanmasına imkân vererek karar yüzeyinin daha kararlı bir biçimde oluşmasını sağlar (Burges, 1998). Doğrusal olarak ayrılamayan veri kümelerinde ise yöntem, çekirdek fonksiyonları aracılığıyla örnekleri daha yüksek boyutlu bir özellik uzayına dönüştürerek bu uzayda doğrusal bir ayrım yapılmasını mümkün kılar. Bu "kernel yöntemi", doğrusal olmayan örüntülerin temsilini kolaylaştırarak sınıflandırma performansını önemli ölçüde artırmaktadır. (Salcedo vd 2014)



Şekil 1: Destek Vektör Makineleri (SVM) için hard margin ve soft margin yaklaşımlarının gösterimi

2.1.2. K-en yakın komşu algoritması (k-nearest neighbors- knn)

K-en yakın komşuluk (KNN) algoritması, örnekler arasındaki benzerlik ilişkisine dayalı çalışan ve sınıflandırma ile regresyon problemlerinde yaygın biçimde kullanılan temel bir gözetimli öğrenme yöntemidir. Algoritmanın altında yatan temel varsayım, özellik uzayında birbirine yakın örneklerin benzer özellikler taşıdığı ve dolayısıyla aynı sınıfa ait olma olasılıklarının yüksek olduğudur. (Kramer, 2013) Bu yaklaşımda sınıfı bilinmeyen bir örnek, eğitim kümesindeki noktalarla olan uzaklıkları üzerinden değerlendirilir ve en yakın k komşusunun ait olduğu sınıflar dikkate alınarak çoğunluk kuralı ile etiketlenir. K değerinin seçimi, modelin karar yüzeyinin ne kadar ayrıntılı veya ne kadar genelleştirilmiş olacağını belirleyen önemli bir parametredir; çok küçük k değerleri gürültüye duyarlılığı artırırken, büyük k değerleri baskın sınıfın etkisini güçlendirebilmektedir. (Mucherino vd., 2009) KNN'in model varsayımı gerektirmeyen yapısı, yöntemi hem anlaşılır hem de farklı veri türlerine kolayca uygulanabilir hâle getirirken; karar mekanizmasının doğrudan veri dağılımıyla şekillenmesi, yöntemin sezgisel yönünü güçlendirmektedir.



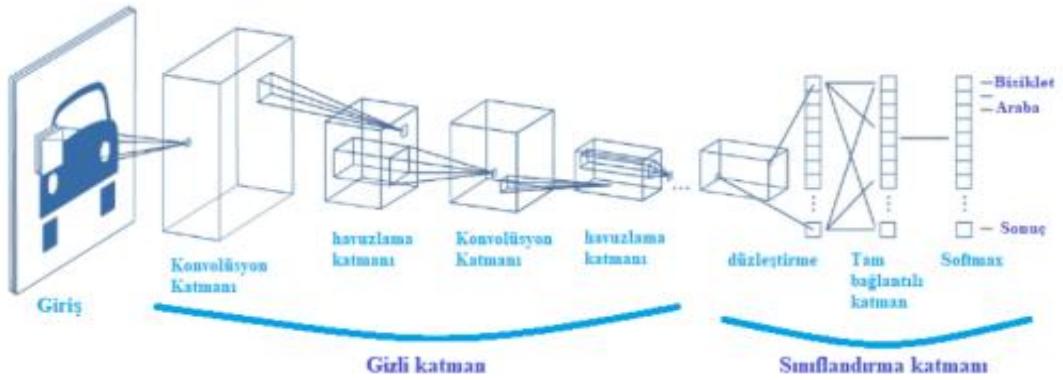
Şekil 2: K-En Yakın Komşu (KNN) algoritmasının çalışma prensibi: yeni örneğin en yakın k komşuya göre sınıflandırılması

2.2. Derin Öğrenme

Derin öğrenme (deep learning), makine öğrenmesinin bir alt dalı olup, verilerdeki karmaşık örüntüleri keşfetmek için çok katmanlı yapay sinir ağlarını kullanan bir yapay zekâ yöntemidir. İnsan beyninin öğrenme biçiminden ilham alan bu yaklaşım, sistemlerin ham verilerden (görüntü, ses, metin vb.) otomatik olarak özellikler çıkararak anlamlı sonuçlar üretmesini sağlar. Her katman, veriyi bir öncekinden daha soyut şekilde temsil eder; bu sayede düşük düzeydeki bilgiler (örneğin kenarlar veya tonlar) üst katmanlarda daha karmaşık yapılar (örneğin nesneler veya kelime anlamları) hâline gelir. Derin öğrenme, özellikle büyük veri ve yüksek hesaplama gücü sayesinde, insan müdahalesine gerek duymadan öğrenme, tanıma ve tahmin yapabilen sistemler geliştirmeye olanak tanır. (LeCun vd. 2015) (Goodfellow, 2016) (Sharifani vd. 2023)

2.1.3. Evrişimsel sinir ağı (convolutional neural network- cnn)

Konvolüsyon/Evrişimsel sinir ağları algoritmaları görüntü, ses işleme, doğal dil işleme, biyomedikal görüntü işleme gibi birçok alanda çalışmalar yapılmıştır ve bu çalışmalarda kullanılan birçok mimaride mevcuttur. Son zamanlarda kullanılan bazı mimariler; LeNet, AlexNet, VGGNet, GoogleNet sayabiliriz. (Erdoğan, 2021)



Şekil 3: CNN'in katman yapısı

Evrişimsel sinir ağlarının birçok katmanı vardır. Verilerimiz her katmanda işlenerek istediğimiz sonucu elde ederiz.

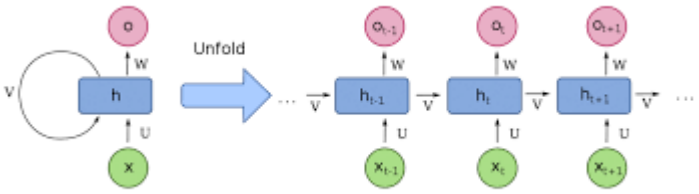
- Convolutional layer- Ana görüntü üzerinde özellikleri çıkarmak için kullanılır.
- Non-Linearity layer- ReLU katmanı – Aktivasyon işleminin bulunduğu katmandır. Sistemde doğrusal olmayan (non-linearity) bir fonksiyon kullanır.

- Pooling (Downsampling) layer-Ağırlık sayısı azaltma ve uygunluk kontrolü yapılır.
- Flattening layer – Klasik Sinir Ağı için tek boyutlu vektör verilerini hazırlar.
- Fully-Connected layer – Sınıflandırmada kullanılan Standart Sinir Ağı.

2.1.4. Tekrarlayan sinir ağı (recurrent neural network- rnn)

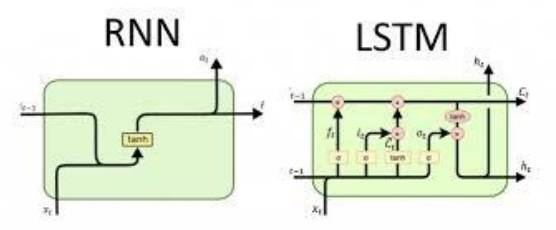
RNN'ler bir yapay sinir ağı (YSA) çeşitlidir. Klasik YSA'lar ileri beslemeli bir yöntem kullanır. RNN'ler ise bir önceki çıkışındaki bilgiyi hafızasında tutup bir sonraki adımın girişi olarak kullanılmasıdır yani geri beslemeli bir yöntem kullanılır. RNN'ler doğal dil işleme, konuşma tanıma yüz tanıma gibi uygulamalarda sıklıkla kullanılır.

Geleneksel sinir ağları, sadece anlık girdi bilgilerini kullanarak nesneleri tanımlar ve önceden sisteme gönderilen bilgiler kullanılmadan tahmin yapar. Bununla birlikte, RNN'ler anlık verilerin yanı sıra daha önce sisteme giren veya daha sonra yüklenen bilgileri de kullanır. Bu sayede RNN'ler, geçmiş ve şu anki verileri bir araya getirerek sonuçlara varabilirler. Bu özellikleriyle RNN'ler, hafızalı ağlar olarak bilinir. (Şenel, 2023)



Şekil 4: RNN yapısının şematik gösterimi

Tekrarlayan sinir ağı, zaman içinde birçok farklı yineleme gerçekleştirebilir. Bu yinelemeler, veriler üzerinde çok az etkiye sahip olduğundan, verilerle ilişkili bazı parametrelerin sistemden kaldırılmasına neden olabilir. Bu durumda, uzun zaman önce eklenen bilgilere erişim kaybedilebilir. Bu tür atılan ve unutulan bilgileri toplamak için geliştirilmiş bir mimari vardır. LSTM, bu mimarinin bir örneğidir. (Cho, 2014)

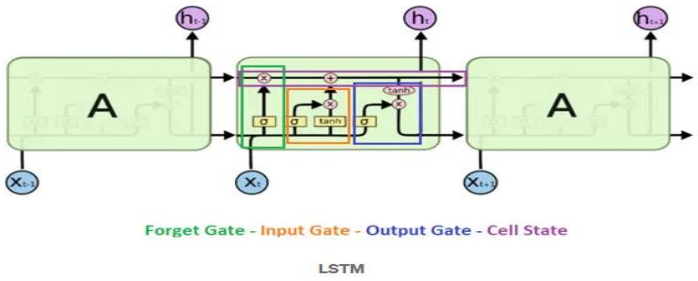


Şekil 5: RNN ve LSTM'nin karşılaştırmalı gösterimi

2.1.5. Uzun kısa süreli bellek ağı (long short-term memory- lstm)

Bir önceki sonuçları rastgele hatırlayan bir RNN mimarisidir. RNN'lerin eksiği olan uzun vadeli bağlılık problemini çözmek için geliştirilmiştir. LSTM, bellek hücreleri, giriş, çıkış ve hatırlama kapıları sayesinde bilgi akışını kontrol eder.

- Forget Gate (Unutma Kapısı): LTM bu kapıya gelir ve yararlı olmayan bilgileri unutur.
- Learn Gate (Öğrenme Kapısı): Yeni Öğrenilen bilgilerin mevcut girdiye uygulanabilmesi için mevcut girdi ve STM birleştirilir.
- Remember Gate (Hatırlama Kapısı): Kalan LTM bilgileri ile STM ve mevcut girdi bu kapıda birleştirilir.
- Use Gate (Kullanma Kapısı): Mevcut olayın çıktısının tahmin edildiği kapı (Onan, 2022)



Şekil 6: LSTM hücresinin şematik diyagramı

2.1.6. Çok katmanlı algılayıcı (multilayer perceptron – mlp)

Çok Katmanlı Algılayıcı (Multilayer Perceptron – MLP), ileri beslemeli yapay sinir ağları sınıfına ait, doğrusal olmayan karmaşık fonksiyonların modellenmesinde yaygın olarak kullanılan temel bir derin öğrenme mimarisidir. MLP, biyolojik sinir sisteminden esinlenerek geliştirilen yapay sinir ağlarının en temel ve en yaygın biçimlerinden biri olup, sınıflandırma, regresyon ve örüntü tanıma problemlerinde etkin biçimde kullanılmaktadır. (Desai vd., 2021).

2.2.4.1. Mlp mimari yapısı

Bir MLP modeli temel olarak üç ana katmandan oluşmaktadır:

- **Girdi katmanı (input layer):**

Sisteme verilen ham veya ön işlenmiş özellik vektörlerini alır. Bu katmandaki nöronlar yalnızca veri iletiminden sorumludur ve herhangi bir hesaplama gerçekleştirmez.

- **Gizli katman (hidden layer):**

Girdi verisi üzerinde doğrusal olmayan dönüşümlerin gerçekleştirildiği katmanlardır. Bir MLP, probleme bağlı olarak bir veya birden fazla gizli katman içerebilir. Gizli katmanların varlığı, ağın karmaşık ve doğrusal olmayan ilişkileri öğrenebilmesini sağlar. (Zare vd. 2013)

- **Çıkış katmanı (output layer):**

Modelin nihai çıktısını üretir. Çıkış katmanındaki nöron sayısı, problemin türüne (ikili sınıflandırma, çok sınıflı sınıflandırma veya regresyon) göre belirlenir.

Bu yapı sayesinde MLP, giriş uzayındaki verileri daha yüksek boyutlu ve soyut temsil uzaylarına dönüştürerek öğrenme sürecini gerçekleştirir.

2.2.4.2. Nöron modeli ve matematiksel gösterim

Bir MLP'de her bir nöron, kendisine bağlı önceki katmandaki nöronlardan gelen sinyalleri ağırlıklandırarak toplar ve bu toplamı bir aktivasyon fonksiyonundan geçirir. Gizli katmandaki bir nöronun çıktısı matematiksel olarak aşağıdaki şekilde ifade edilir:

$$net_j = \sum_{i=1}^{n} w_{ij} x_i$$

$$y_j = f(net_j)$$

Burada $x_i$giriş değerlerini, w_{ij}bağlantı ağırlıklarını ve $f(\cdot)$aktivasyon fonksiyonunu temsil etmektedir.

Aktivasyon fonksiyonu, ağın doğrusal olmayan ilişkileri öğrenebilmesini sağlayan kritik bir bileşendir. MLP mimarilerinde yaygın olarak sigmoid, hiperbolik tanjant (tanh) ve ReLU gibi

fonksiyonlar kullanılmaktadır. Özellikle tanh ve sigmoid fonksiyonları, klasik MLP modellerinde sık tercih edilmektedir. (Desai vd., 2021)

2.2.4.3. İleri yayılım (forward propagation)

MLP'de öğrenme süreci, ileri yayılım adımıyla başlar. Bu aşamada giriş verileri sırasıyla gizli katmanlara ve çıkış katmanına iletilerek ağın tahmini çıktısı hesaplanır. Her katmanda, önce ağırlıklı toplam alınır, ardından aktivasyon fonksiyonu uygulanarak bir sonraki katmana aktarılacak çıktı üretilir.

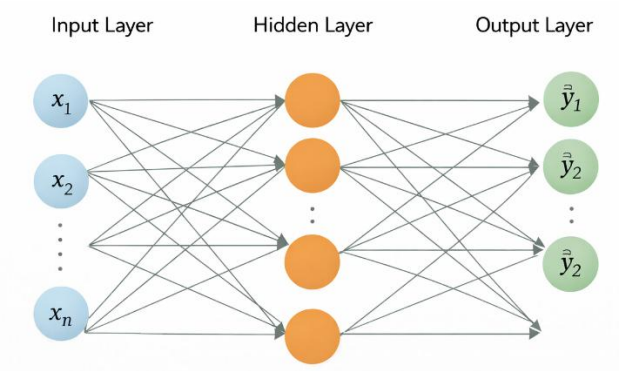
Bu işlem, giriş katmanından çıkış katmanına kadar tek yönlü olarak ilerler ve ağın mevcut ağırlık değerlerine göre bir çıktı oluşturulur.

2.2.4.4. Geri yayılım ve öğrenme süreci

MLP modellerinde öğrenme, geri yayılım (backpropagation) algoritması ile gerçekleştirilir. Geri yayılım sürecinde, ağın ürettiği çıktı ile gerçek hedef değer arasındaki hata hesaplanır ve bu hata, zincir kuralı kullanılarak katmanlar boyunca geriye doğru yayılır. (Zare vd. 2013)
Hata fonksiyonu genellikle ortalama karesel hata (MSE) veya çapraz entropi kaybı şeklinde tanımlanır. Ağırlık güncelleme işlemi aşağıdaki genel form ile ifade edilebilir:

$$w_{ij}^{new} = w_{ij}^{old} - \eta \frac{\partial E}{\partial w_{ij}}$$

Burada ηöğrenme oranını, E ise hata fonksiyonunu temsil etmektedir. Bu iteratif güncelleme süreci, hata belirli bir eşik değerin altına düşene kadar devam eder.



Şekil 7: Çok Katmanlı Algılayıcı (MLP) mimarisi

2.3. Doğal Dil İşleme (Natural Language Processing -NLP)

Doğal dil, yüzeysel biçiminden bağımsız olarak çok katmanlı bir yapıya sahiptir ve NLP uygulamalarında bu katmanların her biri dilin farklı bir yönünü temsil eder. Doğal dil işleme, insan dili ile bilgisayarlar arasındaki etkileşimi inceleyen disiplinler arası bir araştırma alanıdır. Doğal dil, yüzeysel sözcük dizilimlerinden bağımsız olarak çok katmanlı bir yapıya sahiptir ve bu yapı, dilin bilgisayar ortamında modellenmesini karmaşık hale getirmektedir. NLP uygulamalarında dilsel çözümleme genellikle biçimbirimsel (morphology), sözdizimsel (syntax), anlamsal (semantics) ve bağlamsal/pragmatik (pragmatics) düzeylerde ele alınmaktadır.

Biçimbirimsel analiz, kelimelerin kök, ek ve yapı bakımından ayrıştırılmasını kapsarken; sözdizimsel analiz, cümle içerisindeki kelimeler arasındaki gramatik ilişkilerin belirlenmesine odaklanmaktadır. Anlamsal analiz, kelime ve cümle düzeyinde anlam ilişkilerinin modellenmesini amaçlamakta; bağlamsal ve pragmatik analiz ise ifadelerin kullanım bağlamına göre kazandığı anlamları inceleyerek dilin işlevsel yönünü ele almaktadır. Bu çok katmanlı yapı, doğal dilin yalnızca sözdizimsel kurallarla temsil edilmesini yetersiz kılmakta ve veri temelli öğrenme yaklaşımlarını zorunlu hale getirmektedir.

Son yıllarda derin öğrenme tabanlı yöntemlerin gelişmesiyle birlikte, NLP alanında bağlama duyarlı (context-aware) dil temsilleri üretebilen modeller ön plana çıkmıştır. Özellikle Transformer mimarisi üzerine inşa edilen ön-eğitimli dil modelleri, dilin sözdizimsel ve anlamsal yapısını büyük ölçekli metin verileri üzerinden öğrenerek birçok NLP görevinde üstün performans sağlamıştır.

Transformer mimarisi, dizisel verilerdeki bağımlılıkları modellemek amacıyla geliştirilen ve öz-dikkat (self-attention) mekanizmasına dayanan bir yapıdır. Bu mimari, sıralı işlem gereksinimini ortadan kaldırarak paralel hesaplamaya olanak tanımakta ve uzun menzilli bağımlılıkların etkili biçimde öğrenilmesini sağlamaktadır.

Transformer tabanlı dil modelleri, büyük ölçekli metin derlemeleri üzerinde ön-eğitim (pretraining) aşamasında dilin genel yapısını öğrenmekte; ardından belirli görevler için ince ayar (fine-tuning) ile uyarlanabilmektedir. Bu yaklaşım, özellikle sınıflandırma, duygu analizi, soru-cevap ve metin anlama gibi görevlerde önemli başarılar sağlamıştır.

2.3.1. BERT (Bidirectional Encoder Representations from Transformers)

Çift Yönlü Kodlayıcı Temsilleri (Bidirectional Encoder Representations from Transformers), transformer mimarisine dayalı, ön-eğitimli bir dil temsil modelidir. BERT'in temel amacı, doğal dildeki bağlamsal ilişkileri daha etkili biçimde modelleyebilmek için kelimelerin yalnızca önceki ya da sonraki bağlamlarını değil, her iki yönlü bağlamı aynı anda dikkate alan temsiller öğrenmektir. Bu yönüyle BERT, önceki tek yönlü veya sınırlı çift yönlü dil modellerinden ayrılmaktadır.

Model, büyük ölçekli etiketlenmemiş metinler üzerinde ön-eğitim (pretraining) aşamasından geçirilmekte ve bu süreçte dilin anlamsal ve sözdizimsel yapısını derinlemesine öğrenmektedir. Ön-eğitimin ardından BERT, soru cevaplama, metin sınıflandırma, doğal dil çıkarımı ve duygu analizi gibi birçok doğal dil işleme görevinde, yalnızca küçük bir çıktı katmanı eklenerek ince ayar (fine-tuning) yapılabilen esnek bir yapı sunmaktadır. Bu özellik, göreve özgü karmaşık mimari değişikliklerine duyulan ihtiyacı ortadan kaldırmakta ve BERT'in farklı uygulama alanlarında yüksek başarı elde etmesini sağlamaktadır.

2.3.2. RoBERTa (Robustly Optimized BERT Pretraining Approach)

RoBERTa (Robustly Optimized BERT Approach), BERT modelinin Transformer tabanlı encoder mimarisini temel alan bir dil temsil modelidir. RoBERTa'da mimari yapı büyük ölçüde korunmuş; buna karşılık modelin ön-eğitim sürecine ilişkin bazı tasarım tercihleri yeniden ele alınmıştır. Bu yaklaşım, bağlamsal dil temsillerinin daha etkin biçimde öğrenilmesini amaçlamaktadır (Liu vd., 2019).

RoBERTa'nın temel farkı, ön-eğitim aşamasında kullanılan eğitim stratejilerinde ortaya çıkmaktadır. Model, daha uzun eğitim süresi, daha büyük mini-batch boyutları ve daha geniş metin derlemeleri kullanılarak eğitilmektedir. Ayrıca, BERT'te yer alan bazı ön-eğitim bileşenleri RoBERTa yaklaşımında sadeleştirilmiş ve maskeleme işlemi eğitim süreci boyunca dinamik olarak uygulanmıştır. Bu düzenlemeler, Transformer encoder yapısının bağlamsal bilgiyi daha etkili şekilde modellemesine olanak sağlamaktadır (Liu vd., 2019).

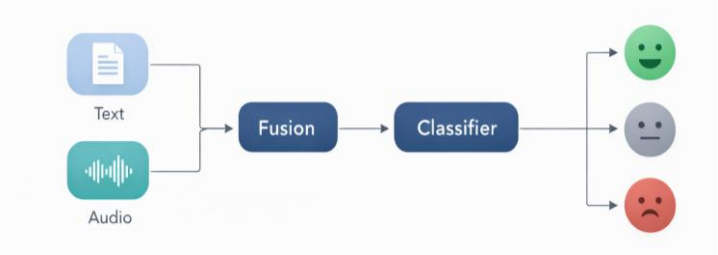
2.4. Füzyon Yaklaşımları

Füzyon, farklı kaynaklardan elde edilen verilerin birleştirilerek tek bir sistemin daha doğru, güvenilir ve bütüncül karar verebilmesini sağlayan bir bilgi bütünleştirme sürecidir. Veri füzyonu; çoklu algılayıcı füzyonu, bilgi füzyonu veya gözlem sentezi gibi çeşitli terimlerle de açıklanmakta olup, temel amaç farklı kaynaklardaki bilgilerin tek bir temsil altında daha anlamlı hâle getirilmesidir. (Biroğul vd., 2007). Bu süreç; tek bir kaynağın sağlayamayacağı kapsam, doğruluk ve belirsizlik azaltma avantajları sunar ve insanın birbiriyle ilişkili duyusal bilgileri birleştirerek dünyayı anlamlandırma biçimiyle benzerlik taşır.

Modern yapay zekâ uygulamalarında füzyon, özellikle çok modlu sistemlerde (metin–ses–görüntü) modelin farklı veri türleri arasındaki ilişkileri daha iyi öğrenmesini sağlayarak performansın kayda değer şekilde artmasına katkıda bulunur.

2.4.1. Erken füzyon (early fusion)

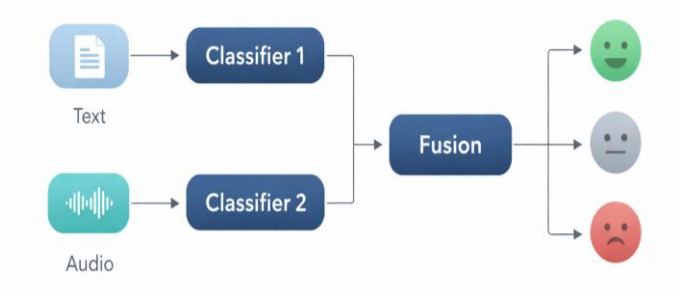
Erken füzyon, farklı modalitelerden elde edilen verilerin veya çıkarılmış özelliklerin modelin ilk aşamalarında birleştirildiği yaklaşımdır. Bu yöntemde her bir modaliteye ait özellikler ortak bir temsil uzayında bir araya getirilerek tek bir girdi vektörü oluşturulur. Model, bu birleşik temsili kullanarak öğrenme sürecini tüm modaliteleri aynı anda dikkate alacak şekilde gerçekleştirir. Erken füzyonda, modaliteler arasındaki ilişkiler doğrudan özellik seviyesinde ele alınır.



Şekil 8: Erken füzyon yaklaşımının şematik gösterimi

2.4.2. Geç füzyon (late fusion)

Geç füzyon, her bir modalitenin bağımsız olarak işlendiği ve birleştirmenin karar aşamasında gerçekleştirildiği yaklaşımdır. Bu yöntemde her modalite için ayrı bir model veya işlem hattı bulunur ve her biri kendi çıktısını üretir. Elde edilen çıktılar, belirli bir birleştirme kuralı doğrultusunda bir araya getirilerek nihai karar oluşturulur. Geç füzyon, modalitelerin birbirinden bağımsız şekilde değerlendirilmesine olanak tanır.



Şekil 9: Geç füzyon yaklaşımının şematik gösterimi

2.5. Performans Değerlendirme Metrikleri

Performans metrikleri, bir makine öğrenmesi veya derin öğrenme modelinin ürettiği tahminlerin doğruluğunu, tutarlılığını ve sınıflandırma başarısını nicel olarak değerlendirmek amacıyla kullanılan ölçütlerdir. Bu metrikler, modelin gerçek veriye ne kadar uygun sonuçlar üretebildiğini belirlemek için tahminler ile gerçek etiketler arasındaki ilişkiyi analiz ederek modelin başarısını ve hata eğilimlerini ortaya koyar. Özellikle sınıflandırma problemlerinde kullanılan metrikler, modelin yaptığı doğru ve hatalı tahminlerin türünü dikkate alarak performansın daha ayrıntılı biçimde yorumlanmasını sağlar.

Bu kapsamda performans değerlendirmesinde kullanılan temel kavramlar aşağıda kısaca tanımlanmıştır:

True Positive (TP): Pozitif sınıfa ait olup doğru tahmin edilen örnekler
True Negative (TN): Negatif sınıfa ait olup doğru tahmin edilen örnekler
False Positive (FP): Negatif olup pozitif olarak tahmin edilen örnekler
False Negative (FN): Pozitif olup negatif olarak tahmin edilen örnekler

2.5.1. Doğruluk (Accuracy)

Doğruluk (accuracy), modelin tüm sınıflar için yaptığı doğru tahminlerin toplam tahmin sayısına oranını ifade eden temel bir performans ölçütüdür. Sınıf dağılımının dengeli olduğu durumlarda modelin genel başarısını yansıtmak için kullanılabilir.

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})}$$

Bununla birlikte, sınıf dağılımının dengesiz olduğu veri setlerinde doğruluk metriği, baskın sınıfların etkisi altında kalarak model performansını olduğundan yüksek gösterebilmektedir. Bu nedenle doğruluk metriği, tek başına kullanıldığında özellikle duygu sınıflandırma gibi dengesiz sınıf yapısına sahip problemlerde sınırlı bir değerlendirme sunmaktadır.

2.5.2. Kesinlik (Precision)

Kesinlik (precision), model tarafından pozitif olarak tahmin edilen örneklerin ne kadarının gerçekte pozitif olduğunu gösteren bir performans ölçütüdür. Özellikle yanlış pozitif tahminlerin maliyetinin yüksek olduğu uygulamalarda model başarısının değerlendirilmesinde önemli bir rol oynamaktadır.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Bununla birlikte, kesinlik metriği tek başına kullanıldığında modelin gerçek pozitifleri ne ölçüde yakalayabildiği hakkında bilgi vermemektedir. Bu nedenle, daha dengeli bir performans değerlendirmesi elde edebilmek için genellikle duyarlılık (recall) metriği ile ele alınmaktadır.

2.5.3. Duyarlılık (Recall / Sensitivity)

Duyarlılık (recall), gerçek pozitif örneklerin ne kadarının model tarafından doğru biçimde tespit edildiğini gösteren bir performans ölçütüdür. Pozitif sınıfın gözden kaçırılmasının kritik olduğu problemlerde model başarısının değerlendirilmesinde önemli bir rol oynamaktadır.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

False Negative (FN) değerinin artması, modelin pozitif örnekleri kaçırdığı anlamına gelmekte ve bu durum duyarlılık değerinin düşmesine neden olmaktadır. Ancak recall metriği tek başına kullanıldığında, yanlış pozitif tahminleri dikkate almadığı için modelin genel performansını tam olarak yansıtmayabilmektedir.

2.5.4. F1-Skoru (F1-Score)

Kesinlik (precision) ve duyarlılık (recall) metrikleri, bazı durumlarda birbirine zıt yönlerde değişebilmektedir. F1-skoru, bu iki metriğin harmonik ortalamasını alarak model performansını tek bir ölçüt altında dengeli biçimde değerlendirmeyi amaçlamaktadır. Bu yönüyle F1-skoru, özellikle sınıf dağılımının dengesiz olduğu sınıflandırma problemlerinde doğruluk metriğine kıyasla daha anlamlı bir performans göstergesi sunmaktadır.

$$\text{F1}_\text{i} = \frac{2 \cdot \text{Precision}_\text{i} \cdot \text{Recall}_\text{i}}{\text{Precision}_\text{i} + \text{Recall}_\text{i}}$$

Harmonik ortalama yapısı gereği F1-skoru hem kesinlik hem de duyarlılık değerlerinin yüksek olmasını gerektirmekte; bu da modelin yalnızca bir ölçüt üzerinde değil, her iki açıdan da tutarlı bir performans sergilemesini zorunlu kılmaktadır.

2.5.5. Weighted F1 skoru

Weighted F1 (WF1) skoru, her bir sınıf için hesaplanan F1 değerinin, ilgili sınıfa ait örnek sayısı (support) ile ağırlıklandırılmasıyla elde edilen bir performans metriğidir. Bu yaklaşım, sınıf dağılımının dengesiz olduğu veri setlerinde, model performansının daha gerçekçi ve dengeli bir biçimde değerlendirilmesini sağlamaktadır.

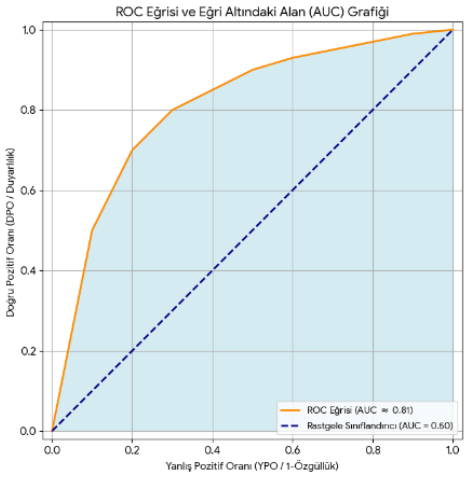
Öncelikle her bir sınıf için F1 skoru, kesinlik (precision) ve duyarlılık (recall) değerlerinin harmonik ortalaması olarak hesaplanır. Daha sonra Weighted F1 skoru, sınıf bazlı F1 değerlerinin, ilgili sınıflara ait örnek sayıları ile ağırlıklandırılması yoluyla aşağıdaki şekilde elde edilir:

$$WF1 = \sum_{i=1}^{C} \frac{N_i}{N} \cdot F1_i$$

2.5.6. Roc eğrisi ve Auc (receiver operating characteristic / area under curve)

ROC eğrisi, farklı eşik (threshold) değerleri altında modelin doğru pozitif oranı (True Positive Rate, TPR) ile yanlış pozitif oranı (False Positive Rate, FPR) arasındaki ilişkiyi göstermektedir. Eğri altında kalan alan olan AUC (Area Under Curve), modelin pozitif ve negatif sınıfları ayırt edebilme yeteneğini özetleyen bir ölçüt olarak kullanılmaktadır.

Bu çalışmada ROC eğrisi ve AUC değerleri, özellikle sınıf bazlı (one-vs-rest) değerlendirmelerde model davranışını görsel olarak incelemek amacıyla kullanılmıştır.

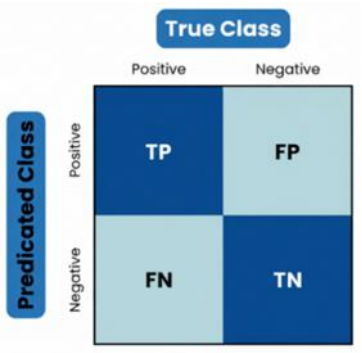


Şekil 10: ROC eğrisi örnek gösterimi

2.5.7. Karışıklık matrisi (confusion matrix)

Karışıklık matrisi, bir sınıflandırma modelinin yaptığı doğru ve hatalı tahminleri sınıf bazında gösteren temel bir analiz aracıdır. Bu yapı, modelin hangi sınıflarda daha başarılı olduğunu ve hangi tür hatalara daha yatkın olduğunu açık biçimde ortaya koymaktadır.

Bu çalışmada karışıklık matrisi, elde edilen sonuçların yorumlanmasını desteklemek ve sınıf bazlı hata dağılımlarını incelemek amacıyla kullanılmıştır.



Şekil 11. Karışıklık matrisi örnek gösterimi

2.6. Literatür Araştırması

2.6.1. Duygu Analizi

Duygu analizi, insan–bilgisayar etkileşiminin geliştirilmesinde temel bir araştırma alanı olarak öne çıkmaktadır. Günümüzde insansı robotların daha doğal tepkiler verebilmesi, e-ticarette müşteri memnuniyetinin artırılması, sağlıkta hasta deneyimlerinin izlenmesi, eğitimde öğrencilerin derse katılımının güçlendirilmesi ve finans sektöründe yatırım eğilimlerinin incelenmesi gibi pek çok alanda duygu analizinden yararlanılmaktadır. Farklı sektörlerde duygu bilgisinin doğru şekilde çıkarılması, karar verme süreçlerinin hızlanmasına ve daha kişiselleştirilmiş hizmetlerin sunulmasına katkı sağlamaktadır. Son yıllarda dijital ortamda üretilen veri miktarının artmasıyla birlikte, duyguların otomatik olarak çözümlenmesine yönelik yöntemlere duyulan ihtiyaç belirginleşmiş ve bu eğilim duyguların sistemli biçimde analiz edilmesine yönelik çalışmaları desteklemiştir.

2.6.2. Tek Modlu Duygu Analizi

2.6.1.1. Ses

Casale vd. (2008), konuşma tabanlı duygu sınıflandırması için EMO-DB ve SUSAS veri setlerini kullanarak makine öğrenmesi temelli bir çalışma gerçekleştirmiş. Araştırmada ETSI ES 202 211 standardına dayalı bir ön işlem süreci uygulanmış ve her konuşma örneğinden 3809

istatistiksel özellik çıkarılmış. EMO-DB veri setinde öfke, bıkkınlık, iğrenme, korku, mutluluk, üzüntü ve nötr olmak üzere yedi duygu sınıfı değerlendirilmiş. Çalışmada SVM ve SMO sınıflayıcıları kullanılmış; özellik seçimi ve normalizasyon–discretization işlemlerinin ardından en yüksek doğruluk oranı %92 olarak raporlanmış. (Casale vd., 2008)

Yun vd. (2011) konuşma duygu sınıflandırması için Large-Margin Gaussian Mixture Models (LM-GMM) tabanlı bir yöntem kullanmışlar. Çalışmada MFCC, log enerji, perde (pitch) ve sıfır geçiş oranı gibi akustik özellikler ile bu özelliklerin delta ve ivme katsayılarından oluşan vektörler kullanılmış. Yöntem EMO-DB, SUSAS, DES ve VAM veri setleri üzerinde değerlendirilmiş ve EMO-DB veri seti için %87,80 doğruluk oranı elde edilmiş. (Yun vd., 2011)

Gumelar ve arkadaşları (2019), konuşma duygu tanıma amacıyla prosodik ve spektral özellikleri kullanarak derin sinir ağları tabanlı bir yapı oluşturmuş. Çalışmada RAVDESS veri setinden elde edilen toplam 4320 ses örneği kullanılmış; MFCC, perde (pitch), enerji ve süre gibi özellikler libROSA kütüphanesi ile çıkarılmış. Elde edilen özellikler çok katmanlı bir derin sinir ağına (DNN) ve alternatif olarak bir CNN modeline giriş olarak verilmiş, sekiz duygu sınıfı üzerinde sınıflandırma yapılmış. Çalışmada DNN tabanlı model için %78,83 doğruluk değeri elde edilmiş. (Gumelar vd., 2019)

Donatus vd. (2024) konuşma sinyallerinden duygu tespiti için spektrogram ve MFCC (Mel-Frekans Kepstrum Katsayıları) temsilleri karşılaştırılmıştır. Farklı makine öğrenimi algoritmaları (SVM, kNN, Random Forest) üzerinde yapılan deneyler yapılmış ve en yüksek SVM+Spektogram ile %54 doğruluk oranı elde edilmiş. (Donatus vd., 2024)

Mustaqeem vd. (2020), konuşma tabanlı duygu tanıma (SER) için kümeleme tabanlı bir derin öğrenme yaklaşımı önermiştir. Çalışmada, RBF (Radial Basis Function) tabanlı K-ortalamalar kümeleme yöntemiyle belirlenen “anahtar ses segmentleri” kullanılarak işlem maliyeti azaltılmıştır. Seçilen segmentler STFT ile spektrograma dönüştürülmüş, ResNet101 ile öznitelikler çıkarılmış ve derin BiLSTM ağıyla zamansal ilişkiler öğrenilmiştir. Model, IEMOCAP (%72,25), EMO-DB (%85,57) ve RAVDESS (%77,02) veri setlerinde yüksek doğruluk elde ederek mevcut yöntemleri geride bırakmıştır. (Mustaqeem vd., 2020)

2.6.1.2. Metin

Zhuang ve vd (2025), uzun metinli çevrim içi yorumlarda duygu analizini geliştirmek amacıyla ATNPWC_BiLSTM adını verdikleri dikkat mekanizmalı ve duygu ağırlıklı Word2Vec tabanlı bir model önermiştir. Önerilen yaklaşımda, kısa menzilli semantik ve duygu bilgisi sentiment-weighted Word2Vec, uzun menzilli bağlamsal ve derin özellikler ise attention-based BiLSTM ile öğrenilmiş ve bu temsiller birleştirilmiştir. Otel ve e-ticaret yorumları üzerinde yapılan deneylerde model, e-ticarette %99 doğruluk ve %98 F1 skoru, otel yorumlarında ise %96 doğruluk ve %97 F1 skoru elde edilmiş. (Zhuang vd., 2025)

Sun ve vd (2020), metin duygu analizinde önemli özelliklerin çıkarılamaması sorununu çözmek amacıyla IMDB ve DouBan veri setleri için CNN-BiLSTM-Attention tabanlı bir model önermiştir. Bu modelde CNN ile yerel özellikler, BiLSTM ile bağlamsal (küresel) anlamsal özellikler elde edilmekte, dikkat (attention) katmanı ise önemli kelimelere daha fazla ağırlık vererek sınıflandırma doğruluğunu artırmaktadır. Çalışma sonuçlarına göre önerilen model, IMDB için %90,67 DouBan için %92,45 doğruluk oranlarını elde etmiştir. (Sun vd., 2020)

Xie vd. (2024), eğitim alanına özgü metinlerde duygu analizi yapabilmek amacıyla BERT ve FastText modellerini birleştiren hibrit bir yaklaşım önermiştir. Çalışmada, BERT'in bağlamsal anlam yakalama gücü ile FastText'in yüksek hız ve verimliliği birleştirilmiş, iki modelin özellikleri RCNN sınıflayıcı üzerinde füzyon edilmiştir. Sonuçlar, önerilen BERT–FastText modelinin doğruluk oranı %88 gelmiştir. (Xie vd., 2024)

Chen (2024), tıbbi metinlerde duygu analizinin doğruluğunu artırmak amacıyla BERT tabanlı bir model önermiş ve çıkış katmanında CNN, FCN ve GCN gibi derin öğrenme mimarilerini karşılaştırmıştır. METS-CoV veri seti üzerinde yapılan deneylerde, BERT ile CNN kombinasyonu %72,97 doğruluk elde ederek küçük ölçekli tıbbi veri kümelerinde en etkili yaklaşım olarak belirlenmiştir. Bu çalışma, tıp alanında metin tabanlı duygu analizinde uygun model seçiminin performans üzerindeki belirleyici rolünü vurgulamaktadır. (Chen, 2024)

Mekala ve vd. (2023), sosyal medya metinlerinde duygu analizinin doğruluğunu artırmak için doğal dil işleme (NLP) ve derin öğrenme tabanlı bir yaklaşım geliştirmiştir. Çalışmada BERT ve Bi-GRU gibi modeller kullanılarak kısa, orta ve uzun metinler üzerinde karşılaştırmalar yapılmış, önerilen yöntemin kısa metinlerde %97,1 doğruluk oranına ulaştığı görülmüştür. (Mekala vd., 2023)

Fujihira vd. (2020), çok dilli web metinlerinde duygu analizi yapmak amacıyla kelime kelime çeviri (word-to-word translation) temelli bir yöntem önermiştir. Çalışmada, morfolojik çözümleme, duygu sözlüğü ve kelime vektör modelleri (fastText) kullanılarak İngilizce, Almanca, Fransızca ve İspanyolca tweetler üzerinde sınıflandırma yapılmıştır. Sonuçlar sırasıyla %54, %54, %60 ve %64 doğruluk oranları elde edilmiş. (Fujihira vd., 2020)

Zhan (2024), üniversite öğrencilerinin ruh sağlığını izlemek amacıyla sosyal metin duygu analizi temelli bir psikolojik sağlık izleme sistemi geliştirmiştir. Çalışmada, LSTM tabanlı derin öğrenme modeli ve özel olarak oluşturulmuş bir duygu sözlüğü kullanılarak öğrencilerin sosyal platformlarda paylaştığı metinler gerçek zamanlı olarak analiz edilmiş. Sonuçlar, LSTM modelinde %97,6 doğruluk ve daha kısa tepki süresi sağladığını göstermiştir. (Zhan, 2024)

Le (2020), metin tabanlı duygu analizinde doğruluğu artırmak amacıyla SADL (Sentiment Attention-based Deep Learning) adını verdiği dikkat mekanizmalı bir derin öğrenme modeli önerilmiş. Bu modelde, Word2Vec ve duygu sözlükleri (SentiWordNet, Liu Lexicon) birlikte kullanılarak hem anlamsal hem duygusal özellikler öğrenilmiş. BiLSTM ve attention mekanizması içeren mimari, Amazon, IMDb ve Yelp veri setlerinde ortalama %86,7 doğruluk elde ederek klasik gömülü modellerden (Word2Vec, GloVe) daha yüksek performans göstermiş. (Le, 2020)

Omurca ve vd. (2017), Türkçe dilinde duygu analizi çalışmalarına kaynak oluşturmak amacıyla cümle düzeyinde açıklamalı (annotated) bir duygu veri kümesi geliştirmiştir. Çalışmada, Zemberek kütüphanesi kullanılarak kelimelerin kökleri, türleri ve kutupları (pozitif, negatif, nötr) belirlenmiş; otel yorumlarından oluşan 1000 inceleme ve 5364 cümle etiketlenmiştir. Elde edilen veri kümesi, JSON formatında hazırlanarak aspect-based sentiment analysis çalışmalarında kullanılabilecek ilk Türkçe kaynaklardan biri haline getirilmiştir. (Omurca vd., 2017)

S. Mian Qaisar (2020) ise çevrim içi film yorumları üzerinde metin tabanlı duygu analizi gerçekleştirmiştir. Çalışmada LSTM (Long Short-Term Memory) modeli kullanılmış, yorumlar pozitif ve negatif olarak iki sınıfa ayrılmıştır. Analiz sonucunda modelin duygu sınıflandırma başarısı %89,9 olarak rapor edilmiştir (Qaisar S. , 2020).

2.6.1.3. Görüntü

2021 yılında Zbancioc ve arkadaşları, derin öğrenme yöntemleri kullanarak yüz ifadelerinden duygu analizi üzerine bir çalışma gerçekleştirmiştir. Bu çalışmada, CNN (Convolutional Neural Network) yöntemi kullanılarak MFSC veri seti üzerinde sınıflandırma yapılmış ve model %85 doğruluk oranı elde etmiştir. (Zbancioc vd., 2021)

2022 yılında Smirnov ve arkadaşları, yüz ifadelerine dayalı duygu analizi için KDEF ve RaFD veri setlerini kullanarak bir çalışma yürütmüştür. Araştırmada toplam 11 farklı algoritma karşılaştırılmış, bunlar arasında LinearSVC yöntemi %83 doğruluk oranı ile en başarılı sonuçları vermiştir. (Smirnov, 2022)

Akar ve arkadaşları, çalışmalarında KDEF ve PICS veri setlerini kullanarak yüz ifadelerinden duygu tanımaya yönelik bir model geliştirmiştir. Görüntü sayısının sınırlı olduğu PICS veri setinde veri artırma yöntemleri kullanılarak eğitim için yeterli veri miktarı elde edilmiştir. Bu kapsamda VGGNet mimarisinden uyarlanmış ESA tabanlı bir derin öğrenme modeli oluşturulmuş ve yedi temel duygu sınıfı üzerinde test edilmiştir. Model, KDEF veri setinin geçerleme kümesinde %97,44 doğruluk oranına ulaşırken, veri artırma uygulanmış PICS veri setinde ise %98,24 doğruluk değeri elde etmiştir. (Akar vd., 2022)

Tek modlu çalışmalar incelendiğinde, konuşma tabanlı duygu tanıma yaklaşımlarının genellikle akustik özelliklere ve geleneksel makine öğrenmesi ya da derin öğrenme modellerine dayandığı; metin tabanlı çalışmaların ise çoğunlukla bağlamsal dil modelleri veya dikkat mekanizmalarıyla anlamsal temsil gücünü artırmayı hedeflediği görülmektedir. Görüntü tabanlı yaklaşımlar ise yüz ifadeleri ve mimiklere dayalı görsel ipuçları üzerinden duygusal durumu modellemekte ve özellikle temel duygu sınıflarının ayırt edilmesinde etkili sonuçlar sunabilmektedir.

Bununla birlikte, duygunun çok boyutlu yapısını yalnızca tek bir modalite üzerinden ele almakta ve bu nedenle bağlam, tonlama ve anlamsal içerik gibi birbirini tamamlayan ipuçlarını eş zamanlı olarak değerlendirememektedir.

2.6.3. Çok Modlu Duygu Analizi

Çam ve arkadaşları (2023), az veriyle gerçekleştirilen duygu analizinde doğruluğu artırmak amacıyla ses ve metin tabanlı çok modlu bir model önermiştir. Çalışmada, CNN ile ses verilerinden ve LSTM ile metin verilerinden çıkarılan özellikler özellik düzeyinde

birleştirilerek (erken füzyon) multimodal bir model oluşturulmuştur. Deneysel sonuçlara göre, yalnızca metin analiziyle elde edilen %78 doğruluk oranı, ses ve metin verilerinin birlikte kullanılmasıyla %98'e yükselmiştir (Çam vd., 2023).

Zhao (2025), iş görüşmeleri bağlamında ses ve metin verilerini birleştiren çok modlu bir duygu analizi sistemi tasarlamıştır. Çalışmada MFCC ve LSTM tabanlı ses özellikleri ile BERT ve duygu sözlüklerine dayalı metin temsilleri kullanılmış; özellik düzeyi (erken füzyon) ve karar düzeyi (geç füzyon) stratejileri karşılaştırılmıştır. Deneysel sonuçlar, karar düzeyi füzyonun daha yüksek doğruluk (%92) ve F1 skoru (%91) sağladığını göstermiştir. (Zhao, 2025)

Kumar ve arkadaşları (2025), metin duygu analizi ile konuşma duygu tanıma süreçlerini birleştiren çok modlu bir iletişim analizi modeli önermiştir. Çalışmada, BERT tabanlı metin temsilleri ile VGGish tabanlı ses özellikleri transfer öğrenme yöntemiyle ayrı ayrı modellenmiş ve karar düzeyinde birleştirilmiştir (geç füzyon). Elde edilen sonuçlar, tek modlu sistemlere kıyasla doğruluk ve F1 skorunda %5–6 oranında artış sağlandığını göstermektedir (Kumar vd., 2025).

Hu ve arkadaşları (2025), metin, ses ve görsel verileri bütünleştiren MOTIF adlı metin odaklı çok modlu bir duygu analizi mimarisi geliştirmiştir. Önerilen sistemde, BiMamba-X tabanlı çift yönlü zaman kodlayıcısı kullanılarak uzun süreli bağımlılıklar yakalanmış; metin rehberli (attention tabanlı) hibrit bir füzyon stratejisi uygulanmıştır. Model, CMU-MOSI ve CMU-MOSEI veri setlerinde %85,6 doğruluk elde etmiştir (Hu vd., 2025).

Poria ve arkadaşları (2016), ses, görüntü ve metin modalitelerini içeren çok modlu bir duygu analizi çalışması gerçekleştirmiştir. Çalışmada YouTube verileri, SenticNet ve EmoSenticNet kaynakları birlikte kullanılmış; SVM, ELM ve ANN tabanlı sınıflandırıcılar uygulanmıştır. Modalitelerin birleştirilmesi karar düzeyinde gerçekleştirilmiş olup (geç füzyon) en yüksek başarı %77,13 ile ELM modeli üzerinden elde edilmiştir. (Poria vd., 2016)

Dixit ve arkadaşları (2024), gerçek zamanlı çok modlu duygu analizi için CNN tabanlı bir yaklaşım önermiştir. CMU-MOSEI veri seti üzerinde gerçekleştirilen çalışmada, farklı modalitelere ait çıktılar geç füzyon (late fusion) stratejisiyle birleştirilmiş ve yedi temel duygu için %85,85 doğruluk oranı elde edilmiştir. (Dixit vd., 2024)

Hakdağlı ve arkadaşları (2024), ses, görüntü ve metin verilerini birlikte kullanan çok modlu bir duygu analizi yaklaşımı önermiştir. Çalışmada Xception, VGG16/VGG19 ve BERT–ALBERT

modelleri kullanılmış; modalite çıktıları ağırlıklandırılmış karar düzeyi füzyon (geç füzyon) yöntemiyle birleştirilmiştir. Görüntü, metin ve ses modaliteleri için sırasıyla %98,25, %94,30 ve %90,71 F1 skorları rapor edilmiştir (Hakdağlı vd., 2024).

Bilotti ve arkadaşları (2024), konvolüsyonel sinir ağları kullanarak çok modlu duygu tanıma üzerine karşılaştırmalı bir çalışma gerçekleştirmiştir. BAUM-1 ve RAVDESS veri setleri üzerinde yapılan deneylerde, farklı modalitelere ait özellikler özellik düzeyinde birleştirilmiş (erken füzyon) ve CNN tabanlı model ile RAVDESS veri setinde %95,5 doğruluk elde edilmiştir (Bilotti, 2024)

Bu çalışmalar birlikte değerlendirildiğinde, çok modlu duygu analizi yaklaşımlarının tek modlu sistemlere kıyasla daha yüksek doğruluk ve genellenebilirlik sunduğu açıkça görülmektedir. Literatürde, erken füzyon yaklaşımlarının modaliteler arası etkileşimleri doğrudan öğrenme avantajı sağladığı; buna karşın, farklı veri türlerine ait özelliklerin heterojen yapısı nedeniyle gürültüye daha duyarlı olabildiği ve modalite katkılarının ayrı ayrı analiz edilmesini zorlaştırdığı rapor edilmiştir. Öte yandan, karar düzeyinde gerçekleştirilen geç füzyon yaklaşımlarının, her bir modalitenin bağımsız olarak modellenmesine olanak tanıması sayesinde daha esnek bir birleştirme sunduğu ve modalite bazlı performans analizini kolaylaştırdığı görülmektedir

Tablo 1'de sunulan çalışmalar, kullanılan veri setleri, duygu sınıfı sayıları, model mimarileri ve değerlendirme metrikleri açısından önemli farklılıklar göstermektedir. Bu nedenle tabloda yer alan başarı değerleri, doğrudan sayısal karşılaştırma amacıyla değil; literatürde kullanılan yaklaşımların çeşitliliğini ve genel performans eğilimlerini ortaya koymak amacıyla sunulmuştur. Her çalışma, kendi deneysel kurulumları altında yazarlar tarafından raporlanan en iyi sonuçlar ile temsil edilmektedir.

Tablo 1. Duygu Analizi Alanında Ses, Metin, Görüntü ve Çok Modlu Yaklaşımlara Ait Literatür Özeti

Yazar (Yıl)	Uygulama Alanı	Modalite	Veri Seti	Dil	Duygu	Model	Başarı
Casale vd. (2008)	Genel	Ses	EMO-DB, SUSAS	Almanca/ İngilizce	7	SVM, SMO	%92
Yun vd. (2011)	Genel	Ses	EMO-DB, SUSAS, DES, VAM	Çok dilli	7	LM-GMM	%87,80

Gumelar vd. (2019)	Genel	Ses	RAVDESS	İngilizce	8	DNN, CNN	%78,83
Donatus vd. (2024)	Genel	Ses	RAVDESS	İngilizce	8	MFCC & Spectrogram + SVM/RF	%54
Mustaqeem vd. (2020)	Genel	Ses	IEMOCAP, EMO-DB, RAVDESS	İngilizce	4,8	ResNet101 + BiLSTM	%85,57
Zhuang vd. (2025)	E-Ticaret / Online Yorumlar	Metin	Online review datasets	İngilizce/ Çince	2,5	ATNPWC-BiLSTM	%99
Sun vd. (2020)	E-Ticaret / Online Yorumlar	Metin	IMDB, DouBan	İngilizce/ Çince	2	CNN-BiLSTM-Attention	%92,45
Xie vd. (2024)	Eğitim	Metin	Eğitim metinleri	İngilizce	3	BERT + FastText	%88
Chen (2024)	Sağlık	Metin	METS-CoV	İngilizce	3	BERT + CNN	%72,97
Mekala vd. (2023)	Sosyal Medya	Metin	Sosyal medya metinleri	İngilizce	3	BERT, Bi-GRU	%97,1
Fujihira vd. (2020)	Sosyal Medya (Çok Dilli)	Metin	Twitter	Çok dilli	3	fastText + sözlük	%64
Zhan (2024)	Sağlık / Psikoloji	Metin	Öğrenci sosyal metinleri	İngilizce	3	LSTM	%97,6
Le (2020)	E-Ticaret / Online Yorumlar	Metin	Amazon, IMDb, Yelp	İngilizce	2	SADL (BiLSTM + Att)	%86,7
Omurca vd. (2017)	Hotel Yorum	Metin	Türkçe otel yorumları	Türkçe	3	Zemberek tabanlı	Veri kümesi
Qaisar (2020)	Online Yorumlar	Metin	Film yorumları	İngilizce	2	LSTM	%89,9
Zbancioc vd. (2021)	Yüz İfadeleri	Görüntü	MFSC	—	7	CNN	%85
Smirnov vd. (2022)	Yüz İfadeleri	Görüntü	KDEF, RaFD	—	7	LinearSVC	%83
Akar vd. (2022)	Yüz İfadeleri	Görüntü	KDEF, PICS	—	7	ESA-VGGNet	%98

Tablo 1.(Devamı)

Çam vd. (2023)	Genel	Çok Modlu S+M	Özel Türkçe veri seti	Türkçe	3	LSTM + CNN-2	%98
Zhao (2025)	Genel	Çok Modlu S+M	Görüşme veri seti	İngilizce	3	MFCC + LSTM + BERT	%92
Kumar vd. (2025)	HCI / Duygusal İletişim	Çok Modlu S+M	—	İngilizce	3	VGGish + BERT	%81,3
Hu vd. (2025)	Genel Amaçlı	Çok Modlu S+M+G	CMU-MOSI, CMU-MOSEI	İngilizce	7,5,2	MOTIF (BiMamba-X)	%85,6
Poria vd. (2016)	Sosyal Medya	Çok Modlu S+M+G	YouTube	İngilizce	3	SVM, ELM, ANN	%77,10
Dixit vd. (2024)	Gerçek Zamanlı Sistemler	Çok Modlu S+M+G	RAVDESS, FER-2013 CMU-MOSEI	İngilizce	6	FastText +CNN	%85,85
Hakdağlı vd. (2024)	Perakende / Müşteri Deneyimi	Çok Modlu S+M+G	RAVDESS, TESS, FER2013, LFW, Beyazperde	TR/EN	7	Xception + BERT/ALBERT	S-%90,7 G-98,25 M-%94,3
Bilotti vd. (2024)	Genel Amaçlı	Çok Modlu S+G	BAUM-1, RAVDESS	İngilizce	5	CNN	%95,95

Tablo 1.(Devamı)

Tablo 1'de özetlenen çalışmalar incelendiğinde, tek modlu duygu analizi yaklaşımlarının belirli veri türlerinde yüksek başarılar elde edebildiği görülmektedir. Ses tabanlı çalışmalar, spektral özellikler sayesinde duygusal tonlamayı yakalamada etkili olurken; metin tabanlı yaklaşımlar, özellikle bağlamsal dil modellerinin kullanımıyla anlamsal içeriği güçlü biçimde temsil edebilmektedir. Görüntü tabanlı yöntemler ise yüz ifadeleri ve mimiklere dayalı duygusal ipuçlarını başarılı bir şekilde modelleyebilmektedir.

Bununla birlikte, tek modlu yaklaşımların her biri duygunun yalnızca belirli bir boyutunu ele almakta ve duygusal durumun çok boyutlu doğasını tam olarak yansıtamamaktadır. Ses tabanlı modeller gürültü ve konuşmacı farklılıklarına duyarlı olabilmekte, metin tabanlı modeller konuşma sırasında ortaya çıkan tonlama ve vurgu gibi işitsel ipuçlarını göz ardı etmekte, görüntü tabanlı yöntemler ise bağlamsal ve dilsel bilgileri dikkate almamaktadır. Bu

sınırlılıklar, özellikle karmaşık ve bağlama duyarlı duyguların doğru biçimde sınıflandırılmasını zorlaştırmaktadır.

Bu nedenle, literatürde tek modlu yaklaşımların sınırlılıklarını gidermek amacıyla, farklı modalitelerin tamamlayıcı özelliklerinden yararlanan çok modlu duygu analizi yöntemlerine yönelimin giderek arttığı görülmektedir. Çok modlu yaklaşımlar, duygunun hem anlamsal hem işitsel hem de görsel boyutlarını birlikte değerlendirerek daha dengeli ve güvenilir sonuçlar sunmayı hedeflemektedir.

Literatürde çok modlu duygu analizi çalışmalarında kullanılan füzyon stratejileri incelendiğinde, erken füzyon ve geç füzyon yaklaşımlarının öne çıktığı görülmektedir. Erken füzyon yöntemleri, modaliteler arası etkileşimlerin özellik düzeyinde öğrenilmesine olanak tanımasına rağmen, farklı veri türlerine ait özelliklerin heterojen yapısı nedeniyle gürültüye daha duyarlı olabilmekte ve modalite katkılarının ayrı ayrı analiz edilmesini zorlaştırabilmektedir. Buna karşılık, özellikle son yıllardaki çalışmalarda, her bir modalitenin bağımsız olarak modellenmesine imkân tanıması, karar düzeyinde daha esnek birleştirme sunması ve modalite bazlı performans analizini mümkün kılması nedeniyle geç füzyon yaklaşımlarının daha yaygın biçimde tercih edildiği gözlemlenmektedir.

Bu literatür bulgularından hareketle, bu tez çalışmasında metin ve ses modaliteleri, her birinin duygu sınıflandırmasına olan katkısını ayrı ayrı koruyabilmek ve karar düzeyinde daha esnek bir birleştirme sağlamak amacıyla geç füzyon yaklaşımı kullanılarak birleştirilmiştir.

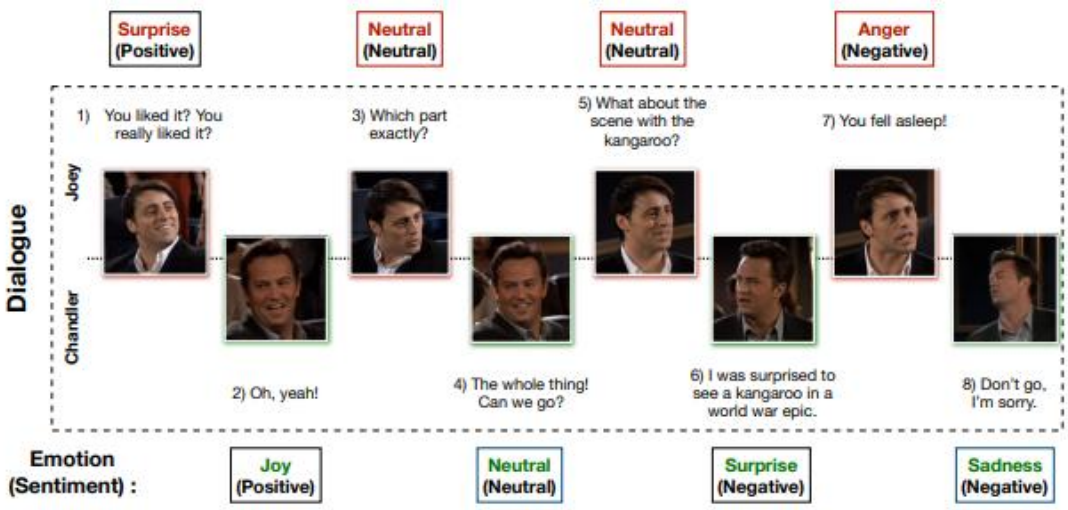
Literatür incelemesi, duygu analizi alanında kullanılan veri türlerinin, modelleme yaklaşımlarının ve temsil çıkarım yöntemlerinin çeşitliliğini ortaya koymaktadır. Ancak bu çalışmaların sağlıklı biçimde yorumlanabilmesi ve kullanılan yöntemlerin kuramsal temellerinin anlaşılabilmesi için, alana ait temel kavramların açık biçimde tanımlanması gerekmektedir. Bu doğrultuda bir sonraki bölümde, duygu analizi, temsil öğrenimi, Transformer mimarileri, çok modlu öğrenme yaklaşımları ve füzyon stratejileri gibi çalışmanın dayandığı kavramsal altyapı sistematik biçimde ele alınmaktadır.

Bu bölümde, duygu analizi problemine yönelik temel kavramlar ile makine öğrenmesi ve derin öğrenme yaklaşımları, Transformer tabanlı dil modelleri, füzyon stratejileri ve performans değerlendirme ölçütleri kavramsal çerçevede ele alınmış; ayrıca literatürde gerçekleştirilen ilgili çalışmalar incelenerek mevcut yaklaşımların güçlü yönleri ve sınırlılıkları değerlendirilmiştir. Sunulan bu teorik ve literatürel altyapı, çalışmanın yöntemsel yapısının anlaşılabilmesi için gerekli kavramsal zemini oluşturmaktadır. Bir sonraki bölümde ise, bu çerçeve temel alınarak çalışmada kullanılan veri seti, veri hazırlama süreçleri ve uygulanan yöntem ayrıntılı ve sistematik biçimde sunulmaktadır.

3. YÖNTEM

3.1. MELD (Multimodal EmotionLines Dataset)

MELD (Multimodal EmotionLines Dataset), Friends televizyon dizisinin diyaloglarından elde edilen çoklu konuşma sahnelerini içeren, çok modlu duygu tanıma araştırmalarında sıkça kullanılan bir veri setidir. Her diyalog; metin (text), ses (audio) ve video (visual) olmak üzere üç farklı modda sunulmakta ve konuşmacının duygusal durumuna göre yedi temel duygu sınıfından (anger, disgust, fear, joy, neutral, sadness, surprise) biriyle etiketlenmektedir. Veri seti toplamda 13.000'den fazla diyalog ifadesi (utterance) içermekte olup, her bir örneğe konuşmacı kimliği (speaker ID), diyalog kimliği (dialogue ID) ve konuşma sırası gibi ek bağlam bilgileri eşlik etmektedir. MELD'in en önemli özelliği, konuşmalar arası bağlam bilgisini koruması sayesinde çoklu konuşmacı ve bağlama dayalı duygu tanıma modellerinin geliştirilmesine olanak sağlamasıdır.



Şekil 12: Diyaloglarda konuşmacıların önceki duygularına göre duygu değişimlerinin gösterimi.

3.2. Veri Hazırlama ve Ön İşleme Süreci

Bu çalışmada MELD veri seti, deneysel aşamaya geçilmeden önce modele uygun ve tutarlı bir yapıya dönüştürülmüştür. Veriler konuşma parçası (utterance) düzeyinde ele alınmış; her utterance için tek bir duygu etiketi kullanılmıştır. Metin ve ses verileri aynı utterance kimliği üzerinden eşleştirilerek modaliteler arası tutarlılık sağlanmış, böylece tekil modalite (yalnız

metin / yalnız ses) ve çoklu modalite (metin + ses) deneyleri için ortak bir veri yapısı oluşturulmuştur.

Model performansının güvenilir biçimde değerlendirilebilmesi amacıyla veri seti; eğitim, doğrulama (dev/validation) ve test olmak üzere üç alt kümede kullanılmıştır. MELD veri seti için literatürde yaygın biçimde kullanılan hazır veri bölünmeleri korunmuş; doğrulama kümesi model seçimi ve erken durdurma gibi kararlar için, test kümesi ise yalnızca nihai performans değerlendirmesi için ayrılmıştır. Bu süreçte veri sızıntısını önlemek amacıyla, eğitim kümesinde yer alan örneklerin doğrulama veya test kümelerine taşınmamasına özen gösterilmiştir.

MELD veri setinde duygu sınıfları arasında belirgin bir dengesizlik bulunmaktadır. Bu durum, modellerin çoğunluk sınıflara yönelmesine ve az temsil edilen sınıflarda düşük performans göstermesine neden olabilmektedir. Bu tür durumlarda yeniden örnekleme (resampling) yaklaşımları sıklıkla tercih edilmektedir. Oversampling, az temsil edilen sınıflara ait örnek sayısının artırılması yoluyla sınıf dağılımının dengelenmesini hedefler ve bu yaklaşımın, azınlık sınıflarının daha iyi öğrenilmesine katkı sağladığı gösterilmiştir (Belhaouari vd., 2024). Ancak yalnızca kopyalama temelli çoğaltma işlemleri bazı durumlarda aşırı öğrenmeye yol açabilmektedir. Bu nedenle, veri çeşitliliğini artırmak ve genelleme yeteneğini güçlendirmek amacıyla veri artırma (data augmentation) yöntemlerinden yararlanılmaktadır. Veri artırma, mevcut örnekler üzerinde uygulanan kontrollü dönüşümler aracılığıyla daha zengin bir eğitim kümesi oluşturmayı amaçlamakta ve özellikle derin öğrenme modellerinde aşırı öğrenme riskini azaltmaya yardımcı olmaktadır (Shorten vd, 2019) (Cubuk vd., 2019).

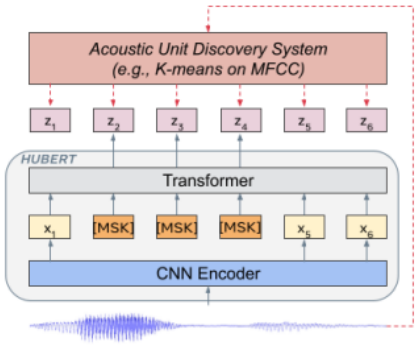
Bu çalışmada sınıf dengesizliğini azaltmak amacıyla Random Oversampling (ROS) yöntemi kullanılmıştır. ROS yaklaşımıyla, az temsil edilen duygu sınıflarına ait örnekler rastgele seçilerek çoğaltılmış ve her sınıfın örnek sayısı eğitim kümesindeki en büyük sınıfın örnek sayısına dengelenmiştir.

Oversampling işlemi sonrasında veri çeşitliliğini artırmak ve aşırı öğrenme (overfitting) riskini azaltmak amacıyla veri artırma (data augmentation) uygulanmıştır. Ses verileri üzerinde akustik yapıyı bozmayan dönüşümler (ör. perde kaydırma, zaman ölçekleme, gürültü ekleme, genlik değişimi) kullanılmış; metin verileri için ise anlamsal tutarlılığı koruyan artırma yaklaşımlarından yararlanılmıştır.

Tüm oversampling ve augmentation işlemleri yalnızca eğitim kümesi üzerinde gerçekleştirilmiş, doğrulama ve test kümeleri bu süreçlere dahil edilmemiştir. Böylece veri sızıntısı (data leakage) önlenmiş ve model performans değerlendirmesinin güvenilirliği korunmuştur.

3.3. Ses Verisi

Konuşma sinyalleri, zaman içerisinde değişen ve farklı frekans bileşenlerini aynı anda barındıran karmaşık yapılar içermektedir. Bu tür verilerin doğrudan kullanımı, zamansal ve bağlamsal ilişkilerin yeterince temsil edilememesi nedeniyle zorluklar barındırmaktadır. Bu nedenle ses verilerinin, konuşmanın akustik ve zamansal özelliklerini yansıtacak biçimde sayısal temsillere dönüştürülmesi yaygın bir yaklaşımdır. Son yıllarda, etiketli veriye ihtiyaç duymadan ham ses sinyallerinden temsil öğrenebilen kendi kendine denetimli (self-supervised) öğrenme tabanlı modeller, konuşma temsili çıkarımı alanında önemli bir ilerleme sağlamıştır. HuBERT, kendi kendine denetimli öğrenme yaklaşımıyla ham ses sinyallerinden konuşma temsili öğrenen bir modeldir. Bu yaklaşımda, ham ses sinyali üzerinden elde edilen akustik özellikler öncelikle kümeleme yöntemiyle ayrık gizli birimlere dönüştürülmekte, ardından model maskelenmiş zaman adımlarındaki bu gizli birimleri tahmin etmeye zorlanmaktadır. Model, yalnızca maskelenmiş bölgelere odaklanan bu öğrenme mekanizması sayesinde hem yerel akustik örüntüleri hem de uzun menzilli zamansal ve bağlamsal ilişkileri öğrenebilmektedir. HuBERT modelinin genel mimarisi ve maskelenmiş zaman adımlarında gizli akustik birimlerin tahminine dayalı öğrenme süreci Şekil 14'te gösterilmektedir (Hsu vd., 2021).



Şekil 13: Hubert modelinin genel mimarisi ve öğrenme süreci (Hsu vd., 2021)

HuBERT modelinin öğrenme süreci, maskelenmiş zaman adımlarındaki gizli birimlerin doğru şekilde tahmin edilmesine dayalı bir kayıp fonksiyonu ile optimize edilmektedir. Bu süreçte

model, yalnızca maskelenmiş zaman adımlarını dikkate alarak hedef gizli birimlerin olasılığını eniyilemeye çalışmaktadır. Bu öğrenme mekanizması genel olarak aşağıdaki şekilde ifade edilebilir:

$$L_m(f; X, M, Z) = \sum_{t \in M} \log p_f(z_t \mid \tilde{X}, t),$$

Burada $\mathcal{M}$ maskelenmiş zaman adımlarını, z_t ilgili zaman adımına ait hedef gizli birimi ve x giriş ses sinyalini temsil etmektedir (Hsu vd., 2021).

Bu çalışmada ham ses sinyallerinden temsil elde etmek amacıyla önceden eğitilmiş HuBERT (Hidden-Unit BERT) modeli kullanılmıştır. HuBERT modeli, ham ses sinyalinden elde edilen temsil katmanlarının çıktıları üzerinden yüksek seviyeli akustik ve zamansal bilgileri içeren gömme (embedding) vektörleri üretmektedir. Bu çalışmada, önceden eğitilmiş HuBERT modelinin son Transformer katmanına ait çıktı temsilleri zaman boyutunda özetlenerek (pooling), her bir konuşma parçası (utterance) için sabit boyutlu ses gömme (embedding) vektörleri elde edilmiştir.

HuBERT modeli bu çalışmada sabit özellik çıkarıcı (feature extractor) olarak kullanılmış, model üzerinde herhangi bir fine-tuning işlemi uygulanmamıştır. Elde edilen ses gömme vektörleri, daha sonraki aşamada eğitilen hafif birçok katmanlı algılayıcı (MLP) sınıflandırıcıya girdi olarak verilmiş ve sınıflandırma süreci bu yapı üzerinden gerçekleştirilmiştir.

3.4. Metin Verisi

Doğal dil verileri, sözcüklerin tekil anlamlarının yanı sıra bağlam içerisinde kazandıkları anlamsal ilişkileri de içeren karmaşık yapılardan oluşmaktadır. Bu nedenle metin verilerinin doğrudan sayısal biçimde kullanımı yerine, bağlamsal bilgiyi yansıtan temsil vektörlerine dönüştürülmesi yaygın bir yaklaşımdır. Özellikle konuşma metinlerinde, kelimelerin anlamı önceki ve sonraki ifadelerle güçlü biçimde ilişkili olduğundan, bağlam bilgisinin doğru biçimde modellenmesi duygu tanıma görevleri açısından kritik önem taşımaktadır.

Bu çalışmada metin verilerinin bağlamsal temsillerinin elde edilmesi amacıyla, önceden eğitilmiş Transformer tabanlı RoBERTa (Robustly Optimized BERT Approach) modeli kullanılmıştır. RoBERTa modeli, girdi metinleri için bağlamsal temsilleri son Transformer

katmanı (last hidden layer) üzerinden üretmekte ve bu temsiller sınıflandırma başlığı (classification head) aracılığıyla duygu sınıflarına dönüştürülmektedir.

RoBERTa modeli, duygu sınıflandırma görevine özgü olacak şekilde fine-tuning yöntemiyle yeniden eğitilmiştir. Bu süreçte yalnızca sınıflandırma katmanı değil, modelin tüm ağırlıkları görev verisi üzerinde güncellenmiş; böylece duyguya özgü bağlamsal ve anlamsal dilsel temsillerin uçtan uca bir yapı içerisinde doğrudan model tarafından öğrenilmesi sağlanmıştır. Modelin son katman çıktıları, her bir konuşma parçası (utterance) için yüksek boyutlu metin temsilleri olarak elde edilmiştir.

Elde edilen bu temsiller, sınıflandırma başlığı üzerinden işlenerek her bir konuşma metni için duygu sınıflarına ait logits değerleri üretilmiş; ardından bu logits çıktıları softmax fonksiyonu aracılığıyla olasılık dağılımına dönüştürülerek her örnek için duygu sınıflarını temsil eden olasılık temsilleri elde edilmiştir.

3.5. Çoklu Modalite Birleştirme

Bu çalışmada duygu tanıma problemi, metin ve ses modalitelerinin birlikte değerlendirilmesiyle ele alınmıştır. Farklı modalitelerin duyguya ilişkin tamamlayıcı bilgiler sunduğu göz önünde bulundurularak, çoklu modalite birleştirme yaklaşımı benimsenmiştir. Metin ve ses modaliteleri için elde edilen temsiller, birbirinden bağımsız sınıflandırıcılar kullanılarak işlenmiş; modalitelere ait çıktı olasılıkları, geç füzyon yaklaşımı doğrultusunda bir araya getirilmiştir.

3.5.1. Geç füzyon (late fusion) yöntemi

Geç füzyon (late fusion) yaklaşımı, çok modlu sistemlerde her bir modalitenin ayrı ve bağımsız modeller aracılığıyla işlenmesi ve bu modellere ait çıktıların karar düzeyinde birleştirilmesi esasına dayanmaktadır. Bu yaklaşımda her modalite, kendi özelliklerine uygun bir öğrenme süreci izlemekte; çoklu modalite entegrasyonu ise doğrudan özellik seviyesinde değil, model çıktıları üzerinden gerçekleştirilmektedir. Böylece modaliteler arası etkileşim, karar aşamasında sağlanmakta ve her bir modalite bağımsız biçimde modellenmektedir. (Pereira vd., 2023) (Tadas vd., 2018)

Bu çalışmada geç füzyon yaklaşımı, metin ve ses modaliteleri için ayrı ayrı eğitilen sınıflandırma modellerinin çıktı olasılıklarının ağırlıklı toplamı şeklinde uygulanmıştır. Çok

modlu karar aşamasında, metin ve ses modalitelerine ait sınıflandırma çıktıları aşağıdaki biçimde birleştirilmiştir:

$$F = \alpha \cdot P_{metin} + (1 - \alpha) \cdot P_{ses}$$

Burada P_{metin}ve P_{ses}sırasıyla metin ve ses modalitelerine ait sınıf olasılık vektörlerini, αise modaliteler arası katkı ağırlığını temsil etmektedir.

Bu çalışmada metin ve ses modalitelerinin geç füzyon ile birleştirilmesinde kullanılan ağırlık katsayısı α, sabit bir değer olarak tanımlanmamış; her bir duygu sınıflandırma görevi için geliştirme (DEV) veri kümesi üzerinde grid search yaklaşımı ile belirlenmiştir. Bu kapsamda α değeri, 0.70–0.98 aralığında önceden tanımlanan değerler üzerinden taranmış ve her bir α değeri için doğrulama kümesindeki sınıflandırma performansı hesaplanmıştır. Deneyler, üç farklı seed (42, 777 ve 2024) kullanılarak yürütülmüş; aynı α değeri için elde edilen sonuçlar seed'ler üzerinden ortalama alınarak değerlendirilmiştir. Doğrulama kümesinde elde edilen sonuçlar karşılaştırılarak, en uygun sınıflandırma performansını sağlayan α değeri ilgili görev için optimum füzyon katsayısı olarak seçilmiştir. Test kümesi üzerindeki nihai değerlendirmeler ise yalnızca DEV kümesinde bu şekilde belirlenen α değeri kullanılarak gerçekleştirilmiştir. Bu seçim süreci, 3 duygu, 5 duygu ve 7 duygu sınıflandırma görevlerinin her biri için bağımsız olarak uygulanmış; denenen α değerleri ve bunlara karşılık gelen doğrulama (DEV) kümesi sonuçları Tablo 2'de verilmektedir.

Tablo 2: DEV Kümesi Üzerinde Grid Search ile Belirlenen α Değerlerinin Performans Sonuçları

	3-Duygu		**5-Duygu**		**7-Duygu**	
α	**WF1**	**ACC**	**WF1**	**ACC**	**WF1**	**ACC**
0.70	0.6867	0.6895	**0.5048**	0.5033	**0.5530**	0.5738
0.80	0.6888	0.6913	0.4937	0.4976	0.5522	0.5741
0.84	0.6916	0.6940	0.4818	0.4925	0.5505	0.5729
0.88	0.6914	0.6937	0.4726	0.4903	0.5510	0.5735
0.90	0.6912	0.6934	0.4690	0.4903	0.5514	0.5738
0.92	0.6907	0.6928	0.4661	0.4906	0.5508	0.5735
0.94	0.6907	0.6928	0.4603	0.4871	0.5508	0.5735
0.96	0.6910	0.6931	0.4472	0.4791	0.5505	0.5732
0.98	**0.6920**	0.6940	0.4382	0.4763	0.5501	0.5729

3.5.2. Seed tabanlı çoklu çalıştırma (ensemble) stratejisi

Derin öğrenme modellerinin eğitimi sırasında kullanılan rastgele ağırlık başlatma ve eğitim sürecindeki rastgelelik, aynı mimari yapı ve aynı veri bölünmeleri kullanılsa dahi farklı model davranışlarının ortaya çıkmasına neden olabilmektedir. Bu durum, tek bir eğitim çalıştırmasına dayalı olarak elde edilen sonuçların, modele özgü rastlantısal etkilerden etkilenmesine yol açabilmektedir.

Bu çalışmada seed tabanlı ensemble stratejisi kapsamında üç farklı rastgele başlangıç tohumu (seed = 42, 2024, 777) kullanılmıştır. Metin ve ses modalitelerine ait modeller, aynı mimari yapı, aynı hiperparametreler ve aynı veri bölünmeleri korunarak, bu üç farklı seed değeri ile bağımsız olarak eğitilmiştir.

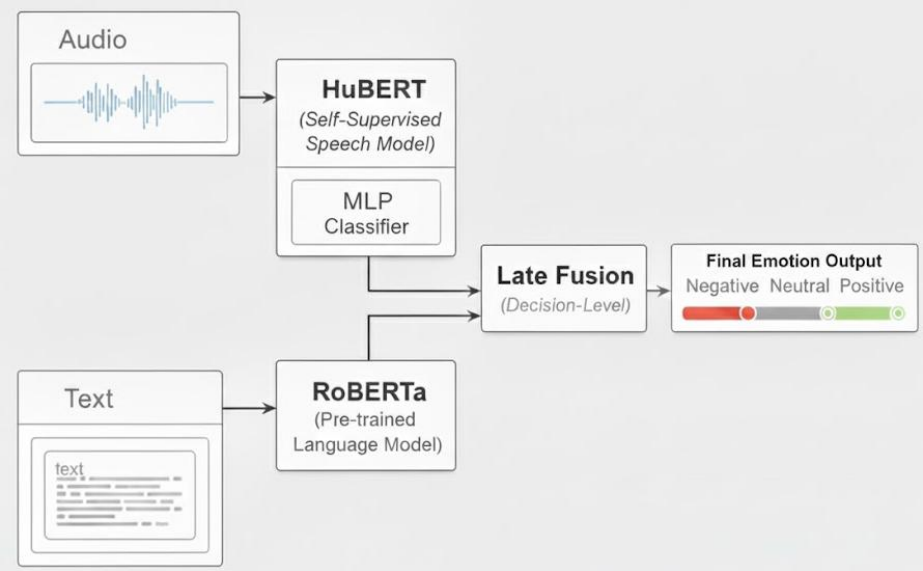
Her bir seed için elde edilen modeller, konuşma parçalarına ait duygu sınıfları için ayrı olasılık çıktıları üretmiştir. Nihai duygu tahmini, bu bağımsız çalıştırmalardan elde edilen çıktıların olasılık düzeyinde ortalaması alınarak oluşturulmuştur. Başka bir ifadeyle, model ağırlıkları birleştirilmemiş; her bir modelin ürettiği sınıf olasılık vektörleri üzerinde aritmetik ortalama (mean ensemble) işlemi uygulanmıştır.

Bu yaklaşımda herhangi bir ek öğrenilebilir meta-model kullanılmamış, ensemble işlemi yalnızca çıktı düzeyinde gerçekleştirilmiştir. Böylece farklı rastgele başlangıçlardan kaynaklanan varyans etkisi azaltılmış, daha kararlı ve genellenebilir bir karar mekanizması elde edilmesi hedeflenmiştir.

3.5.3. Ses ve metin özelliklerinin birleştirilmesi

Bu çalışmada ses ve metin modalitelerine ait bilgiler, doğrudan ham özellikler düzeyinde birleştirilmemiş, bunun yerine her bir modalite için öğrenilen temsillerin sınıflandırma çıktıları esas alınmıştır. Metin ve ses verileri, modaliteye özgü temsil öğrenimi ve sınıflandırma aşamalarından ayrı ayrı geçirilmiş; her bir modalite, konuşma parçasına ait duygu sınıfları için birer çıktı vektörü üretmiştir. Bu bağlamda “özellik birleştirme”, her bir modalitenin duyguya ilişkin ayırt edici bilgisini yansıtan çıktı temsillerinin birlikte değerlendirilmesi şeklinde gerçekleştirilmiştir. Ses ve metin modalitelerine ait bu çıktılar, geç füzyon yaklaşımı doğrultusunda karar düzeyinde bir araya getirilmiş ve modalitelerin sınıflandırma sürecine olan katkıları deneysel olarak dengelenmiştir. Bu yaklaşım sayesinde, her bir modalitenin kendi

temsil gücü korunmuş, erken aşamada gerçekleştirilen birleştirmelerde ortaya çıkabilen boyut artışı ve hizalama problemlerinin önüne geçilmiştir.



Şekil 14: Çok modlu duygu analizi örnek gösterimi

3.6. Derin Öğrenme Mimarisinin Yapısı

Bu çalışmada kullanılan derin öğrenme mimarisi, metin ve ses modalitelerinin farklı temsil özellikleri dikkate alınarak modaliteye özgü şekilde yapılandırılmıştır. Metin ve ses verileri için temsil öğrenimi aşamaları önceden eğitilmiş modeller aracılığıyla gerçekleştirilmiş, duygu sınıflandırma işlemi ise bu temsiller üzerine eklenen sınıflandırma katmanları üzerinden yürütülmüştür. Çoklu modalite yapısı kapsamında, metin ve ses modalitelerine ait sınıflandırma çıktıları geç füzyon yaklaşımı ile birleştirilmiştir.

3.6.1. Kullanılan sinir ağı mimarisi

Bu çalışmada kullanılan derin öğrenme mimarisi, metin ve ses modalitelerinin farklı temsil özellikleri dikkate alınarak modaliteye özgü şekilde yapılandırılmıştır. Metin ve ses verileri için temsil öğrenimi aşamaları önceden eğitilmiş modeller aracılığıyla gerçekleştirilmiş, duygu sınıflandırma işlemi ise bu temsiller üzerine eklenen sınıflandırma katmanları üzerinden yürütülmüştür. Çoklu modalite yapısı kapsamında, metin ve ses modalitelerine ait sınıflandırma çıktıları geç füzyon yaklaşımı ile birleştirilmiştir.

3.6.2. Aktivasyon fonksiyonları ve düzenlileştirme

Model mimarisinde doğrusal olmayan ilişkilerin öğrenilebilmesi amacıyla uygun aktivasyon fonksiyonları kullanılmıştır. Transformer tabanlı RoBERTa modeli, iç yapısında doğrusal olmayan dönüşümler aracılığıyla bağlamsal temsil öğrenimini gerçekleştirmektedir (Liu vd., 2019) Ses modalitesi için kullanılan sınıflandırma yapısında ise, doğrusal olmayan aktivasyonlar sayesinde HuBERT temsilleri daha ayırt edici bir sınıflandırma uzayına taşınmıştır.

Aşırı öğrenmenin önüne geçmek ve modelin genelleme yeteneğini artırmak amacıyla düzenlileştirme yöntemlerinden yararlanılmıştır. Bu kapsamda, sınıflandırma katmanlarında dropout uygulanmış ve eğitim süreci boyunca doğrulama kümesi performansı izlenerek erken durdurma (early stopping) mekanizması kullanılmıştır. Dropout tekniği, ağın belirli nöronlara aşırı bağımlılık geliştirmesini engelleyerek genelleme performansını artırmayı amaçlamaktadır. (Srivastava vd., 2014)

3.6.3. Çıkış katmanı ve sınıflandırma yapısı

Metin ve ses modaliteleri için oluşturulan sınıflandırma yapılarının çıkış katmanları, çalışmada kullanılan duygu sınıfı sayısına uygun olacak şekilde yapılandırılmıştır. Her bir modalite için çıkış katmanında, duygu sınıflarına karşılık gelen skorlar üretilmiş ve bu skorlar olasılık dağılımlarına dönüştürülmüştür.

Geç füzyon yaklaşımı kapsamında, metin ve ses modalitelerine ait bu çıktı olasılıkları karar düzeyinde birleştirilmiş ve her bir konuşma parçası için duygu tahmini elde edilmiştir. Bu yapı sayesinde, tekil modalite çıktıları korunarak birlikte değerlendirilmiş ve çoklu modalite kullanımının sınıflandırma sürecine olan katkısı analiz edilebilir hâle getirilmiştir.

Bu bölümde, çalışmada kullanılan veri seti, veri hazırlama süreçleri, temsil çıkarım yaklaşımları, sınıflandırma yapıları ve çoklu modalite birleştirme stratejileri ayrıntılı olarak sunulmuştur. Tanımlanan bu yöntemsel çerçeve doğrultusunda gerçekleştirilen deneysel çalışmaların sonuçları bir sonraki bölümde sistematik biçimde raporlanmakta ve elde edilen bulgular performans ölçütleri üzerinden sunulmaktadır.

4. BULGULAR

Bu bölümde, MELD veri seti üzerinde metin (RoBERTa), ses (HuBERT) ve çoklu modalite (geç füzyon) yaklaşımları ile elde edilen sınıflandırma sonuçları sunulmuştur. Deneyler, üç farklı etiketleme senaryosu altında gerçekleştirilmiştir: 3 duygu (Negative/Neutral/Positive), 5 duygu *(*Anger/Sadness/Joy/Neutral/Surprise*)* ve 7 duygu (Anger/Sadness/Joy/Neutral/Surprise/ Fear/Disgust). Performans değerlendirmesinde sınıf dengesizliğinin etkisini dikkate almak amacıyla Weighted F1 (WF1) temel ölçüt olarak ele alınmış; buna ek olarak accuracy, ROC-AUC (OvR) ve karışıklık matrisleri üzerinden sınıflar arası karışma eğilimleri analiz edilmiştir. Metin, ses ve çoklu modalite yaklaşımlarının 3, 5 ve 7 duygu senaryolarındaki genel WF1 performansları Tablo 3'te sunulmaktadır.

Tablo 3: 3, 5 ve 7 Duygu Senaryoları İçin WF1 Karşılaştırması

Model	**3 Duygu**	**5 Duygu**	**7 Duygu**
Roberta (Text)	70.54	63.45	56.31
Hubert + Mlp (Audio)	43.67	59.13	50.56
Hubert + Cnn (Audio)	42,99	33,87	32,43
Roberta + Hubert + Mlp (Late Fusion)	**72**	**66**	**62**

Tablo 3 incelendiğinde, geç füzyon yaklaşımının üç senaryonun tamamında en yüksek WF1 değerlerini ürettiği görülmektedir. Geç füzyon modelinde 3 duygu için WF1=72, 5 duygu için WF1=66 ve 7 duygu için WF1=62 elde edilmiştir. Bu bulgu, literatürde çok modlu duygu analizine ilişkin çalışmalarda sıklıkla raporlanan "çoklu modalite > tekil modalite" eğilimiyle tutarlıdır. Önceki çalışmalarda, metin modalitesinin tek başına genellikle ses modalitesine kıyasla daha yüksek performans ürettiği, ancak metin ve sesin birlikte kullanıldığı çok modlu yapılarda sınıflandırma başarısının daha dengeli ve kararlı hale geldiği gösterilmiştir (Poria vd., 2017; Tadas vd., 2018). Bu çalışma kapsamında elde edilen sonuçlar da aynı eğilimi desteklemekte; özellikle geç füzyon yaklaşımının, tekil modalitelerin zayıf kaldığı örneklerde dengeleyici bir rol üstlendiğini ortaya koymaktadır.

Ses modalitesi özelinde iki farklı sınıflandırıcı denenmiştir. HuBERT + CNN yaklaşımının özellikle 5 ve 7 duygu senaryolarında daha düşük performansta kaldığı görülmüştür, sırasıyla WF1=33.87 ve WF1=32.43 bu durum, HuBERT gömmelerinin sabit boyutlu ve özet temsiller olması nedeniyle CNN'in sağladığı yerel örüntü yakalama avantajının bu temsil yapısı üzerinde

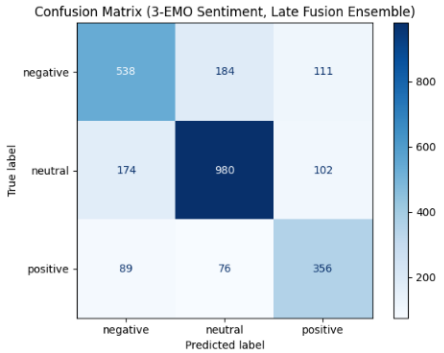
etkili biçimde kullanılamamasından kaynaklanmaktadır. Buna karşılık, gömme vektörleri üzerinde doğrudan çalışan MLP tabanlı sınıflandırıcı, temsil uzayına daha uygun bir öğrenme dinamiği sunmuş ve daha kararlı sonuçlar üretmiştir.

3 duygu senaryosuna ait sınıf bazlı precision, recall ve F1-score değerleri Tablo 4'te sunulmuştur. Bu senaryoda geç füzyon modeli için accuracy=72 ve weighted F1=72 elde edilmiştir. Neutral sınıfı (F1=79) en yüksek performansı gösterirken, negative (F1=66) ve positive (F1=65) sınıfları birbirine yakın sonuçlar üretmiştir.

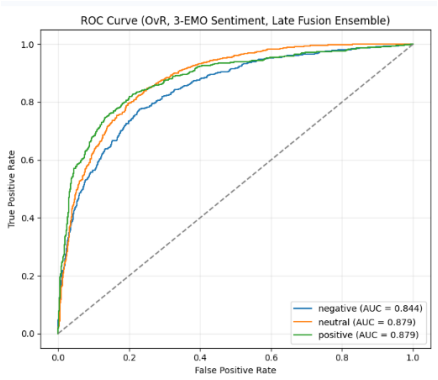
Tablo 4: 3 Duygu Sınıfı İçin Sınıflandırma Performans Sonuçları

Duygu	**Precision**	**Recall**	**F1-Score**
Negative	67	65	66
Neutral	79	78	79
Positive	63	68	65
Accuracy			**72**
Weighted Avg	**72**	**72**	**72**

Şekil 15'te verilen karışıklık matrisi incelendiğinde, hataların büyük ölçüde neutral sınıfı etrafında yoğunlaştığı görülmektedir. Özellikle negative–neutral ve positive–neutral geçişleri belirgin olup, bu durum, neutral sınıfının semantik ve akustik özellikler açısından iki sınıf arasında konumlanması nedeniyle karar sınırlarının bu sınıf etrafında yoğunlaşmasından kaynaklanmaktadır. Şekil 16'da sunulan ROC eğrileri ise üç sınıfın da kabul edilebilir düzeyde ayrıştırılabildiğini, ancak neutral sınıfının karar sınırlarının daha belirgin olduğunu göstermektedir.



Şekil 15: 3 duygu senaryosu için karışıklık matrisi



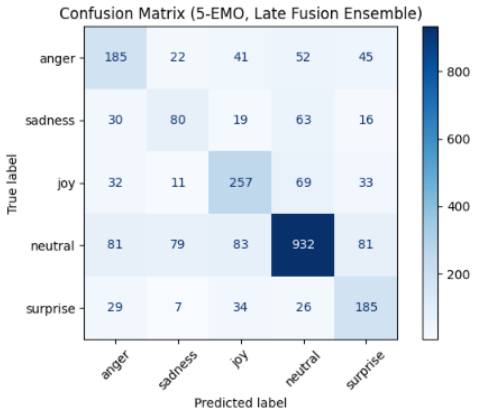
Şekil 16: 3 duygu senaryosu için roc eğrisi

5 duygu senaryosuna ait sınıf bazlı performans sonuçları Tablo 5'te verilmiştir. Bu senaryoda geç füzyon modeli için accuracy=66 ve weighted F1=67 elde edilmiştir. Neutral (F1=79) ve joy (F1=61) sınıfları görece yüksek performans üretirken, sadness (F1=39) belirgin biçimde daha düşük kalmıştır.

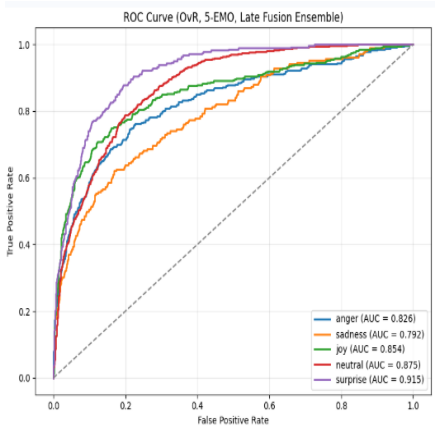
Tablo 5: 5 Duygu Sınıfı İçin Sınıflandırma Performans Sonuçları

Duygu	Precision	Recall	F1 score
Anger	52	54	53
Sadness	40	38	39
Joy	59	64	61
Neutral	82	74	79
Surprise	51	66	58
Accuracy			66
Weighted avg	67	66	67

Şekil 17'de sunulan karışıklık matrisi, sadness ve anger sınıflarının sıklıkla neutral sınıfı ile karıştığını göstermektedir. Şekil 18'deki ROC eğrileri incelendiğinde ise surprise (AUC=0.915) ve neutral (AUC=0.875) sınıflarının daha net ayrıştırılabildiği görülmektedir.



Şekil 17: 5 duygu senaryosu için karışıklık matrisi



Şekil 18: 5 duygu senaryosu için roc eğrisi

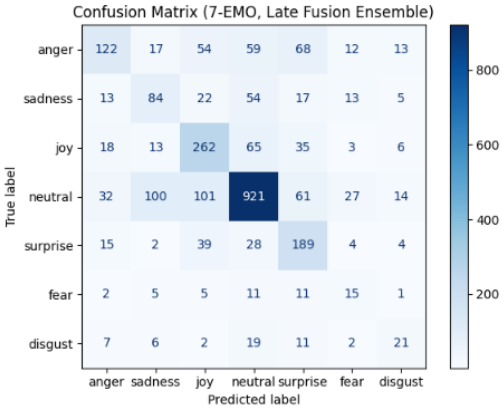
Sınıf sayısının artmasıyla birlikte karar sınırları daha karmaşık hale gelmekte, bu durum özellikle semantik olarak yakın sınıflarda karışmayı artırmaktadır. Sadness sınıfındaki düşük performans hem örnek sayısının görece azlığı hem de sadness–neutral sınırının belirsizliği ile ilişkilidir. Geç füzyon, bu karmaşıklığa rağmen tekil modalitelere kıyasla daha dengeli bir performans üretmiştir.

7 duygu senaryosu, sınıf sayısının artması ve bazı sınıfların düşük örnek sayısına sahip olması nedeniyle en zor problem tanımını oluşturmaktadır. Bu senaryoda geç füzyon modeli için accuracy=62 ve weighted F1=62 elde edilmiştir. Neutral (F1=76) ve joy (F1=59) sınıfları görece güçlü sonuçlar verirken, fear (F1=24) ve disgust (F1=32) sınıflarında belirgin performans düşüşü gözlenmiştir. Bu senaryoya ait sınıf bazlı performans sonuçları Tablo 6'da sunulmuştur.

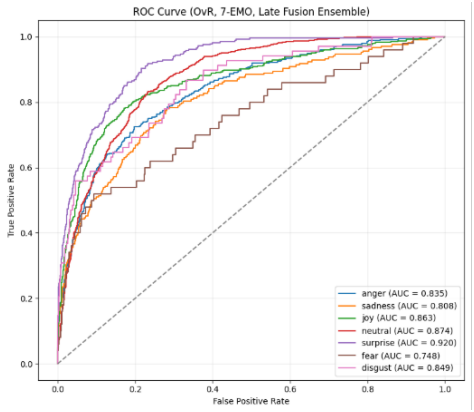
Tablo 6: 7 Duygu Sınıfı İçin Sınıflandırma Performans Sonuçları

Duygu	Precision	Recall	F1-Score
Anger	58	35	44
Sadness	37	40	39
Joy	54	65	59
Neutral	80	73	76
Surprise	48	67	56
Fear	20	30	24
Disgust	33	31	32
Accuracy			62
Weighted Avg	64	62	62

Şekil 19'daki karışıklık matrisi, özellikle fear ve disgust sınıflarının sadness ve neutral sınıflarıyla yoğun biçimde karıştığını göstermektedir. Şekil 20'deki ROC eğrileri de bu zorluğu yansıtmaktadır; fear (AUC=0.748) sınıfında ayrıştırma gücünün düşük kaldığı, buna karşılık surprise (AUC=0.920) ve neutral (AUC=0.874) sınıflarında daha net karar sınırlarının öğrenildiği görülmektedir.



Şekil 19: 7 duygu senaryosu için karışıklık matrisi



Şekil 20: 7 duygu senaryosu için roc eğrisi

7 sınıflı yapı, karar uzayını belirgin biçimde karmaşıklaştırmakta ve sınıflar arası örtüşmeyi artırmaktadır. Fear ve disgust sınıflarındaki düşük performans, hem sınıf dengesizliği hem de bu duyguların akustik ve anlamsal olarak diğer sınıflarla örtüşmesiyle ilişkilidir. Geç füzyon yapısı, bu zorlu senaryoda dahi tekil modalitelere kıyasla daha kararlı sonuçlar üretmiştir.

Bu bölümde, farklı duygu senaryoları altında metin, ses ve çoklu modalite yaklaşımları ile elde edilen deneysel bulgular sistematik biçimde sunulmuştur. Elde edilen performans sonuçları ve sınıf bazlı analizler, model davranışına ilişkin önemli ipuçları ortaya koymaktadır. Bir sonraki bölümde ise, bu bulgular literatür bağlamında değerlendirilerek çalışmanın genel sonuçları tartışılmakta ve gelecekte yapılabilecek çalışmalara yönelik öneriler sunulmaktadır.

5. TARTIŞMA VE SONUÇ

Üç senaryo birlikte değerlendirildiğinde, sınıf sayısının artmasıyla birlikte sınıflandırma performansının azalması, problem uzayının yapısal olarak karmaşıklaşmasının doğrudan bir sonucu olarak ortaya çıkmaktadır. Bu düşüşün temel nedenleri şu şekilde özetlenmektedir: sınıf sayısının artmasıyla karar sınırlarının daha karmaşık hale gelmesi, sınıf dengesizliğinin özellikle azınlık sınıflarda genelleme yeteneğini doğrudan sınırlaması, bazı duyguların (ör. sadness–neutral, fear–sadness) semantik ve akustik olarak birbirine yakın olması nedeniyle sınıf sınırlarının yapısal olarak belirsizleşmesi, veri miktarının artan sınıf sayısına oranla sınırlı kalması.

Bununla birlikte, her üç senaryoda da geç füzyon yaklaşımının en iyi sonuçları vermesi, metin ve ses modalitelerinin birbirini tamamlayan bilgiler sunduğunu açık biçimde göstermektedir. Metin modalitesi anlamsal ipuçlarını güçlü biçimde taşırken, ses modalitesi tonlama, vurgu ve ritim gibi akustik sinyallerle bu bilgiyi desteklemekte; karar düzeyinde birleştirme ise tekil modalitelerin zayıf kaldığı örneklerde dengeleyici bir rol üstlenmektedir. Bu yapı, çok modlu duygu analizinin tekil modalite temelli yaklaşımlara kıyasla daha kararlı ve daha yüksek genelleme gücüne sahip çözümler ürettiğini ortaya koymaktadır.

Bu çalışmada, MELD veri seti üzerinde metin ve ses modalitelerine dayalı duygu sınıflandırma problemi ele alınmış; metin modalitesi için RoBERTa, ses modalitesi için HuBERT temsilleri kullanılmış ve bu temsiller üzerinde çalışan hafif sınıflandırıcılar aracılığıyla tekil ve çoklu modalite senaryoları değerlendirilmiştir. Çoklu modalite yaklaşımı olarak geç füzyon (late fusion) yöntemi benimsenmiş; metin ve ses modellerinin ürettiği çıktı olasılıkları karar düzeyinde birleştirilerek nihai duygu tahmini elde edilmiştir.

Deneyler 3, 5 ve 7 duygu sınıflı senaryolarda gerçekleştirilmiş ve performans değerlendirmesinde ağırlıklı olarak Weighted F1 (WF1) metriği kullanılmıştır. Elde edilen sonuçlar, geç füzyon yaklaşımının tüm senaryolarda en yüksek genel performansı sağladığını göstermektedir. Buna göre, 3 duygu sınıflı senaryoda WF1=72, 5 duygu sınıflı senaryoda WF1=66 ve 7 duygu sınıflı senaryoda WF1=62 değerleri elde edilmiştir. Sınıf sayısının artmasıyla birlikte performansta gözlenen düşüş, duygu sınıflarının daha ince ayrımlar gerektirmesi ve sınıf dengesizliğinin etkisinin artması nedeniyle beklenen bir durumdur. Özellikle 7 duygu senaryosunda fear ve disgust gibi az temsil edilen sınıflarda belirgin performans kayıpları gözlenmiştir. Buna karşılık, neutral sınıfı hem sınıf dağılımının görece

dengeli olması hem de daha belirgin örüntüler içermesi nedeniyle tüm senaryolarda en yüksek başarıya ulaşan sınıflardan biri olmuştur.

ROC-AUC analizleri, sınıfların ayrıştırılabilirliğinin duygu türüne ve senaryoya bağlı olarak değiştiğini ortaya koymuştur. Özellikle neutral ve surprise sınıflarında daha yüksek AUC değerleri elde edilirken, fear gibi sınıflarda ayrıştırma gücünün daha sınırlı kaldığı görülmüştür. Karışıklık matrisleri incelendiğinde ise hataların büyük ölçüde neutral sınıfı etrafında yoğunlaştığı ve duygu yoğunluğu düşük ifadelerin sıklıkla neutral sınıfına yönlendirildiği tespit edilmiştir.

Genel olarak elde edilen bulgular, metin ve ses modalitelerinin birlikte kullanılmasının, tekil modalite yaklaşımlarına kıyasla daha dengeli ve tutarlı sonuçlar ürettiğini göstermektedir. Geç füzyon yaklaşımı, modalitelerin bağımsız biçimde modellenmesine olanak tanıyarak hem uygulama açısından esneklik sağlamış hem de karar düzeyinde gerçekleştirilen birleştirme sayesinde genel sınıflandırma başarısını artırmıştır.

Bu çalışma MELD veri seti ile sınırlandırılmıştır. Gelecek çalışmalarda, farklı çok modlu duygu veri setlerinin birlikte kullanılmasıyla daha geniş, çeşitli ve temsil gücü yüksek bir eğitim ortamı oluşturulabilir. Bu kapsamda IEMOCAP, CMU-MOSEI ve benzeri çok modlu veri setlerinin bir arada kullanıldığı bütünleşik eğitim senaryoları oluşturularak, modellerin farklı konuşma tarzları, bağlam yapıları ve duygu ifade biçimleri karşısındaki genelleme yeteneği sistematik olarak incelenebilir.

Bunun yanı sıra, bu çalışmada yalnızca metin ve ses modaliteleri kullanılmıştır. Gelecek çalışmalarda görüntü (visual) modalitesinin sisteme entegre edilmesiyle yüz ifadeleri, mimikler ve beden dili gibi görsel ipuçlarının da duygu tanıma sürecine dâhil edilmesi mümkündür. Bu doğrultuda, yüz ifadelerine dayalı temsil çıkarımı için CNN tabanlı modeller veya Vision Transformer (ViT) tabanlı mimariler kullanılarak metin + ses + görüntü üçlü çok modlu sistemler geliştirilebilir.

Bu çalışmada çoklu modalite entegrasyonu geç füzyon yaklaşımıyla gerçekleştirilmiştir. Gelecek çalışmalarda erken füzyon (early fusion) ve hibrit füzyon (hybrid fusion) stratejileri de deneysel olarak değerlendirilerek, özellik düzeyinde birleştirme ile karar düzeyinde birleştirmenin karşılaştırmalı analizleri yapılabilir.

Buna ek olarak, son yıllarda geliştirilen Transformer tabanlı çok modlu mimariler kullanılarak modaliteler arası etkileşimin doğrudan model mimarisi içerisinde öğrenildiği bütünleşik yapılar oluşturulabilir.

Ayrıca, çalışmaya özgü çok modlu duygu veri setlerinin oluşturulması önemli bir araştırma yönü sunmaktadır.

Son olarak, farklı dillerde toplanacak veri setleri ile çok dilli ve çapraz-dil duygu analizi çalışmalarının gerçekleştirilmesi, modelin dil bağımsızlığı ve kültürler arası genellenebilirliğinin incelenmesi açısından önemli bir araştırma alanı oluşturmaktadır.

KAYNAKÇA

Akar, F., & Akgül, İ. (2022). Derin Öğrenme Modeli ile Yüz İfadelerinden Duygu Tanıma. *Journal of the Institute of Science and Technology, 12*(1), s. 69-79.

Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41(2)*, s. 423-443.

Belhaouari, S. B., Islam, A., Kassoul, K., & Al-Fuqaha, A. (2024). Oversampling techniques for imbalanced data in regression. *Expert systems with applications*, s. 124118.

Bilotti, U. B. (2024). Multimodal Emotion Recognition via Convolutional Neural Networks: Comparison of different strategies on two multimodal datasets. *Engineering Applications of Artificial Intelligence, 130*, s. 107708.

Biroğul, S., Sönmez, Y., & Güvenç, U. (2007). Veri Füzyonuna Genel Bir Bakış. *Politeknik Dergisi, 10(3)*, s. 235-240.

Burges, J. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery, 2(2)*, s. 121–167.

Casale, S., Russo, A., Scebba, G., & Serrano, S. (2008, August 12). Speech Emotion Classification Using Machine Learning Algorithms. *2008 IEEE International Conference on Semantic Computing*, s. 158-165.

Chen, Y. (2024, May). Sentiment Analysis of Medical Text Based on Deep Learning. *2024 9th International Symposium on Computer and Information Processing Technology (ISCIPT)*, s. 478-482.

Cho, K. V. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv preprint arXiv:1406.1078*.

Chopra, S., Agarwal, P., Ahmed, J., Biswas, S. S., & Obaid, A. (2024). Roberta and BERT: Revolutionizing Mental Healthcare Through Natural Language. *SN Computer Science, 5(7)*, s. 888-889.

Cubuk, E., Zoph, B., Mane, D., Vasudevan, V., & Le, Q. (2019). AutoAugment:Learning Augmentation Strategies from Data. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, s. 113-123.

Çam, N., Dönmez, İ., Bitkiçioğlu, Ö. F., Yediparmak, F. B., Bektaş, E., & Haklıdır, M. (2023, September). Multimodal Speech Emotion and Text Sentiment Analysis. *In 2023 8th International Conference on Computer Science and Engineering (UBMK)*, s. 157-162.

Desai, M., & Shah, M. (2021). An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (MLP) and Convolutional neural. *Clinical eHealth, 4*, s. 1-11.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers*, s. 4171–4186.

Dixit, C., & Satapathy, S. M. (2024). Deep CNN with late fusion for real time multimodal emotion recognition. *Expert Systems With Applications, 240*, s. 122579.

Donatus, R. E., Pal, B., Donatus, I. H., & Chiedu, U. O. (2024). Comparative Analysis of Spectrogram and MFCC Representations for Speech Emotion Recognition Using Machine Learning. *Asian Journal of Computer Science and Technology*, s. 41-47.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion, 6*(3-4), s. 169–200.

El Ayadi, M. K. (2011). Survey on speech emotion recognition. *Pattern Recognition, 44*(3), s. 572–587.

Erdoğan, A. A. (2021, TEMMUZ). YÜKSEK LİSANS TEZİ- Bilgisayar Mühendisliği Anabilim Dalı. *DERİN ÖĞRENME YÖNTEMİ İLE EL YAZISI*. KONYA, TÜRKİYE: NECMETTİN ERBAKAN NİVERSİTESİ FEN BİLİMLERİ ENSTİTÜSÜ.

Fujihira, K., & Horibe, N. (2020, September). Multilingual Sentiment Analysis for Web Text Based on Word to Word Translation. *2020 9th International Congress on Advanced Applied Informatics (IIAI-AAI)*, s. 74-79.

Gadzicki, K., Khamsehashari, R., & Zetzsche, C. (2017). Early vs. late fusion in multimodal convolutional neural networks., (s. 1-6).

Gadzicki, K., Khamsehashari, R., & Zetzsche, C. (2020). Early vs Late Fusion in Multimodal Convolutional Neural Networks. *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*.

Go, A., Bhayani, R., & Huang, L. (2009). *Twitter sentiment classification using distant supervision.* CS224N project report, Stanford.

Goodfellow, I. B. (2016). Deep learning. *Cambridge: MIT press, 1*(2).

Gumelar, A. B., Kurniawan, A., Sooai, A. G., Purnomo, M. H., Yuniarno, E. M., Sugiarto , I., . . . Fahrudin , T. M. (2019, August). Human Voice Emotion Identification Using. *2019*

IEEE 7th International Conference on Serious Games and Applications for Health (SeGAH), s. 1-8.

Hakdağlı, Ö., & Hakdağlı, E. (2024). Perakende Sektöründe Müşteri Duygularını Anlamak için Çok Modlu Duygu Analizi. *Ekonomi, İşletme ve Bilişim Dergisi*, s. 44-64.

Hsu, W.-N., Bolte, B., Tsai, Y.-H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM transactions on audio, speech, and language processing*(29), s. 3451-3460.

Hu, B., & Li, Y. (2025). MOTIF: A text-aware and disentangled Mamba-based architecture for multimodal sentiment analysis. *Knowledge-Based Systems*, s. 114651.

Jamjoom, A., Karamti, H., Umer, M., Alsubai, S., Kim, T.-H., & Ashraf, I. (2024). RoBERTaNET: Enhanced RoBERTa Transformer Based Model for Cyberbullying Detection With GloVe Features. *IEEE Access*(12), s. 58950 - 58959.

Kaelbling, L., Littman, M., & Moore, A. (1996). Reinforcement Learning: A Survey. *Journal of Arti*, s. 237-285.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (s. 1746–1751). arXiv preprint arXiv:1408.5882.

Kramer, O. (2013). K-Nearest Neighbors. *In Dimensionality reduction with unsupervised nearest neighbors* (s. 13–23). içinde Berlin, Heidelberg: Springer Berlin Heidelberg.

Kumar, P., Agrawal, K., & Dwivedi, N. (2025). Multimodal Affective Communication Analysis: Integrating Text Sentiment with Speech Emotion using Transfer Learning. *14th IEEE International Conference on Communication Systems and Network Technologies 2025*, s. 251-255.

Le, T. (2020, December). An attention-based deep learning method for text sentiment analysis. *2020 International Conference on Computational Science and Computational Intelligence (CSCI)*, s. 282-286.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521(7553)*, s. 436–444.

Liu, B. (2022). *Sentiment analysis and opinion mining.* Springer Nature.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., . . . Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907*.

Mekala, Sreenivas, Bhuvana, M., Gupta, D., Bhatt, V., Kunekar, P., & Manoharan, G. (2023, September). Natural Language Processing and Deep Learning Techniques to Improve

Sentiment Analysis in Social Media Texts. *2023 6th International Conference on Contemporary Computing and Informatics (IC3I)*, s. 1751-1755.

Muataqeem, Sajjad, M., & Kwon, S. (2020). Clustering-Based Speech Emotion Recognition by. *IEEE Access*, s. 79861-79875.

Mucherino, A., Papajorgji , P., & Pardalos , P. (2009). K-nearest neighbor classification. *In Data mining in agriculture* (s. 83-106). içinde New York, NY: Springer New York.

Neumann, M., & Vu, N. T. (2017). Attentive convolutional neural network based speech emotion recognition. *IEEE Spoken Language Technology Workshop (SLT)* (s. 427-433). arXiv preprint arXiv:1706.00612.

Nwe, T. L., Foo, S. W., & De Silva, L. C. (2003). Speech emotion recognition using hidden Markov models. *Speech Communication, 41*(4), s. 603-623.

Omurca, S., Ekinci, E., & Türkmen, H. (2017, September). An annotated corpus for Turkish sentiment analysis at sentence level. *2017 International Artificial Intelligence and Data Processing Symposium (IDAP)*, s. 1-5.

Onan, A. (2022). Türkçe Metin Madenciliği için Çalışan Bellek Bağlantıları Tabanlı Uzun Kısa Süreli Bellek Mimarisi. *Avrupa Bilim ve Teknoloji Dergisi*(33), s. 239-246.

Pandey, D., Niwaria, K., & Chourasia, B. (2019). Machine Learning Algorithms: A Review. *International Research Journal of Engineering and Technology (IRJET), 6*(2), s. 916-922.

Pandey, S., Bhagawati, C., Dandapat , S., & Chakraborty, A. (2019). Surface chlorophyll anomalies associated with Indian Ocean Dipole and El Niño Southern Oscillation in North Indian Ocean: a case study of 2006–2007 event. *Environmental monitoring and assessment*, s. 806-807.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). humbs up? Sentiment classification using machine learning techniques. *Conference on Empirical Methods in Natural Language Processing* (s. 79-87). arXiv preprint cs/0205070.

Pereira, L. M., Salazar, A., & Vergara, L. (2023). A Comparative Analysis of Early and Late Fusion for the Multimodal Two-Class Problem. *IEEE Access*, s. 84283-84300.

Picard, R. W. (1997). *Affective computing.* MIT Press.

Poria, S. C. (tarih yok). Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing, 174*, s. 50-59.

Poria, S., Cambria, E., Howard, N., Huang, G.-B., & Hussain, A. (2016). Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing*, s. 50-59.

Qaisar, S. (2020, October). Sentiment Analysis of IMDb Movie Reviews Using Long Short-Term Memory. *2020 2nd International Conference on Computer and Information Sciences (ICCIS)*, s. 1-4.

Qaisar, S. M., Sadiq, M. S., & Malik, N. (2020). Deep learning for sentiment analysis on textual data. *Journal of Artificial Intelligence and Soft Computing Research, 1*(10), s. 23-34.

Salcedo-Sanz, S., Rojo-Álvarez, J., Martínez-Ramón, M., & Camps-Valls, G. (2014). Support vector machines in engineering: an overview. *WIREs Data Mining and Knowledge Discovery*, s. 234-267.

Schuller, B., Steidl, S., & Stefan, A. (2009). The INTERSPEECH 2009 Emotion Challenge. *INTERSPEECH*, (s. 312-315).

Sharifani, K., & Amini, M. (2023). Machine Learning and Deep Learning: A Review of Methods and Applications. *World Information Technology and Engineering Journal*, s. 3897-3904.

Shorten, C., & Khoshgoftaar, T. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, s. 1-48.

Smirnov, N. V. (2022). Emotion recognition from facial images. *In 2022 International Russian Automation Conference(RusAutoCon)*, s. 116-121.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, s. 1929-1958.

Sun, F., & Chu, N. (2020, November). Text Sentiment Analysis Based on CNN-BiLSTM-Attention Model. *2020 International Conference on Robots & Intelligent System (ICRIS)*, s. 749-752.

Sutton, C. (1992). *Words, science and learning.* McGraw-Hill Education (UK).

Şenel, E. (2023, Ekim). *Derinlemesine Ayrılabilir Evrişim ve Lstm Ağları ile Görüntülerden Anlamsal ifade çıkarma*. KONYA, TÜRKİYE: Master's thesis, Konya Teknik Üniversitesi.

Tadas, Ahuja, C., & Morency, L.-P. (2018). Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, s. 423 - 443.

U. Bilotti, C. Bisogni, M. De Marsico, & S. Tramonte. (2024). Multimodal Emotion Recognition via Convolutional Neural Networks:Comparison of different strategies on two multimodal datasets. *Engineering Applications of Artificial Intelligence*.

Xie, P., Gu, Gu, H., & Zhou, D. (2024, April). Modeling Sentiment Analysis for Educational Texts by Combining BERT and FastText. *2024 6th International Conference on Computer Science and Technologies in Education (CSTE)*, s. 195-199.

Yun, S., & Yoo, C. (2011, February 2). Loss-Scaled Large-Margin Gaussian Mixture Models for Speech Emotion Classification. *IEEE Transactions on Audio, Speech, and Language Processing*, s. 585-598.

Zare, M., Pourghasemi, H. R., Vafakhah, M., & Pradhan, B. (2013). Landslide susceptibility mapping at Vaz Watershed (Iran) using an artificial neural network model: a comparison between multilayer perceptron (MLP) and radial basic function (RBF) algorithms. *Arabian Journal of Geosciences*, s. 2873-2888.

Zbancioc, M., & Feraru, S. (2021, November). Emotion Recognition for Romanian Language Using MFSC Images with Deep-Learning Neural Networks. *2021 International Conference on e-Health and Bioengineering (EHB)*, s. 1-4.

Zhan, Z. (2024, Decenber). Construction of a Psychological Health Monitoring System based on Social Text Sentiment Analysis. *2024 4th International Conference on Mobile Networks and Wireless Communications (ICMNWC)*, s. 1-5.

Zhang, Z., Deng, J., & Schuller, B. (2014). Cooperative learning and its application to audio emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 11*(4), s. 115-126.

Zhao, L. (2025). Design of a multimodal emotion analysis system based on speech and text in business negotiations. *2025 IEEE 3rd International Conference on Image Processing and Computer Applications (ICIPCA)*, s. 1677-1681.

Zhuang, Z., & Yu, C. (2025). Text Sentiment Classification Model using Attention Mechanism and Word Vector Sentiment Computation. *2025 International Conference on Intelligent Computing and Knowledge Extraction (ICICKE)*, s. 1-6.