

**BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ**

**YAPAY SİNİR AĞLARI TEKNOLOJİLERİNİN MULTİMODÜLER VERİ SETLERİ İLE DUYGU ANALİZİ ÜZERİNE PERFORMANS DEĞERLENDİRMESİ**

**BİTİRME PROJESİ**

Bilgisayar Mühendisliği Bölümü

**DANIŞMAN**

Doç. Dr. Ayşe Berna ALTINEL

İSTANBUL, 2025

**MARMARA ÜNİVERSİTESİ**

**TEKNOLOJİ FAKÜLTESİ**

**BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ**

Marmara Üniversitesi Teknoloji Fakültesi Bilgisayar Mühendisliği Öğrencileri Rabia Şevval AYDIN ve Sude Nur TUNGAÇ tarafından “**YAPAY SİNİR AĞLARI TEKNOLOJİLERİNİN MULTİMODÜLER VERİ SETLERİ İLE DUYGU ANALİZİ ÜZERİNE PERFORMANS DEĞERLENDİRMESİ**” başlıklı proje çalışması, 19/06/2025 tarihinde savunulmuş ve jüri üyeleri tarafından başarılı bulunmuştur.

**Jüri Üyeleri**

|  |  |  |
| --- | --- | --- |
| Doç. Dr. Ayşe Berna ALTINEL  Marmara Üniversitesi | **(Danışman)** | ………………… |
| Dr. Öğr. Üyesi Neşe ÖZDEMİR  Marmara Üniversitesi | (Üye) | ………………… |
| Arş. Gör. Nursaç KURT  Marmara Üniversitesi | (Üye) | ………………… |
|  |  |  |

**ÖNSÖZ**

Bitirme tezi süresi boyunca karşılaştığımız tüm zorluklarda sabırla yanımızda olan, bilgi ve deneyimleriyle bize yol gösteren, tez konumuzun şekillenmesinde yönlendirmeleri ile katkı sağlayan çok değerli danışman hocamız Sayın Doç. Dr. Ayşe Berna ALTINEL’e en içten teşekkürlerimizi sunarız.

Projenin hazırlanması ve lisans sürecimiz boyunca yardımını esirgemeyen tüm öğretim üyelerimize, maddi ve manevi desteklerini esirgemeyen ailelerimize ve çalışma arkadaşlarımıza teşekkür ederiz.

**İÇİNDEKİLER**

**1. GİRİŞ 1**

**1.1.** **Problem Tanımı 1**

**1.2. Projenin Amacı ve Hedefleri 1**

**1.3. Proje Takvimi 2**

**2.**  **MULTİMODÜLER YAPIDAKİ DİYALOG VERİSİ ÜZERİNDE YAPAY SİNİR AĞLARI KULLANILARAK DUYGU ANALİZİ GERÇEKLEŞTİRME YÖNTEMLERİ 2**

**2.1.** **Literatür Taraması 3**

**2.2.** **Veri Seti Analizi 4**

**2.2.1 MELD Veri Seti 4**

**2.2.2 IEMOCAP Veri Seti 7**

**2.3. Önerilen Yöntem ve Modeller 8**

**2.4. Model Eğitimi ve Değerlendirme Teknikleri 11**

**2.4.1. Deney Ortamı ve Kullanılan Araçlar 11**

**2.4.2. Model Eğitim Süreci 12**

**2.4.3 Model Değerlendirme Süreci 13**

**3.** **BULGULAR VE TARTIŞMA 14**

**3.1.** **Yapılan Çalışmalar 14**

**3.2. Sonuçların Değerlendirilmesi 15**

**4.** **SONUÇLAR 16**

**5. GELECEK ÇALIŞMALAR İÇİN ÖNERİLER 16**

**6. KAYNAKÇA 18**

**ÖZET**

**YAPAY SİNİR AĞLARI TEKNOLOJİLERİNİN MULTİMODÜLER VERİ SETLERİ İLE DUYGU ANALİZİ ÜZERİNE PERFORMANS DEĞERLENDİRMESİ**

Bu tez çalışmasında, diyaloglarda duygu analizi yapılmasının zorluğu ile ilgili problemlere çözüm olarak yapay sinir ağları teknolojilerinin performansları incelenmiştir. Metin, ses ve görüntü olmak üzere farklı modaliteleri sağlayan veri setleri, insan duygularının daha doğru ve kapsamlı bir şekilde analiz edilebilmesi için gereklidir. Bu yapıya sahip MELD ve IEMOCAP veri seti üzerinde farklı sinir ağı modelleri kullanılarak deneyler gerçekleştirilerek duygu analizi konusunda en optimal modellerin tespit edilmesi için çalışmalar yapılmıştır.

Elde edilen sonuçlar genellikle duygu analizi çalışmalarında birden çok modalitenin aynı anda kullanıldığı durumların tekli modaliteye göre daha başarılı performans gösterdiğini kanıtlamıştır.

**Haziran, 2025 Sude Nur TUNGAÇ**

**Rabia Şevval AYDIN**

**ABSTRACT**

**PERFORMANCE EVALUATION OF ARTIFICIAL NEURAL NETWORK TECHNOLOGIES ON EMOTION ANALYSIS WITH MULTIMODAL DATASETS**

In this thesis, performance of artifical neural network Technologies was evaluated as a solution to challenges of emotion analysis in dialogues. Datasets with multiple modalities, such as text, audio and visual data, are essential for achieving more accurate and comprehensive understanding of human emotions. To find optimal model types, experiments were conducted using various neural network models on MELD and IEMOCAP datasets, both having multimodular structures.

The results demonstrate that in most cases, using multiple modalities simultaneoulsy gives better performance compared to single modality approaches.

**June, 2025 Sude Nur TUNGAÇ**

**Rabia Şevval AYDIN**

**SEMBOLLER**

**:** aktivasyon fonksiyonu  
 **:** toplam sembolü

**KISALTMALAR**

**1D-CNN :** one-dimensional convolutional neural network

**ATOMIC :** atlas of machine commonsense

**AVEC :** audio/visual emotion challenge

**BC-LSTM / bcLSTM:** bidirectional contextual long short-term memory

**CMN :** conversational memory network

**CMU-MOSEI :** carnegie mellon university - multimodal opinion sentiment and emotion intensity

**CMU-MOSI :** carnegie mellon university - multimodal opinion sentiment intensity

**CMU-MOUD :** carnegie mellon university - multimodal opinion utterance dataset

**COMET :** commonsense transformers

**ConGCN :** contextual graph convolutional network

**COSMIC :**commonsense knowledge for emotion identification in conversations

**DialogueGCN :** dialogue graph convolutional network

**DialogueRNN :** dialogue recurrent neural network

**EmbraceNet :** embedding-based multimodal fusion network

**ERC :** emotional recognition in conversations

**FaceNet :** face recognition network

**GCN :** graph convolutional network

**GPT :** generative pre-trained transformer

**GRU :** gated recurrent unit

**ICON :** improved context-aware network

**IEMOCAP :** interactive emotional dyadic motion capture

**Masked NLL-Loss :** masked negative log-likelihood loss

**MELD :** multimodal emotion lines dataset

**MFN :** memory fusion network

**MOSEI :** multimodal corpus of sentiment intensity

**MOUD :** multimodal opinion utterances dataset

**openSMILE :** open-source speech and music interpretation by large-space extraction

**RNN :** recurrent Neural Network

**RoBERTa :** robust optimized BERT pretraining approach

**SEMAINE :** sensitive artificial listener multimodal database

**text-CNN :** text-based convolutional neural network

**waveRNN :** waveform recurrent neural network

**ŞEKİL LİSTESİ**

**Şekil 1**. MELD veri setinden diyalog örneği 5

**Şekil 2.** MELD veri setindeki konuşmacıların ifade dağılımı yüzdeleri 6

**Şekil 3.** MELD veri setindeki konuşmacıların duygu dağılımları 6

**Şekil 4.** MELD veri setindeki konuşmacıların his dağılımları 7

**Şekil 5.** IEMOCAP veri setindeki duyguların metot gruplarına göre dağılımları 8

**Şekil 6.** DialogueRNN mimarisi ve t zamanını ifade etmek üzere bir diyalogdaki 9

küresel, konuşmacı, dinleyici ve duygu durumlarının güncellenme şeması

**TABLO LİSTESİ**

**Tablo 1.** Proje takvim tablosu 2

**Tablo 2.** MELD veri setinin duygu sınıflarına ait örnek sayıları5

**Tablo 3.** MELD veri setinde base modellerin ve DialogueRNN modelinin 14

duygu sınıflandırmasında ait doğruluk ve f1-skorları

**Tablo 4.** MELD ve IEMOCAP veri setinde DialogueRNN ve COSMIC 15

modellerinin duygu sınıflandırmasına f1-skor değerleri

# GİRİŞ

## Problem Tanımı

Günümüzde duygu analizi sağlık alanında psikolojik rahatsızlıkların tespiti ve tedavisi, hukuk alanında suç analizi ve ifade analizi, pazarlama alanında kullanıcı tepkisi, kişi davranış tespiti gibi çeşitli amaçlar ile kullanılmaktadır. Geleneksel duygu analizi çalışmaları genellikle tek tip veri türüne odaklanmaktadır. Konuşmacılardan alınan ifadelerde metin verisi üzerine duygu analizi, mimikler kullanılarak görsel veri ile duygu analizi veya konuşma kayıtlarından ses verisi ile duygu analizi yapılması gibi. Bu tez çalışması duygu analizinin diyaloglar üzerine uygulanmasına odaklanacaktır.

Diyaloglar insan doğası gereği çoklu modüler yapıdadır. Konuşmacılar sözlü ifadelerine ek olarak mimik, ses tonu, ve vücut hareketleri kullanarak duygularını karşı tarafa aktarırlar. Bu sebeple karşılıklı konuşmalarda duygu tespiti (ERC-Emotional Recognition in Conversations) çalışmalarında bu yapıya uygun veri seti ve yöntemlerinin kullanılması gerekmektedir [1].

Yıllar içerisinde ses, metin ve görsel veriler ile yapılan çalışmalar ile duygu analizi konusunda gelişmeler elde edilmesine rağmen diyaloglar üzerine olan çalışmalar yetersiz kalmaktadır. Bu durumun en büyük nedeni diyaloglar üzerine kapsamlı veri setleri bulunmamasıdır. Veri eksikliği sorununu çözmek için [1] tarafından diyaloglar üzerine çoklu modaliteye sahip geniş örnek sayısına sahip veri seti MELD (Multimodal EmotionLines Dataset) sunulmuştur. Bu çalışmada MELD veri setinin yanı sıra bu alanda önde gelen veri setlerinden biri olan IEMOCAP veri seti kullanılacaktır.

## Projenin Amacı ve Hedefleri

Projenin temel amacı multimodüler duygu analizi için en iyi performans gösteren yapay sinir ağı yöntemlerini bulmaktır. Bu doğrultuda öncelikle MELD ve IEMOCAP veri setinin farklı modaliteleri incelenerek tek bir modalite üzerinden temel analiz yöntemleri araştırılacaktır. Ardından bu modaliteler birleştirilerek, çoklu modaliteler üzerine duygu analizlerinin performansları incelenecektir. Proje süresince her aşamada, her bir modalite için çeşitli yapay sinir ağları modelleri test edilerek analiz değerlendirme raporu oluşturulacaktır. Bu rapor ile her bir modalite için optimal yöntemin bulunması ve farklı yöntemlerin birleştirilerek en iyi sonucu sağlayacak olası multimodal model yapısına dair çıkarımlar yapılabilmesi hedeflenmektedir.

## Proje Takvimi

Projenin geliştirilmesi esnasında ana görevlerin gerçekleştirilme sırası ve ne kadar sürecekleri ile ilgili hazırlanan proje takvimi aşağıdaki gibidir.

**Tablo 1.** Proje takvim tablosu

|  |  |
| --- | --- |
| **Proje Aşaması** | **Tarih** |
| Literatür taraması | 13/01/2025-26/01/2025 |
| Veri setinin analizi | 27/01/2025-09-02-2025 |
| Modalitelerden özellik çıkarılması | 10/02/2025-02/03/2025 |
| Tekli modaliteye sahip modellerin oluşturulması | 03/03/2025-30/03/2025 |
| Multimodal model yapısının oluşturulması | 31/03/2025-03/05/2025 |
| Analiz performans raporunun oluşturulması | 04/05/2025-18/05/2025 |
| Sonuçların değerlendirilmesi | 19/05/2025-04/06/2025 |
| Rapor ve sunum hazırlığı | 05/06/2025-19/06/2025 |

# MULTİMODÜLER YAPIDAKİ DİYALOG VERİSİ ÜZERİNDE YAPAY SİNİR AĞLARI KULLANILARAK DUYGU ANALİZİ GERÇEKLEŞTİRME YÖNTEMLERİ

Multimodüler duygu analizinde farklı modalitelerden (ses, metin ve görüntü) gelen verilerin birleştirilmesiyle tekli modalite yaklaşımlarından elde edilen başarıyı arttırması hedeflenir. Ses verisi tonlama, perde ve vurgu gibi akustik özellikler taşırken, metin verisi diyalog hakkında anlamsal bilgiler sunar. Görüntü verisinin içeriğinde ise jest ve mimikler bulunur. Bu modalitelerin entegre edilerek bir arada kullanılması ile duygu analizinde daha yüksek doğruluk elde edilir. Bu bağlamda, temel olarak kullanılan üç çıktı modeli bulunur. Bu çıktı modelleri şu şekildedir:

* Erken Birleşme (Early Fusion): Farklı modaliteler için ham verilerin işlenmeden birleştirilmesinin ardından modelin girdi katmanına verilmesi şeklinde gerçekleşir.
* Geç Birleşme (Late Fusion): Her modalitenin ayrı ayrı işlenerek özelliklerinin çıkarılmasının ardından birleştirilmesi durumudur.
* Orta Düzeyde Birleştirme (Hybrid Fusion): Modalitelerin belirli bir seviyeye kadar ayrı ayrı işlenmesinin ardından daha sonrasında bir arada işlenmesiyle gerçekleştirilen yöntemdir [2].

## Literatür Taraması

Yapay sinir ağları ile çoklu modüler yapıda duygu analizi yapılması güncel ve önemli görülen bir konudur. Literatürde bu konuda yapılmış birçok araştırma ve yöntem bulunur.

Duygu analizinde diyalog sürecini anlamak ve duygu akışını takip etmek için [2] tarafından yapılan çalışmada geliştirilmiş DialogueRNN modelinden bahsedilir. Geliştirilen DialogueRNN modeli ve araştırmada incelenen diğer modeller IEMOCAP ve AVEC veri setleri üzerinde test edildiğinde en başarılı performansı DialogeRNN varyantlarından biri olan BiDialogRNN+Attn yapısı gösterir.

İkili diyalogların yanında çok kişili diyaloglar üzerinde duygu analizi için [1] tarafından yapılan çalışmada MELD veri seti üzerinde test edilen üç ana model vardır. Bu modeller text-CNN, bcLSTM ve DialogueRNN modelleridir. text-CNN konuşmanın bağlamını dikkate almaz, yani ifadelerin sırasını veya önceki ifadelerin durumlarını kullanmaz. İki yönlü RNN kullanan bcLSTM ise konuşmayı bir bütün olarak işler, konuşmacı değişikliğini dikkate almaz. DialougeRNN ise çok kişili diyaloglarda konuşmacı durumlarını takip ederek her bir konuşmacı için özel bağlam oluşturur.

[3] tarafından yapılan çalışmada bağlam duyarlılığı bağımlılıklarının yanı sıra konuşmacı duyarlılığı bağımlılıkları da dikkate alınır. Bu özelliklere odaklanan bir model olan ConGCN modeli geliştirilir. Modelde graf bazlı evrişimli sinir ağı kullanılır. Graf tabanlı modellemede her bir ifade ile her bir konuşmacı için birer düğüm bulunur. Bağlamsal bağımlılık için aynı konuşmadaki düğümlerin ilgili kenarları birbirine bağlanırken konuşmacı bağımlılığının sağlanmasında ifade düğümü ile konuşmacı düğümü arasında bir bağlantı oluşturulur. Oluşturulan GCN modelinin performansının MFN, BC-LSTM, CMN, ICON ve DialogueRNN modellerinin performansları ile karşılaştırıldığında en yüksek başarıyı verdiği görülür.

[4] tarafından yapılan çalışmada MELD veri setinde bulunan üç veri türü için de derin öğrenme modülleri yapılandırılarak ince ayar yapılır. Metin modalitesinin ön işlemesinde GPT, ses modalitesinin ön işlemesinde WaveRNN, görüntü modalitesinin ön işlemesinde ise FaceNet modalitesinden yararlanılır. Duygunun tahmini aşamasında çapraz modalite füzyon transformatörü ve füzyon için EmbraceNet mimarisi kullanılır.

[5] tarafından yapılan çalışmada ise farklı veri setleri üzerinde metin verisi ile yapılan çalışmalarda metnin bağlam durumu, içsel durum, dışsal durum, niyet durumu ve duygu durumu değişkenleri kullanılır. Ayrıca RoBERTa large modelinin ince ayar edilerek bağımsız özellik çıkarımında kullanılmasının başarıyı artırdığı görülür.

[6] tarafından yapılan çalışmada konuşmalardaki duygu tanıma problemi (ERC) ele alınmıştır ve bu probleme çözüm olarak konuşmayı hem konuşmacılar arasındaki bağımlılık hem de kendi kendine bağımlılık incelemesi ile duygusal olarak sınıflandıran DialogueGCN modeli tanıtılmıştır.

## Veri Seti Analizi

**2.2.1 MELD Veri Seti**

Multimodüler yapıda duygu analizi için pek çok veri seti bulunmaktadır. Fakat bu veri setlerinin çoğunluğu tekli ifadeler içermesi sebebiyle sınırlı çalışma imkânı sunmaktadır. Örneğin CMU-MOSEI, CMU-MOSI ve CMU-MOUD veri setleri sadece tekli ifadelerden oluştuğundan dolayı diyalog üzerine analiz imkânı sağlayamamaktadır. IEMOCAP VE SEMAINE veri setleri ikili ifadeler içerdiklerinden diyaloglar üzerine çalışma imkânı sağlamaktadır ancak örnek sayları MELD veri setine göre çok daha azdır. Örnek sayısına ek olarak IEMOCAP ve SEMAINE veri setleri iki kişilik diyaloglar içerirken MELD veri seti ikiden fazla kişiden oluşan diyaloglara yer verdiği için daha kapsamlı bir çalışma fırsatı sunmaktadır.

MELD veri seti hazırlanırken diyaloglar video klipleri ile beraber değerlendirilmiştir. Veri seti Friends adlı televizyon serisine ait 1433 diyalogdan alınmış 13000 ifadeden oluşur. Her bir ifade için 7 duygu (sinir, tiksinti, üzüntü, mutluluk, nötr, şaşkınlık, korku) etiketine ek olarak duygular pozitif, negatif ve nötr sınıflarında gruplandırılmıştır. Sinir, tiksinti, üzüntü, korku duyguları negatif, mutluluk pozitif, nötr ise nötr sınıfı içerisinde gruplanmıştır. Şaşkınlık hem pozitif hem de negatif olarak ifade edilebilen bir duygu örneğidir. Şekil 1’de veri setine ait örnek bir diyalog yer almaktadır.

metin, insan yüzü, ekran görüntüsü, kişi, şahıs içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil 1. MELD veri setinden diyalog örneği

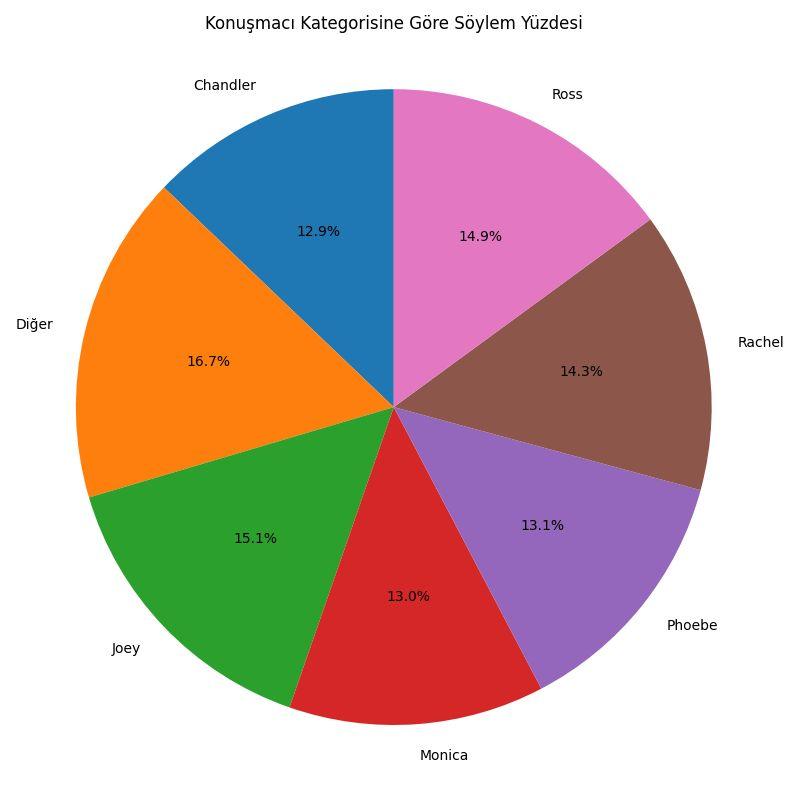
**Şekil 1.** MELD veri setinden diyalog örneği

Tablo 2’de veri setinde her bir duygu sınıfı için kaç adet ifade olduğu gösterilmektedir.

**Tablo 2.** MELD veri setinin duygu sınıflarına ait örnek sayıları

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Duygular | Nötr | Şaşkınlık | Korku | Üzüntü | Sevinç | Tiksinti | Sinir |
| MELD | 6436 | 1636 | 358 | 1002 | 2308 | 361 | 1697 |

Konuşmacılara göre ifade sayısı Şekil 2’de yer almaktadır. Ana karakterlere ait olan konuşmalar haricindeki konuşmalar ‘Diğer’ başlığı altında toplanmıştır. Konuşmacılara göre duygu dağılımı ise Şekil 3’te görülmektedir. Konuşmacılara göre his dağılımı Şekil 4’te bulunmaktadır. Bu görsellerin incelenmesi sonucunda veri serinde nötr sınıfına ait verilerin diğer sınıflara oranla daha fazla olduğu gözlemlenmektedir. Korku ve tiksinme sınıflarında ise bu oranın düşük olduğu görülmektedir.



metin, diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, ekran görüntüsü içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Şekil 2. MELD veri setindeki konuşmacıların ifade dağılımı yüzdeleri

Şekil 3. MELD veri setindeki konuşmacıların duygu dağılımları

metin, diyagram, ekran görüntüsü, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.**Şekil 4.** MELD veri setindeki konuşmacıların his dağılımları

**2.2.2 IEMOCAP Veri Seti**

Interactive Emotional Dyadic Motion Capture (IEMOCAP) veri seti 5 erkek, 5 kadın olmak üzere 10 kişilik bir grup tarafından gerçekleştirilen ikili diyalog çiftlerinden oluşur. 151 diyalog verisinde 2 konuşmacı için de video görüntüleri alınarak 302 video kaydı oluşturulmuştur. Her bir diyalog ifadesi sinir, heyecan, korku, üzüntü, şaşkınlık, mutluluk, memnuniyetsizlik, hayal kırıklığı ve nötr olmak üzere 9 duygu sınıfı ile etiketlenmiştir. Her bir diyalog için ifadelerin yazılı verisi, konuşmacıların ses kayıtları, video kayıtları olmak üzere üç modalite bulunur.

Toplamda 10039 tane veri örneği bulunur. Bu veri örnekleri iki farklı metot altında gruplanır: bir senaryoya bağlı gerçekleşen diyaloglar ve doğaçlama olarak gerçekleştirilen diyaloglar. İfadelerin 4784 tanesi doğaçlama, 5255 tanesi senaryoya bağlıdır. Şekil 5’te her bir grup için duygu etiketlerinin dağılım yüzdeleri gösterilmiştir.[7]

diyagram, daire, renklilik, metin içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**Şekil 5.** IEMOCAP veri setindeki duyguların metot gruplarına göre dağılımları

## Önerilen Yöntem ve Modeller

**DialogueRNN / RoBERTa+DialogueRNN:** DialogueRNN modeli multimodüler duygu sınıflandırmasında diyaloğun bağlamını anlamak ve duygu akışını takip etmekte verimli bir yöntem olarak karşımıza çıkar. Bu modelin üç modülü bulunmaktadır.

* Küresel durum (Global GRU), önceki ifadelerin ve konuşmacının durumunun dikkate alınması ile genel bağlamın temsil edilmesidir. , küresel bağlamdaki zaman adımının durumunu temsil eder. , güncel girdinin vektörüdür. , önceki zaman adımındaki küresel bağlam vektörünü ifade eder.

|  |  |
| --- | --- |
|  | (1) |

* Konuşmacı durumu (Speaker GRU), konuşmacının önceki durumunun ve konuşmasının bağlamının ifadesidir. , belirli bir konuşmacının belirli bir andaki durumudur. Eğer bir konuşmacı , zamanında konuşuyor ise durumu aşağıdaki gibi güncellenir.

|  |  |
| --- | --- |
|  | (2) |

* Duygu durumu (Emotion GRU) kısmında ise konuşmacının durumunun ve önceki ifadelerin duygusal bağlamının birleştirilmesinin sonucunda duygusal temsilin oluşturulmasıdır. , zamanındaki duygusal durumu temsil eder.

|  |  |
| --- | --- |
|  | (3) |

metin, diyagram, ekran görüntüsü, plan içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.metin, ekran görüntüsü, diyagram, yazı tipi içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**Şekil 6.** DialogueRNN mimarisi ve t zamanı ifade etmek üzere bir diyalogdaki küresel, konuşmacı, dinleyici ve duygu durumlarının güncellenmesi şeması

Şekil 2.5’te M katılımcısı olan bir konuşmada kişi 𝑖 konuşmacıdır ve kişiler 𝑗∈[1,M] ve 𝑗≠𝑖 dinleyicidir. Mevcut ifade bu ifadelerin bir fonksiyon ile modellenmesinden elde edilir [2]

DialogueRNN varyantları test edildiğinde en yüksek performansı BiDialogueRNN+Att varyansı elde eder. RoBERTa-large modelinin ince ayar edilmesi ile özellik çıkarılmasının ardından DialogueRNN kullanımında ise performansın artması söz konusudur [5].

**COSMIC:** İkiden fazla konuşmacının yer aldığı diyaloglar üzerinde yapılan duygu analizi görevinin karmaşıklığı için COSMIC bu probleme çözüm olarak duygu analizi işleminde konuşmacının ifadesine ait içerik, bağlam ve sezgisel bilgilerden faydalanmıştır. Mevcut konuşma tabanlı duygu tanıma modellerinin temel üç sorununa odaklanmaktadır: yetersiz bağlam takibi, duygu geçişlerinin yetersiz takibi ve yakın duygular arasında ayrım güçlüğü. COSMIC modeli bu konuşmalarda geçen ifadelerin her birini bir olay olarak ele alır. Bu olayın kişinin niyeti, zihinsel durumu ve olayın sonucu gibi çıkarımların kullanılmasını önerir. ATOMIC veri kümesi üzerinde eğitilmiş COMET adlı bilgi çıkarım modeli aracılığıyla elde edilen çıkarımlar kullanılır. COSMIC mimarisinde dört ana bileşen bulunur:

1. Girdi Temsili: Diyaloglar ( , ), (, ), ... , ( , ) yerleştirilir. Burada bir ifadeyi, ​ ise konuşmacıyı temsil eder. Her bir ​ için hedef, uygun duygu sınıfı ​'nin tahmin edilmesidir.
2. Commonsense Bilgi Çıkarımı: İfadelerin her biri için COMET modeli, üç tür sezgisel çıkarım yapar. Mental state () konuşmacının o anda nasıl hissettiğini temsil eder. Intent () bu ifadenin niyetidir. Event effect (​) ise olayın başkasına veya kendine etkisi nedir.

COMET, ATOMIC bilgi kümesi üzerinde eğitilmiş dil modeli tabanlı bir yapı olup, doğal dil ifadelerinden bu soyut kavramları çıkarabilmektedir.

1. Diyalog Akışı ve Bağlam Temsili: COSMIC, konuşmacı bazlı bağlam izleme için DialogueRNN benzeri bir yapı kullanır. Her konuşmacının geçmişi ayrı ayrı GRU tabanlı durum vektörleriyle takip edilir. Bu durumda, her bir ifadenin kendisiyle birlikte mental durum (mental state), niyet (intent) ve durum etki (event effect) vektörleri de modele dahil edilir. Bu sayede konuşmanın akışında yalnızca kelimeler değil, o ifadenin arka planındaki psikolojik ve sezgisel bilgiler de modele dahil edilir.

|  |  |
| --- | --- |
|  | (4) |
|  | (5) |

1. Duygu Sınıflandırması: Son adımda, her vektörü (bağlamsal temsil), softmax katmanı üzerinden sınıflandırılır:

|  |  |
| --- | --- |
|  | (6) |

Eğitim sırasında klasik çapraz entropy kayıp fonksiyonu (= - ∑ ) kullanılır [5].

Proje kapsımda bahsedilen modellerin veri setlerin üzerindeki performansları test edilecektir.

## Model Eğitimi ve Değerlendirme Teknikleri

Model eğitimi süreci verilerin ön işlenmesi, model mimarisinin belirlenmesi, verilerin bu modellere uygun boyuta getirilmesi, modellerin parametrelerinin uygun şekilde ayarlanması ve optimizasyon işlemlerinin yapılması, model performansının değerlendirilmesi adımlarını içerir. Bu bölümde modelin eğitimi ve değerlendirilmesi kısımları ile ilgili detaylı bilgiler verilecektir.

## Deney Ortamı ve Kullanılan Araçlar

Bu çalışmada gerçekleştirilen deneysel çalışmalar, iki farklı bilgisayar ortamında gerçekleştirilmiştir. Kullanılan sistemlerin özellikleri aşağıda belirtilmiştir:

Bilgisayar-1 (MacBook Pro):

* İşlemci (CPU): Apple M3 Pro
* Bellek (RAM): 18GB
* Grafik Kartı (GPU): Apple GPU
* İşletim Sistemi: macOS

Bilgisayar-2 (Dell G5 5500):

* İşlemci (CPU): Intel® Core™ i7-10750H CPU @ 2.60GHz
* Bellek (RAM): 16 GB
* Grafik Kartı (GPU): NVIDIA GeForce GTX 1660 Ti (6GB), Intel® UHD Graphics
* İşletim Sistemi: Windows 10 64-bit

Çalışma sırasında verilerin işlenmesi, modellerin eğitilmesi ve değerlendirilmesinde kullanılan başlıca kütüphaneler aşağıdakiler gibidir:

Conda: Python ortamlarını ve paket bağımlılıklarını yönetmek için kullanılan bir paket yöneticisidir.

NumPy: Hesaplamalar için kullanılan temel bir Python kütüphanesidir.

Matplotlib: Veri görselleştirme amacıyla kullanılan bir kütüphanedir.

PyTorch: Derin öğrenme modellerinin kurulması, eğitilmesi ve test edilmesi için kullanılan bir yapay sinir ağı kütüphanesidir.

Scikit-learn: Makine öğrenmesi algoritmaları ve model değerlendirme metrikleri için kullanılan bir Python kütüphanesidir.

Pandas: Veri analizi ve veri işleme süreçlerinde kullanılan bir kütüphanedir.

TensorFlow: Yapay sinir ağı modellerinin kurulması ve eğitilmesi amacıyla kullanılan bir derin öğrenme kütüphanesidir.

## Model Eğitim Süreci

MELD ve IEMOCAP veri seti üzerinde uygulanan deneylerde çeşitli yöntemler uygulanmıştır. Verilerden özellik elde etme aşamasında yazılı veriler için önceden eğitilmiş GloVe vektörleri ve 1D-CNN kullanılarak metin özellikleri elde edilmiştir. Ses verileri için openSMILE aracı kullanılmıştır. Deneylerde görsel modülü video bazlı konuşmacı tespiti probleminden dolayı kullanılmamıştır.

Model eğitimi için yapay sinir ağları tabanlı modeller ve çoklu modüler yapıya sahip modeller tercih edilmiştir. Model mimarisinde metinsel ve ses verilerini birlikte ve ayrı ayrı işleyebilen yapılar tercih edilmiştir. Bundan dolayı CNN tabanlı bir text-CNN modeli, çift yönlü RNN tabanlı bir bcLSTM modeli ve 3 adet GRU yapısı kullanan DialogueRNN modeli eğitilmiştir. Aynı zamanda verilerden öznitelik çıkarımı aşamasında RoBERTa modeli kullanılarak elde edilen öznitelikler DialogueRNN modeline verilmiştir. Benzer bir şekilde RoBERTa ile işlenen veriler COSMIC modeli kullanılarak duygu sınıflandırması yapılmasında da kullanılmıştır.

Eğitim sürecinde modellerin aşırı öğrenmesinin önüne geçilmesi adına erken durdurma (early stopping) ve dropout teknikleri kullanılmıştır. text-CNN ve bcLSTM modellerinde optimizasyon algoritması için ‘Adam’ seçilmiştir. Kayıp fonksiyonu olarak ise ‘kategorisel çapraz entropi kaybı (categorical crossentropy)’ tercih edilmiştir. DialogueRNN ve COSMIC modellerinde ise bu modellerden farklı olarak kayıp fonksiyonunda ‘Masked NLL Loss’ tercih edilmiştir.

## Model Değerlendirme Süreci

Modellerin eğitim sürecinin tamamlanmasının ardından model performansının doğru ve kapsamlı bir şekilde değerlendirilebilmesi için model başarısını ve genelleme yeteneğini ölçerken bazı temel ölçüt ve tekniklerden yararlanılmıştır. Bu değerlendirme, modelin yalnızca eğitim verisi üzerindeki başarısını değil aynı zamanda daha önce görmediği veriler üzerinde nasıl genelleme yapabildiğini anlamak açısından önemli bir role sahiptir. Modelin değerlendirilmesi eğitim, doğrulama ve test veri setleri üzerinde gerçekleştirilmiştir. Test verisi, eğitim sürecine dahil edilmeden yalnızca son değerlendirme amacı ile kullanılmıştır.

Modelin başarısını ölçmek amacıyla aşağıdaki temel performans metriklerinden faydalanılmıştır:

* Doğruluk (Accuracy): Modelin doğru sınıflandırıldığı örneklerin toplam örnek sayısına oranıdır. Ancak sınıf dengesizliğinin bulunduğu durumlarda yanıltıcı olabilir.
* Hassasiyet (Precision): Belirli bir duygu sınıfı için yapılan doğru pozitif tahminlerin sayısının o sınıf için yapılan toplam pozitif tahminlerin sayısına oranıdır. Yani modelin tahmin ettiği bir duygu sınıfının gerçekten doğru olup olmadığını ölçmektedir.
* Duyarlılık (Recall): Gerçek pozitif örneklerin, toplam pozitif örnek sayısına oranıdır. Yani modelin o sınıfa ait örnekleri ne kadar yakalayabildiğini göstermektedir.
* F1-skoru (F1-Score): Hassasiyet ve duyarlılığın harmonik ortalamasıdır. Dengesiz veri setlerinde en çok tercih edilen ölçütlerden biridir çünkü hem yanlış pozitif hem de yanlış negatif tahminlerin etkisini dengeleyerek genel bir başarı ölçüsü sunmaktadır.
* Karmaşıklık Matrisi (Confusion Matrix): Modelin hangi sınıfları doğru veya yanlış sınıflandırıldığını görselleştirmek için kullanılmaktadır. Her bir satır gerçek sınıfları, her bir sütun ise modelin tahminlerini göstermektedir.

# BULGULAR VE TARTIŞMA

## Yapılan Çalışmalar

Çok konuşmacılı diyaloglarda yapay sinir ağları ve çoklu modüler yapıların kullanılması ile konuşmanın bağlam takibinin yapılabilmesi önem taşımaktadır. Ayrıca farklı veri türlerinden yararlanılarak duygu sınıflandırmasında ulaşılan başarının tekli modaliteler ile yapılan duygu sınıflandırmasına göre daha yüksek başarı oranına sahip olması beklenmektedir. Çalışmada DialogueRNN modelinin MELD veri setindeki metin, ses ile hem metin hem ses verisi üzerinde çalıştırılması, COSMIC modelinin MELD ve IEMOCAP veri setinde yer alan metin verileri için çalıştırılması, text-CNN ve bcLSTM modellerinin ise MELD veri seti için hem metin hem ses verileri üzerinde çalışmasının ardından multimodal çalıştırılması sağlanmıştır. Base modeller ile MELD veri kümesi üzerinde elde edilen sonuçlar Tablo 3’te karşılaştırılmıştır. DialogueRNN modelinin, RoBERTa+DialogueRNN ve RoBERTa+COSMIC kombinasyonlarının MELD ve IEMOCAP veri seti üzerinde elde ettiği sonuçlar ise Tablo 4’te gösterilmiştir.

**Tablo 3.** MELD veri setinde base modellerin ve DialogueRNN modelinin duygu sınıflandırmasına ait doğruluk ve f1-skorları

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Modeller / Duygular | nötr | şaşkınlık | korku | üzüntü | sevinç | tiksinti | sinir | doğruluk |
| Base model text | 0.6498 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.4812 |
| Base model audio | 0.4578 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.3390 |
| Bimodel base text+audio | 0.6518 | 0.0137 | 0.00 | 0.00 | 0.1240 | 0.00 | 0.272 | 0.4816 |

**Tablo 4.** MELD ve IEMOCAP veri setinde DialogueRNN ve COSMIC modellerinin duygu sınıflandırmasına ait f1-score değerleri

|  |  |  |
| --- | --- | --- |
| Modeller / Veri Setleri | MELD  text audio multimodal | IEMOCAP |
| DialogueRNN | 57.27 43.24 57.68 | 60.60 |
| RoBERTa + DialogueRNN | 47.10 47.47 50.40 | - |
| RoBERTa + COSMIC | 64.36 - - | 66.34 |

Çalışmada base model olarak bcLSTM modeli kullanılmıştır. Base modellerin her biri için epoch sayısı 100, batch sayısı 50 olacak şekilde eğitim gerçekleştirilmiştir. DialogueRNN modeli için ise epoch sayısı 100 iken batch sayısı 30 olacak şekilde eğitim gerçekleştirilmiştir. COSMIC modelinin eğitiminde epoch sayısı 60 olarak belirlenmiştir. Batch size ise 32 olarak kullanılmıştır.

## Sonuçların Değerlendirilmesi

Sonuçlar incelendiğinde duygu sınıfları arasında nötr etiketine sahip verilerin genel olarak sınıflandırmada diğer sınıflara göre baskın olduğu görülmektedir. Şaşkınlık, korku, üzüntü, tiksinti gibi daha karmaşık duygulara ait verilerin sınıflandırılabilmesi için modellerin tek başına yetersiz kaldığı, öznitelik çıkarımı yöntemleri kullanılarak bu verilerden elde edilecek özellikler ile sınıflandırmanın yapılması gerektiği görülmektedir. Modalitelerde tekli kullanım ile ikili kullanım karşılaştırıldığında, tek modalite kullanımında sevinç ve sinir duygularında performans göstermediği, ikili modalite kullanımında ise bu durumun iyileştiği gözlemlenir. Bu duygu analizinde farklı modalitelerin beraber kullanılarak sınıflandırma sürecinin desteklenmesi gerektiği görüşünü desteklemektedir. Aynı zamanda metin verisinin kullanımıyla yapılan sınıflandırılmalarda COSMIC modelinin performansının diğer modellerin performansları ile karşılaştırıldığında çok daha iyi olduğu görülmektedir.

# SONUÇLAR

Tez kapsamında şu ana kadar yapılan çalışmalara ait sonuçlar incelendiğinde duygu sınıfları arasında nötr etiketine sahip verilerin genel olarak sınıflandırmada diğer sınıflara göre daha yüksek başarıya sahip olduğu görülmektedir. Şaşkınlık, korku, üzüntü, tiksinti gibi daha karmaşık duygulara ait verilerin sınıflandırılabilmesi için bcLSTM, textCNN ve DialogueRNN modellerinin kapsamlı duygu analizi için tek başına yetersiz kaldığı görülmüştür.

Her ne kadar DialogueRNN modeli bağlamsal bilgilerin kullanımı açısından başarılı bir yapıya sahip olsa da model tekli modaliteler (sadece ses ya da sadece metin) üzerinde çalıştırıldığında sınırlı performans göstermiştir. Ancak çoklu modaliteler ile çalıştırılarak ses ve metin verileri birlikte kullanıldığında modelin genel doğruluğunun arttığı gözlemlenmiştir. Bu durum, çoklu modalitelerin birlikte kullanılmasının duygu sınıflandırma başarısını olumlu yönde etkilediğini göstermektedir.

COSMIC ile yapılan metin temelli duygu sınıflandırmasının diğer modellere göre çok daha başarılı olduğu gözlemlenmiştir. Bu noktada COSMIC modelinin diğer modellere kıyasla daha yüksek başarı göstermesinin asıl nedeni yalnızca ham metin verisini değil, bununla beraber ortam bilgisi ve kişiler arasındaki ilişkisel bağlam gibi ek bilgileri kullanarak karmaşık duyguların anlaşılmasını sağlamasıdır. Modelin sahip olduğu bu özellik modelin özellikle bağlama duyarlı (şaşkınlık, tiksinti gibi) duyguların sınıflandırılmasında daha yüksek başarıya ulaşmasını sağlamıştır.

# GELECEK ÇALIŞMALAR İÇİN ÖNERİLER

Bu çalışma kapsamında elde edilen bulgular, duygu sınıflandırma görevinde özellikle bazı duygu grupları için iyileştirme gerektiren özelliklerin bulunduğunu işaret etmektedir. Özellikle karmaşık duyguların (şaşkınlık, korku, üzüntü, tiksinti gibi) sınıflandırılmasında, kullanılan modellerin performansı diğer sınıfların sınıflandırmalarına kıyasla düşük kalmıştır. Bundan dolayı mevcut yöntemlerin karmaşık duygu sınıflandırması için yeterli olmadıkları sonucuna ulaşılmıştır. Bu duruma ek olarak, veri dengesizliği ve duygular arası semantik benzerliklerin model performansını etkilediği ortaya çıkmıştır. Bu nedenle gelecekteki çalışmalarda veri dengesizliği ile başa çıkabilecek örnekleme stratejileri kullanılması önerilmektedir.

COSMIC modelinin metin tabanlı sınıflandırmada sağladığı yüksek başarıdan dolayı, bu modelin farklı modalitelerle desteklenerek genişletilmesiyle multimodüler veriler için duygu analizinde en optimal modelin elde edilmesi önerilmektedir. Bu sayede hem duygu bağlamının hem de ortam dinamiklerinin daha güçlü bir şekilde yakalanması hedeflenmektedir.

Geliştirilen modellerin gerçek zamanlı diyaloğu analiz edebilen konuşma analiz sistemlerine uygulanması ise akademik araştırmadan pratik uygulamalara geçiş için değerli olacaktır.

1. **KAYNAKÇA**

[1] S. Poria, D. Hazarika, N. Majumder, G. Naik, R. Mihalcea, and E. Cambria, “MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversation,” 2018. doi: 10.48550/arXiv.1810.02508.

[2] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, “DialogueRNN: An Attentive RNN for Emotion Detection in Conversations,” *arXiv preprint*, Nov. 2018. doi: 10.48550/arXiv.1811.00405

[3] D. Zhang, L. Wu, C. Sun, S. Li, Q. Zhu, and G. Zhou, “Modeling both context- and speaker-sensitive dependence for emotion detection in multi-speaker conversations,” in Proc. 28th Int. Joint Conf. Artif. Intell. (IJCAI-19), 2019, pp. 5415–5421. [Online]. Available: https://www.ijcai.org/proceedings/2019/0752.pdf.

[4] C. Bai, S. Kumar, J. Leskovec, M. Metzger, J. F. Nunamaker, and V. S. Subrahmanian, “Predicting the visual focus of attention in multi-person discussion videos,” in *Proc. 28th Int. Joint Conf. Artif. Intell. (IJCAI-19)*, 2019. doi: 10.24963/ijcai.2019/626.

[5] D. Ghosal, N. Majumder, A. Gelbukh, R. Mihalcea, and S. Poria, “COmmonSense knowledge for eMotion Identification in Conversations,” in *Findings of the Association for Computational Linguistics: EMNLP*, 2020, pp. 2470–2481. doi: 10.48550/arXiv.2010.02795.

[6] A. Gelbukh, N. Majumder, S. Poria, N. Chhaya, and D. Ghosal, “DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation,” *arXiv preprint*, Aug. 2019. doi: 10.48550/arXiv.1908.11540.

[7] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee ve S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," Language Resources and Evaluation, cilt. 42, ss. 335–359, Kasım 2008, doi: 10.1007/s10579-008-9076-6.

[8] D. Zhang, L. Wu, C. Sun, S. Li, Q. Zhu, and G. Zhou, “Modeling both context- and speaker-sensitive dependence for emotion detection in multi-speaker conversations,” in Proc. 28th Int. Joint Conf. Artif. Intell. (IJCAI-19), 2019, pp. 5415–5421. [Online]. Available: https://www.ijcai.org/proceedings/2019/0752.pdf.