T.C.
MİMAR SİNAN GÜZEL SANATLAR ÜNİVERSİTESİ
FEN EDEBİYAT FAKÜLTESİ
İSTATİSTİK BÖLÜMÜ

LİSANS TEZİ

MELD ile Çok modlu duygu analizi üzerine bir çalışma

İlke DERCAN

Tez Danışmanı
Prof. Dr. Ozan KOCADAĞLI

İSTANBUL – 2024

ÖZET

Çok modlu veri, birden fazla modaliteden (örneğin, metin, ses, görüntü) elde edilen bilgileri içeren verileri ifade eder. Bu tür veriler, farklı modaliteler arasındaki ilişkileri ve etkileşimleri analiz ederek daha zengin ve anlamlı bilgiler sağlar. Friends dizisinden elde edilen videolardan oluşan MELD veri setinde pencereleme tekniği kullanılarak multimodal duygu analizi için dört farklı füzyon tekniği incelenmiştir: Erken Füzyon, Geç Füzyon, Hibrit Füzyon yaklaşımı altında Ara Katmanlardan Birleştirme Yaparak Hibrit Füzyon ve Dikkat Mekanizmaları Kullanılarak Çapraz Füzyon yaklaşımları uygunlanmıştır. Bu çalışmada kullanılan pencereleme yöntemi, performansı önemli ölçüde artırmıştır. Pencereleme yöntemi, duygu sınıflandırmasında F1 skorlarının iyileştirilmesini sağlamıştır. MELD verisiyle kurulan 4 modelin karşılaştırılmaları yapılmıştır.

ABSTRACT

Multimodal data refers to data that contains information obtained from multiple modalities (e.g., text, audio, video). This type of data provides richer and more meaningful information by analyzing the relationships and interactions between different modalities. Using the MELD dataset, which consists of videos from the TV show Friends, four different fusion techniques were examined for multimodal emotion analysis using the windowing technique: Early Fusion, Late Fusion, and under the Hybrid Fusion approach, Fusion by Merging Intermediate Layers and Cross Fusion Using Attention Mechanisms. The windowing method used in this study significantly improved performance. The windowing method has contributed to the improvement of F1 scores in emotion classification. Comparisons of the four models built with the MELD data were conducted

ÖNSÖZ

Üniversite eğitimim boyunca da bitirme tezim kapsamında da bilgi ve tecrübelerini eksik etmeden her zaman yardımcı ve destek olan tez danışmanım Prof. Dr. Ozan KOCADAĞLI'ya saygı ve teşekkürü borç bilirim.

İÇİNDEKİLER

1. Giriş

1.1 Konunun Tanıtımı ve Önemi

Canlılar için çoklu modalite son derece doğal bir kavramdır. Canlılar, sinyalleri tespit etmek ve aralarında ayrım yapmak, iletişim kurmak, çapraz doğrulama yapmak, belirsizliği ortadan kaldırmak ve hızlı bir şekilde alınması gereken çok sayıda ölüm kalım tercihi ve tepkisine sağlamlık kazandırmak için bazen "duyular" olarak da adlandırılan harici ve dahili sensörleri kullanır. [5] Duygu sınıflaması için de gerçek dünyaya daha benzer bir yaklaşım izleyebilmek için çok modalite kavramından faydalanılmıştır. Aşağıda çalışma kapsamında bahsedilecek olan terimlerin açıklaması verilmiştir.

1.1.1 Modalite

Bir şeyin var olduğu, deneyimlendiği, ifade edildiği ya da yapıldığı kanal. Terim, aralarında tıp, temel bilim, teknoloji, beşeri bilimler ve dilbilim olmak üzere pek çok alanda kullanılır.

1. duyum modalitesi: Duyumun oluştuğu kanal; duyum türü (örn., görme, işitme),

2. tepki modalitesi: Tepkinin oluştuğu kanal; tepki türü (örn., yazılı, sözlü),

3. tanısal modalite: Tıp doktorunun tanıyı koymada kullandığı teknik veya yöntem (örn., bilgisayarlı beyin tomografisi, radyografi, mammografi). [19]

1.1.2 Çok Modlu Veri

Çok modlu veri, birden fazla türdeki verinin (örneğin, metin, görüntü, ses) bir araya getirilmesiyle elde edilir. Farklı modalitelerin sunduğu çeşitli bilgileri içerir. Modaliteler görsel (görüntüler, videolar), metinsel ve işitsel (ses, sesler, müzik) ve diğer biyometrik ve sinyal verileri olabilir. Sıkça karşımıza çıkan modalite kombinasyonları şöyledir:

- Resim + Metin
- Resim + Ses
- Resim + Metin + Ses
- Metin + Ses

1.1.3 Çok Modlu Makine Öğrenimi

Çok modlu makine öğrenimi, birden fazla modaliteden gelen bilgileri işleyebilen ve ilişkilendirebilen modeller oluşturmayı amaçlayan canlı, çok disiplinli bir alandır. Çok modlu kaynaklardan öğrenmek, modaliteler arasındaki ilişkileri yakalama ve doğal fenomenler hakkında derinlemesine bir anlayış kazanma olanağı sunar. [2]

1.1.4 Füzyon

Füzyon, tüm modalitelerden gelen etkileşimi modellemek amacıyla ortak bir temsil oluşturmak için çeşitli modalitelerden gelen girdilerin birleştirilmesi sürecini ifade eder [1].

1.1.5 Çok modlu veri füzyonu (MMDF)

Daha kullanılabilir bir biçimde bilgi üretme amacıyla farklı modalitelerdeki birden fazla heterojen kanaldan gelen çeşitli verileri entegre etme sürecini tanımlar[12].

Yüksek frekansta birden fazla veri kaynağının (video, günlükler, ses, jestler, biyosensörler) toplanması, verilerin senkronize edilmesi ve kodlanması ve öğrenmenin gerçekçi, ekolojik olarak geçerli, sosyal, karma bir biçimde incelenmesi için kullanılabilecek bir dizi teknik olarak tanımlanmaktadır. [3]

1.1.6 Duygu Analizi

Duygu analizi, metin, konuşma veya diğer veri kaynaklarından duygusal tonları veya hisleri belirlemeyi amaçlayan bir tekniktir. Sosyal medyada duygu analizi son yıllarda oldukça popüler hale gelmiştir. NLP alanında gelişmelere bolca katkı sağlamıştır. Kullanıcıların duygularını veya görüşlerini analiz etmek için kullanılır ve birçok alanda kullanılmaktadır. Sosyal medya, pazarlama ve marka analizi, politik analiz, haber analizi, müşteri geri bildirimleri vs.

1.1.7 Multimodal Duygu Analizi

Multimodal duygu analizi, çeşitli modalitelerden (metin, ses, görüntü) gelen verileri birleştirerek duygu analizi yapma sürecidir. Bu yöntem, tek modaliteye dayalı analizlere göre daha doğru ve kapsamlı sonuçlar elde etmeyi amaçlar. Örneğin, bir kişinin hem söylediklerini (metin) hem de nasıl söylediklerini (ses tonu) ve yüz ifadelerini (görsel) analiz ederek, duygusal durumu hakkında daha kesin çıkarımlar yapılabilir. Çok modlu duygu tanıma görevleri önemli bir etkiye sahiptir ve çok sayıda alanda geniş uygulamalar bulunur. Ancak, birden fazla modaliteden duygusal bilgi çıkarma ve yorumlama prosedürü oldukça karmaşıktır. [13] Son yıllarda multimodal duygu analizi çalışmaların geliştirilebilmesi için, çok modlu veriler sunulmuştur. Çok modlu verilerle duygu analizi için birçok farklı yöntem geliştirilmektedir.

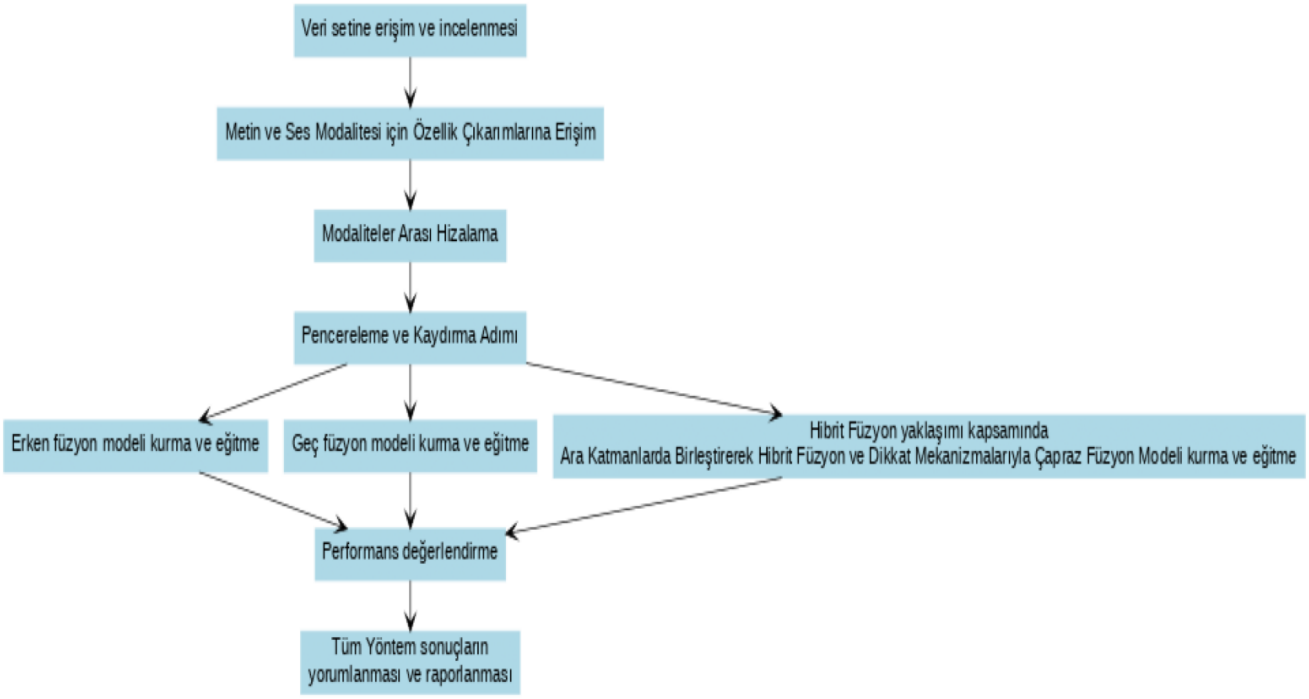
1.1.8 Multimodal Duygu Analizinin Önemi

Multimodal duygu analizi, çeşitli veri türlerini birleştirerek daha zengin ve doğruluğu yüksek analizler yapmamıza olanak tanır. Bu analizler, duygusal tepkilerin daha iyi anlaşılmasını sağlar ve farklı modalitelerden gelen bilgilerin birleşimi, her bir modalitenin tek başına sunabileceğinden daha fazla bilgi sağlamasını amaçlar. [15][16]

2. Çalışmanın Amacı ve Kapsamı

Bu çalışmada, literatürde yaygın olarak kullanılan füzyon mimarileri uygulanmış ve her birinin performansı detaylı bir şekilde analiz edilmiştir. Bununla birlikte, kullanılan pencereleme yöntemi, performansı önemli ölçüde artırmıştır. Pencereleme yöntemi, duygu sınıflandırmasında modaliteler arasındaki ilişkileri daha iyi yakalayarak F1 skorlarının iyileştirilmesini sağlamıştır.

2.1 Çalışmada İzlenecek Yolun Akış Diyagramı



3. Yöntem

3.1 Araştırma Yöntemi ve Tasarımı

Bu çalışmada, MELD (Multimodal EmotionLines Dataset) veri seti kullanılarak duygu tanıma ve farklı füzyon teknikleri incelenmiştir. Bu çalışmada kullanılan temel teknikler ve metodolojiler aşağıda açıklanmıştır.

3.2 Veri Seti

3.2.1 Multimodal EmotionLines Dataset (MELD)

EmotionLines, Chen ve arkadaşları (2018) [17] tarafından geliştirilmiştir. EmotionLines, her bir diyaloğun birden fazla konuşmacıdan ifadeleri içerdiği popüler sitcom Friends'ten diyaloglar içermektedir.[10] MELD, Emo-tionLines veri kümesinin Soujanya Poria, Devamanyu

Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria ve Rada Mihalcea taraflarından yapılan çalışmayla multimodal senaryo için genişletilmesi, iyileştirilmesiyle oluşturulmuştur.

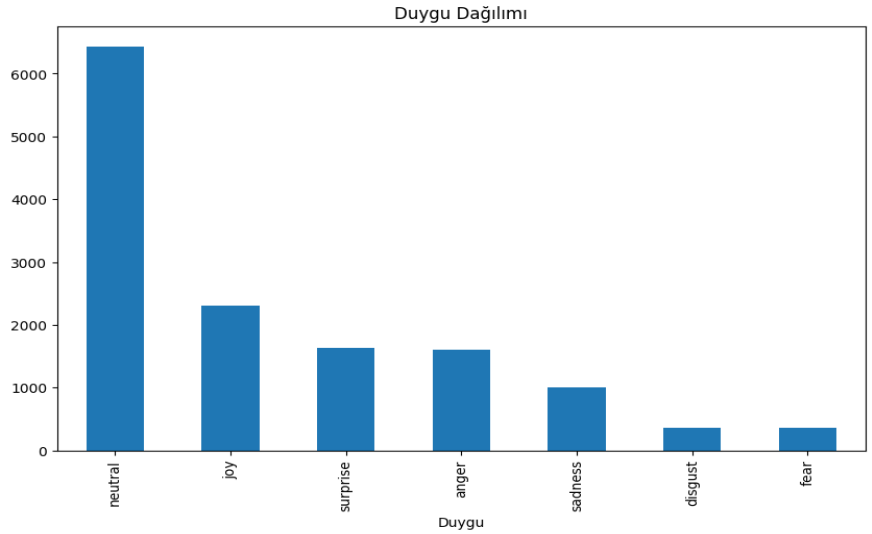
MELD, Friends dizisindeki 1.433 diyalogdan yaklaşık 13.000 ifade içermektedir. Her bir ifade duygu ve duyarlılık etiketleri ile notlandırılmıştır ve işitsel, görsel ve metinsel modaliteleri kapsamaktadır. Veri setinin tamamı http://affective-meld.github.io adresinde kullanıma hazırdır. [10]

3.2.2 Veri Setinin Tanıtımı

MELD veri seti, çeşitli duygusal ifadeler içeren çok modlu diyaloglardan oluşmaktadır. Veri seti, EmotionLines'da bulunan aynı diyalog örneklerini içeren videolardan oluşur. Soujanya Poria ve arkadaşları tarafından MELD veriseti oluşturulurken, videolar izlenmiş ve her bir ifadenin doğru duygu etiketi belirlenmiştir. Bu nedenle, analizlerde metin ve ses modaliteleri için aynı etiketler kullanılmaktadır.

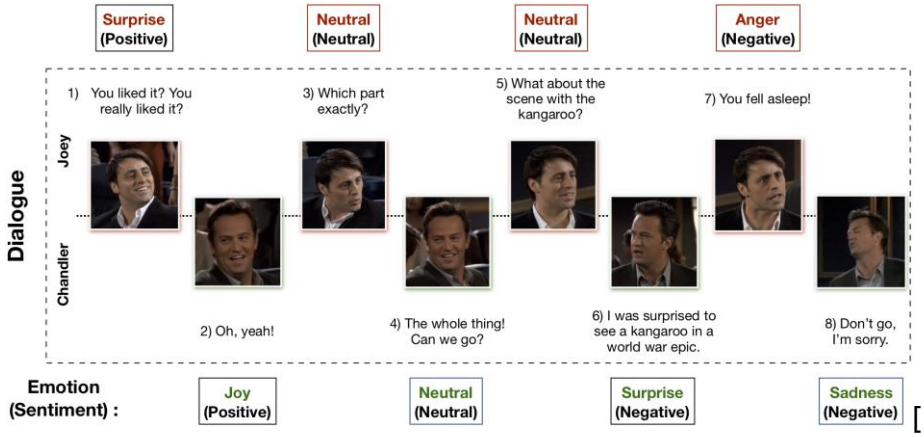
Diyaloglara birden fazla konuşmacı katılmıştır. Bir diyalogdaki her sözce, bu yedi duygudan herhangi biriyle etiketlenmiştir

Bu grafik, MELD veri setindeki duygu sınıflarının dağılımını göstermektedir. Konuşmalar, "neutral" (nötr), "joy" (neşe), "surprise" (sürpriz), "anger" (öfke), "sadness" (üzüntü), "disgust" (tiksinti) ve "fear" (korku) gibi farklı duygu sınıflarına ayrılmıştır. Grafik, nötr duyguların veri setinde en yaygın duygu olduğunu, ardından neşe ve sürpriz duygularının geldiğini göstermektedir. Korku ve tiksinti duyguları ise en az görülen duygular arasında yer almaktadır.



Diyalog Boyunca Duygu Değişimi

Joey ve Chandler arasındaki bir diyalog örneği verilmiştir. Diyalog boyunca duyguların nasıl değiştiğini ve bu değişimlerin metin, ses ve görsel modaliteler aracılığıyla nasıl algılandığını göstermektedir. Bu görsel, çok modlu duygu tanımada bağlamın ve zaman içindeki değişimlerin önemini vurgular.



[10]

Duygu tanıma dilin yapısı, bağlam, jest ve mimik gibi birçok açıdan yorumlanabilmesi zor bir konudur. Örneğin diyalogdaki kinayeyi sadece metin modalitesinden anlayabilmemiz zor bir durumdur. Ancak kişinin ses tonu, söyleyiş biçimi veya o anki görünüş bilgisine erişebildiğimizde o kişinin duygusunu sınıflandırabilmek kolaylaşacaktır. Başka bir açıdan bakacak olursak, görüntü modalitesi için sinirli duygu sınıfına sahip bir kişi gülebilir ve yanlış bir sınıflandırmaya sebebiyet verebilir. Ses ve metin modalitesiyle birlikte değerlendirildiğinde daha doğru yorumlanabilir. Bu görsel, MELD veri setinde yer alan iki farklı cümlenin metin, ses ve görsel modaliteler tek başına değerlendirilseydi nasıl farklı duygular çıkarılabileceğini göstermektedir. Modalitelerden birlikte değerlendirilerek daha doğru ve bağlamsal duygu sınıflandırılması yapılması hedeflenmektedir. Modaliteler arası uyumsuzluk, duygu tanımada zorluklara neden olabilir ve bu da çok modlu öğrenme yöntemlerinin önemini vurgular.

Utterance: *"Become a drama critic!"*

Emotion: *Joy* **Sentiment:** *Positive*

Text	Audio	Visual
Ambiguous	Joyous tone	Smiling Face

Utterance: *"Great, now he is waving back"*

Emotion: *Disgust* **Sentiment:** *Negative*

Text	Audio	Visual
Positive/Joy	Flat tone	Frown

[10]

İlk cümle metin modalitesi düzeyinde, hangi duyguyu taşıdığı konusunda belirsizdir. Diğer modalitelerden edinilen neşeli ton ve gülümseyen yüz bilgisiyle nihai sınıflandırmada neşe ve pozitif duygu olarak sınıflandırılabilmiştir. Metinden olumlu bir duygu taşıyor gibi gözüken ikinci cümle içinse düz tonda ses ve çatık kaş bilgisiyle tiksinti/negatif duygu olarak doğru şekilde sınıflandırılmıştır.

3.2.3 Modaliteler ve Özellik Çıkarımı

Çalışma kapsamında MELD veri setinden elde edilen 2 modalite kullanılacaktır.

I. Metin Modalitesi: Konuşmaların transkriptleri.
II. Ses Modalitesi: Konuşmaların ses kayıtları.

Bu çalışma kapsamında, MELD veri setindeki metin ve ses modaliteleri için Poria ve arkadaşlarının (2017) [10] yaptığı çalışmada çıkarılan özellikler kullanılmıştır. Bu özellikler, MELD resmi web sitesinden erişilebilecek pickle dosyalarında mevcuttur. Poria ve arkadaşlarının yaptığı çalışmada görsel özellikler çıkarılmamıştır. Poria ve arkadaşlarının videolardan elde ettikleri metin ve ses özelliklerinden özellik çıkarımı adımları verilmiştir.

3.2.3.1 Metin Özellikleri

Metin özellikleri, her bir belirteci önceden eğitilmiş 300 boyutlu GloVe vektörleriyle başlatarak ve bu vektörleri bir 1D-CNN'ye besleyerek 100 boyutlu metin özellikleri çıkartılarak elde edilmiştir. [10]

3.2.3.2 Ses Özellikleri

Ses özellikleri, çeşitli düşük seviyeli tanımlayıcılardan ve çeşitli ses ve prozodik özelliklerin istatistiksel fonksiyonellerinden oluşan 6373 boyutlu özellikler çıkaran openSMILE araç seti kullanılarak elde edilmiştir [19]. Ses temsili yüksek boyutlu olduğundan, yoğun bir ses temsili elde etmek için SVM gibi seyrek tahmincilerle L2 tabanlı özellik seçimi yapılmıştır. [10]

3.3 Özelliklerin Hizalanması

Çok modlu verilerle çalışırken farklı modalitelerden elde edilen özelliklerin bir araya getirilmesi ve doğru etiketlerle hizalanması kritik bir adımdır. Modaliteler arası anlamlı bağlılıklar, performans iyileştirmeye yönelik çalışmalara başlayabilmek için farklı kanallardan gelen ama aynı bilgiyi temsil eden veriler hizalanmalıdır.

Hizalama, farklı modaliteler arasındaki doğrudan ilişkileri tanımlama görevini ifade eder. Çok modlu öğrenme alanındaki güncel araştırmalar, modaliteyle değişmeyen temsiller oluşturmayı amaçlamaktadır. Bu, farklı modaliteler benzer bir anlamsal kavrama atıfta bulunduğunda, temsillerinin gizli bir uzayda benzer / birbirine yakın olması gerektiği anlamına gelir. Örneğin, "havuza daldı" cümlesi, bir havuz görüntüsü ve bir sıçrama sesinin ses sinyali, temsil uzayının bir manifoldunda birbirine yakın olmalıdır. [18]

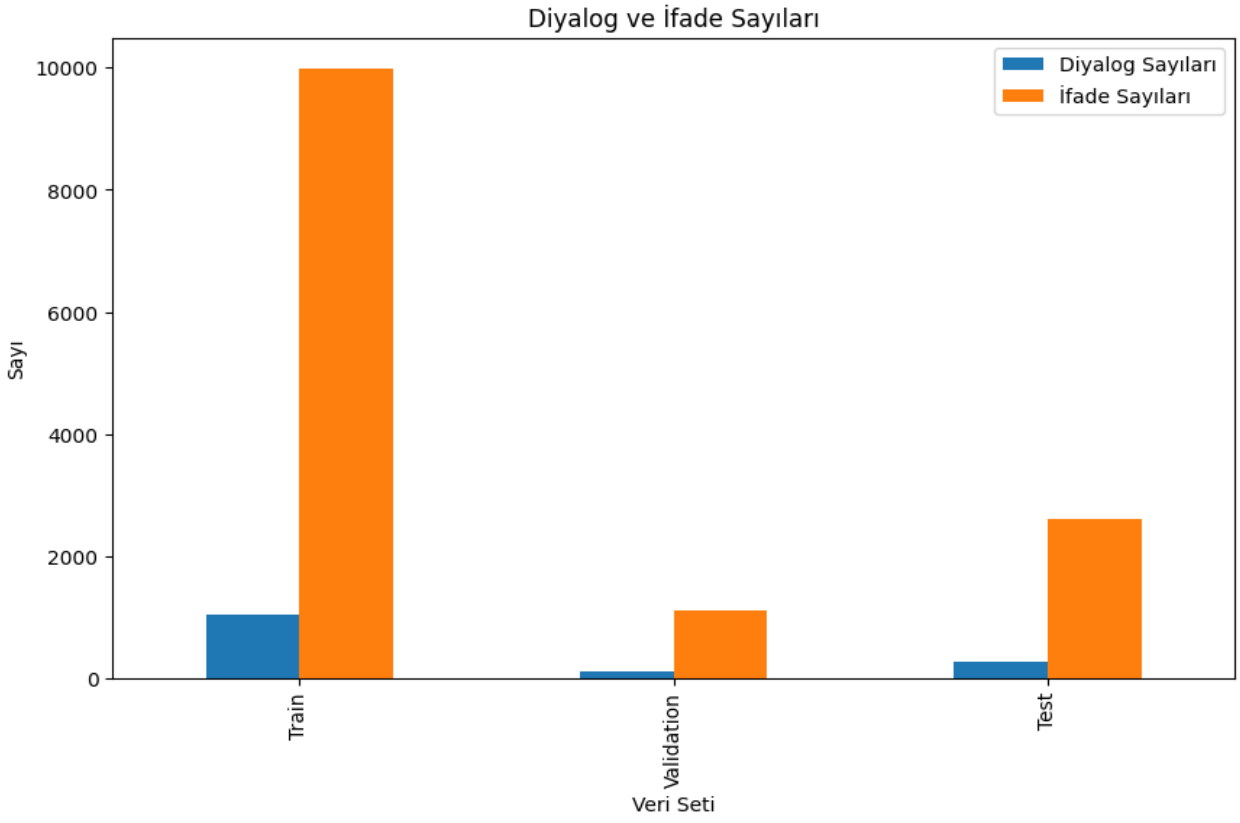
Örneğin, bir görüntü ve bir resim yazısı verildiğinde, görüntünün resim yazısındaki kelimelere veya cümlelere karşılık gelen alanlarını bulmak isteriz [98]. Başka bir örnek olarak, bir film verildiğinde, onu senaryoya veya dayandığı kitap bölümlerine hizalamak [252]. Çok modlu hizalamayı örtük ve açık olarak ikiye ayırıyoruz. Açık hizalamada, alt bileşenleri modaliteler arasında hizalamakla açıkça ilgileniriz, örneğin yemek tarifi adımlarını ilgili talimat videosuyla hizalamak gibi [131]. Örtük hizalama, başka bir görev için ara (genellikle gizli) bir adım olarak kullanılır; örneğin, metin açıklamasına dayalı görüntü bulma, kelimeler ve görüntü bölgeleri arasında bir hizalama adımı içerebilir [99] [2]

Bu çalışma kapsamında Poria ve arkadaşları tarafından çıkarılan özelliklere erişildikten sonra her bir giriş için oluşturulan benzersiz anahtarlar (diyalog ve ifadeler) kullanılarak metin ve ses özellikler doğru şekilde hizalanmıştır.

3.4 Eğitim, Doğrulama ve Test Ayrımı

MELD veri seti ses ve metin modelitesi için, modelin eğitimi, doğrulanması ve test edilmesi amacıyla üç farklı alt kümeye ayrılmıştır. Aynı videolardan çıkartılmış olduğundan ses ve metin modalitelerinin sayıları eşittir. Her iki modalite için de bu alt kümelerdeki veri sayıları ve dağılımlar aşağıdaki gibidir:

	Diyalog Sayıları	İfade Sayıları
Eğitim	**1038**	**9988**
Test	**280**	**2610**
Doğrulama	**114**	**1109**



Eğitim seti, modelin öğrenme süreci için kullanılırken, doğrulama seti modelin performansını izlemek ve hiperparametre ayarlamaları yapmak için kullanılır. Test seti ise modelin genel performansını ve genelleme yeteneğini değerlendirmek için kullanılır.

3.5 Pencereleme ve Kaydırma

Veri setindeki özelliklerin zaman içindeki değişimlerini yakalamak için pencereleme ve kaydırma teknikleri kullanılmıştır.

Pencereleme, zaman serisi verilerinde belirli bir zaman dilimindeki örüntüleri yakalayarak modelin daha etkili öğrenmesini sağlar. Bu teknik, modelin kısa süreli bağımlılıkları öğrenmesine yardımcı olur ve verinin daha anlamlı bir şekilde temsil edilmesine olanak tanır .[7]

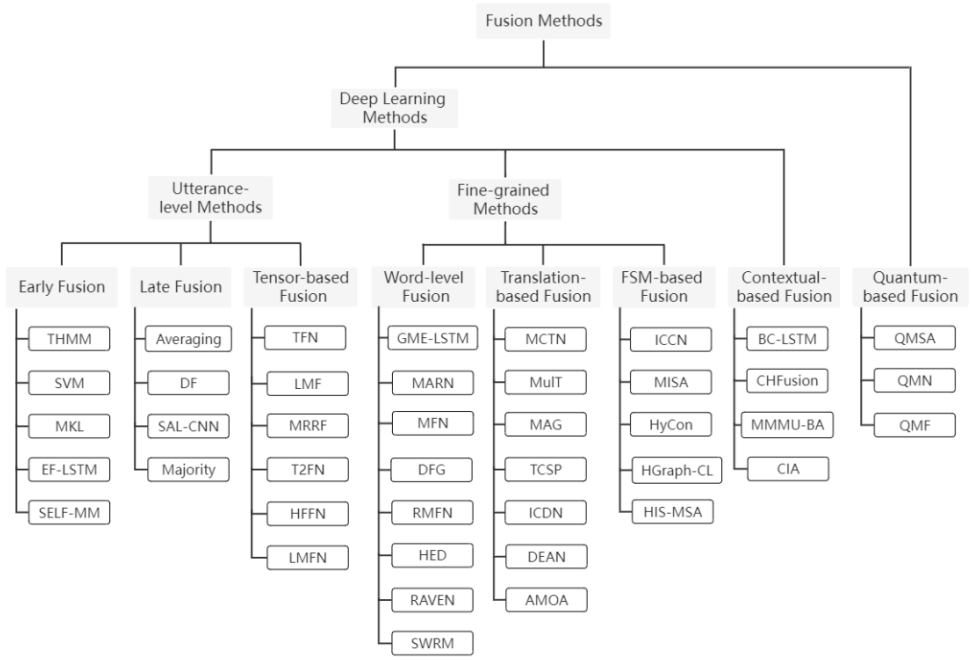
I. Veri Hazırlama: Eğitim, doğrulama ve test verileri ayrı ayrı ele alınarak her bir veri kümesi için kaydırma pencereleri oluşturulmuştur.
II. Pencere Boyutu ve Kaydırma Adımı: Pencere boyutu 5 ve kaydırma adımı 1 olarak belirlenmiştir. Bu, her 5 veriden oluşan alt dizilerin oluşturulmasını ve bir sonraki pencerenin 1 veri kaydırılarak alınmasını sağlar.
III. Etiketleme: Pencerelenmiş verilerin etiketleri, pencerenin sonundaki etiket olarak belirlenmiştir.

Veri	**Pencereleme Sonrası Boyut**
Metin Eğitim	9985, 5, 300
Ses Eğitim	9985, 5, 1611
Metin Doğrulama	1105, 5, 300
Ses Doğrulama	1105, 5, 1611
Metin Test	2606, 5, 300
Ses Test	2606, 5, 1611

3.6 Füzyon Teknikleri

Füzyon modülü, özellik çıkarma işlemi tamamlandıktan sonra her bir modalitenin birleştirilmesinden sorumludur. Füzyon için kullanılan yöntem/mimari muhtemelen başarı için en önemli bileşendir. [18]

Çok modlu füzyon teknikleri oldukça araştırılan bir konudur. Ve çok fazla önerilen füzyon tekniği bulunmaktadır.



Füzyon Çeşitleri [4]

Füzyon tekniklerini temel farklarına göre genel olarak üçe ayırmak mümkündür.

3.6.1 Erken Füzyon

Erken füzyonda, her bir modaliteden gelen çıktı vektörleri bir karar vermek için birleştirilir. Erken füzyonda, her modalitenin öğrenicisinden (veya kodlayıcısından) çıkan çıktı vektörleri bir karar vermek için birleştirilir. [11] Erken füzyonda video özellikleri, ses özellikleri ve metin özellikleri model analizi için doğrudan genel özellik vektörleri olarak birleştirilir. [9] Özetle bu yaklaşım, her modaliteden gelen ham veya çıkarılmış özelliklerin birleştirilerek tek bir özellik vektörü oluşturulması ve bu vektörün model eğitimi için kullanılması şeklinde çalışır. Birden fazla modalitenin ortak temsilini öğrenmesine olanak tanır. Ancak her modalitenin ihtiyacı farklı olabileceğinden tüm modaliteler için tek bir temsille yaklaşmak dezavantajlı olabilir.

Morency ve arkadaşlarının (2011) yaptığı çalışmanın çok modlu verilerle duygu analizi için erken füzyon yaklaşımının ilk örneği olarak söylenebilir. Bu makale, çok modlu duygu analizi görevini ele almakta ve görsel, işitsel ve metinsel özellikleri entegre eden ortak bir modelin Web videolarındaki duyguları tanımlamak için etkili bir şekilde kullanılabileceğini gösteren kavram kanıtlama deneyleri yürütmektedir. Üç modlu duygu analizi görevini ilk kez ele almakta ve bunun görsel, işitsel ve metinsel modalitelerin ortak kullanımından faydalanabilecek uygulanabilir bir görev olduğunu göstermektedir. [8]

3.6.2 Geç Füzyon

Geç füzyonda, her modalitenin kodlayıcısından gelen ara anlamsal bilgiler birleştirilir.[11] Bu yaklaşım, her modaliteye özel modelin kendi giriş türü için en iyi temsili öğrenmeye odaklanmasına olanak tanır. Modaliteye özgü modeller kurulacağından bireysel modalitelere özgü gürültüye karşı daha dayanıklı olabilir. Çıktıların birleştirilmesiyle elde edileceğinden modaliteler arası etkileşimleri kaçırabilir.

3.6.3 Hibrit Füzyon

Hibrit füzyonda, çoklu modal girdileri birleştirmek için entegrasyon prosedürü, farklı seviyelerde erken ve geç füzyonun karışımını içerir.[11] Hem erken hem de geç füzyonun avantajlarından yararlanmayı amaçlar. Bu yaklaşım, modaliteye özel modellerin ara katmanlarındaki özelliklerin önemine göre birleştirilmesini veya her modalitenin önemini dinamik olarak öğrenmek için dikkat mekanizmaları kullanma veya birbirinden bilgileri öğrenebilen çapraz füzyon yaklaşımları gibi birçok yöntem önerilmiştir.
Bazı önerilen yöntemlerden bahsedilmiştir:

Zadeh ve arkadaşlarının Tensor Fusion Network (TFN) modeli, üç modaliteyi tensör işlemleriyle birleştirerek modaliteler arası etkileşimleri etkili bir şekilde yakalamış ve duygu sınıflandırma performansını artırmıştır.

Önerilen TFN modeli üç ana bileşenden oluşur: 1)Modalite Gömme Alt Ağları (Modality Embedding Subnetworks): Tek modlu özellikleri giriş olarak alır ve zengin bir modalite gömme (embedding) çıktısı verir. 2)Tensor Füzyon Katmanı (Tensor Fusion Layer): Modalite gömmelerinden üç katlı Kartezyen çarpım kullanarak tek modlu, çift modlu ve üç modlu etkileşimleri açıkça modeller. 3)Duygu Çıkarımı Alt Ağı (Sentiment Inference Subnetwork): Tensor Füzyon Katmanı'nın çıktısına bağlı olarak çalışır ve duygu çıkarımı yapar. Göreve bağlı olarak ağın çıktısı ikili sınıflandırma, 5 sınıflı sınıflandırma veya regresyon olabilir. (TFR) [14]

RMFN (Tekrarlayan Çok Aşamalı Füzyon Ağ) [40]
Diğer bir yaklaşım ise farklı modalitelerden, yerelden küresele, düşük seviyeden yüksek seviyeye özellikleri kademeli olarak birleştirmek ve nihayetinde kapsamlı bir duygu temsili elde etmek için çoklu tekrarlayan sinir ağı katmanları kullanmaktır. Bu model, farklı modaliteler için anlamsal uzaydaki özelliklerin konumunu ayarlamak için bir dikkat mekanizması kullanır ve aynı kelimenin farklı sözel olmayan davranışlar altında farklı duygular sergilemesine izin verir. [4]

RAVEN (Tekrarlayan Katılımlı Varyasyon Gömme Ağı) [41]
Çok modlu duygu analizi için yerel bir füzyon modülü ve küresel bir füzyon modülü içeren bir Hiyerarşik Füzyon Ağı önerilmiştir. Yerel çapraz modal füzyon, hesaplama karmaşıklığını etkili bir şekilde azaltan kayan bir pencere aracılığıyla araştırılmıştır. [11]

CM-RoBERTa

Bu makalede, konuşulan ses ve ilgili transkriptlerden duygu tespiti için Cross-Modal RoBERTa (CM-RoBERTa) modelini öneriyoruz. CM-RoBERTa'nın çekirdek birimi olarak, paralel kendi-kendine ve çapraz dikkat mekanizması, ses ve metnin modaliteler arası ve içi etkileşimlerini dinamik olarak yakalamak için tasarlanmıştır. [6]

4. Bulgular

4.1 Erken Füzyon Modeli : Multimodal Duygu Analizi

4.1.1 Modalite Özelliklerinin Birleştirilmesi ve Boyutları

Özellik çıkarımı yapılmış metin ve ses verilerini pencereleme işlemi sonrası birleştirerek elde edilen veri setlerinin boyutları aşağıdaki gibidir:

Modelite Türleri	**Alt Küme**	**Boyutlar**
Metin ve Ses	Eğitim	9985, 5, 1911
Metin ve Ses	Doğrulama	1105, 5, 1911
Metin ve Ses	Test	2606, 5, 1911
Etiketler	Eğitim	9985
Etiketler	Doğrulama	1105
Etiketler	Test	2606

4.1.2 Modelin Tanımlanması ve Eğitimi

Erken füzyon modelinde, metin ve ses özelliklerinden oluşan birleşik veri kümesi üzerinde derin öğrenme modelleri kullanılarak duygu analizi yapılmaktadır. Model, metin ve ses özelliklerini işleyerek, verideki örüntüleri öğrenir ve duygu sınıflandırması yapar.

```
Layer (type)                 Output Shape              Param #   Connected to
==================================================================================================
input_2 (InputLayer)         [(None, 5, 1911)]         0         []

bidirectional_2 (Bidirecti   (None, 5, 256)            2088960   ['input_2[0][0]']
onal)

conv1d_2 (Conv1D)            (None, 3, 512)            2935808   ['input_2[0][0]']

batch_normalization_4 (Bat   (None, 5, 256)            1024      ['bidirectional_2[0][0]']
chNormalization)

batch_normalization_3 (Bat   (None, 3, 512)            2048      ['conv1d_2[0][0]']
chNormalization)

dropout_4 (Dropout)          (None, 5, 256)            0         ['batch_normalization_4[0][0]'
                                                                  ]

dropout_3 (Dropout)          (None, 3, 512)            0         ['batch_normalization_3[0][0]'
                                                                  ]

bidirectional_3 (Bidirecti   (None, 128)               164352    ['dropout_4[0][0]']
onal)

conv1d_3 (Conv1D)            (None, 1, 256)            393472    ['dropout_3[0][0]']

batch_normalization_5 (Bat   (None, 128)               512       ['bidirectional_3[0][0]']
chNormalization)

global_average_pooling1d_1   (None, 256)               0         ['conv1d_3[0][0]']
 (GlobalAveragePooling1D)

dropout_5 (Dropout)          (None, 128)               0         ['batch_normalization_5[0][0]'
                                                                  ]

concatenate_1 (Concatenate   (None, 384)               0         ['global_average_pooling1d_1[0
)                                                                 ][0]',
                                                                   'dropout_5[0][0]']

dense_1 (Dense)              (None, 7)                 2695      ['concatenate_1[0][0]']

==================================================================================================
Total params: 5588871 (21.32 MB)
Trainable params: 5587079 (21.31 MB)
Non-trainable params: 1792 (7.00 KB)
```

4.1.3 Model Eğitimi ve Değerlendirmesi

Model, categorical_crossentropy kayıp fonksiyonu ve adam optimizer kullanılarak eğitilir. Erken durdurma yöntemi ile modelin eğitim sırasında aşırı öğrenme yapmasının önüne geçilir. Eğitim sırasında doğrulama veri seti kullanılarak modelin performansı izlenir ve en iyi performans gösterdiği noktada eğitim durdurulur.

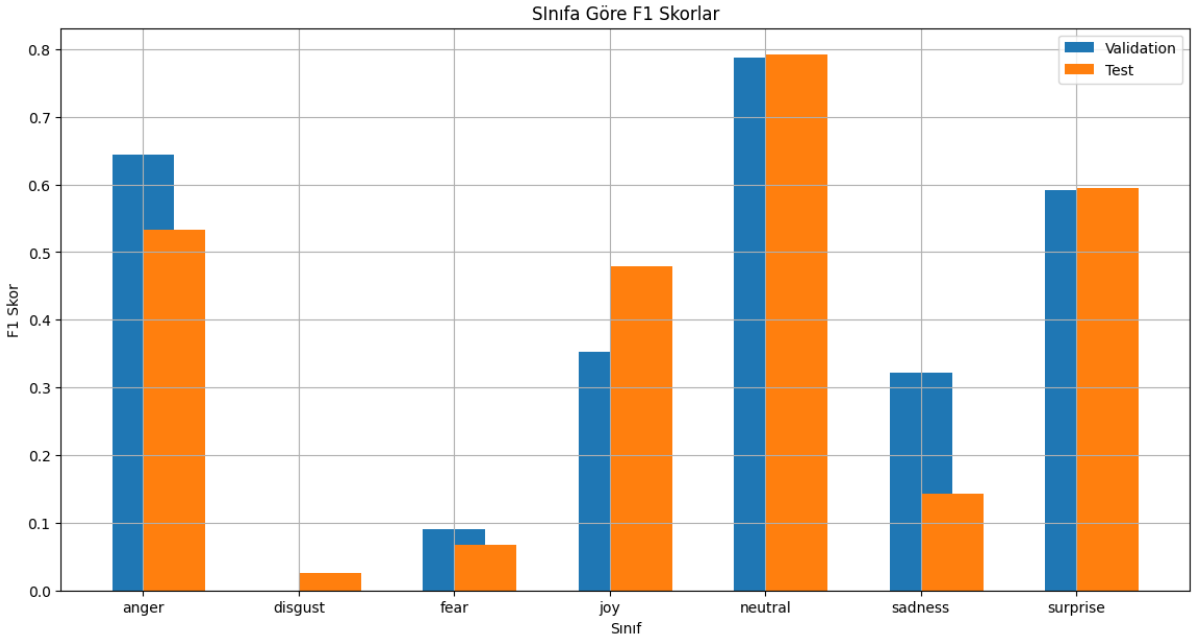
Modelin performansı, test veri setinde F1 skoru kullanılarak değerlendirilir ve sınıflandırma raporu hazırlanır. Bu raporlar, modelin farklı duygu sınıflarında ne kadar başarılı olduğunu gösterir.

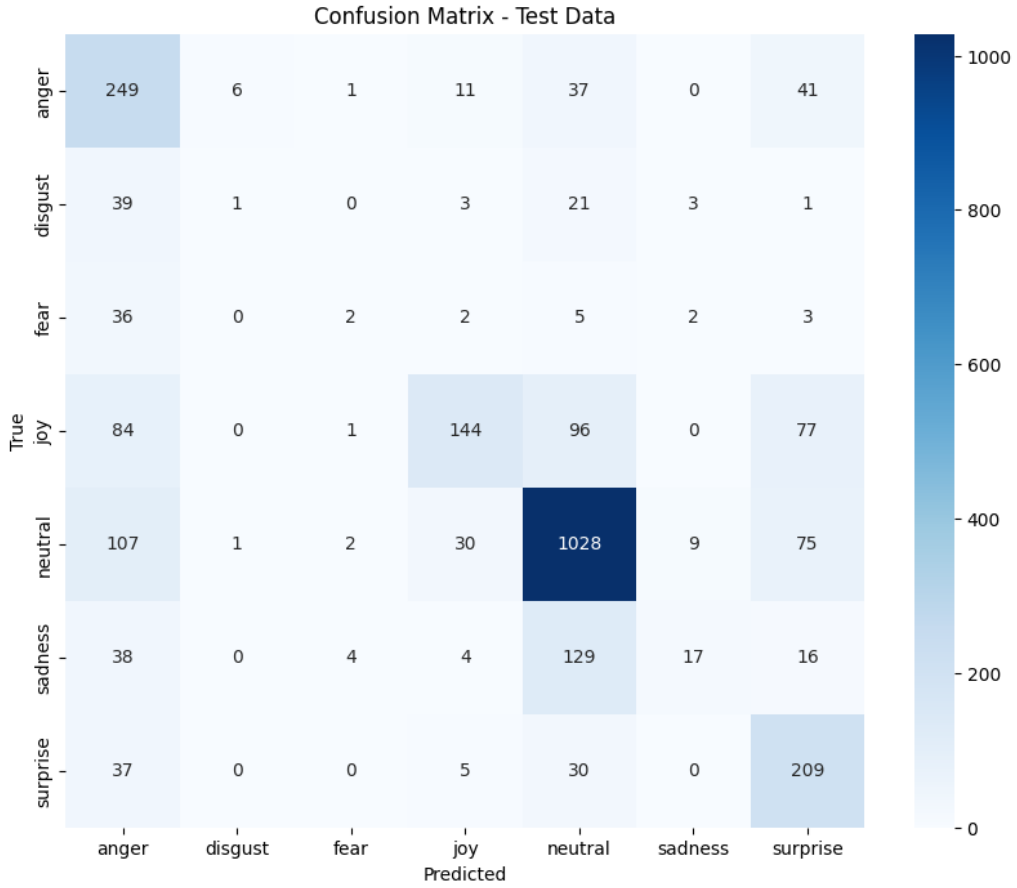
4.1.4 Sonuç

Sınıf	Kesinlik	Duyarlılık	F1-Skoru	Destek
Öfke	0.42	0.72	0.53	345
İğrenme	0.12	0.01	0.03	68
Korku	0.2	0.04	0.07	50
Neşe	0.72	0.36	0.48	402
Nötr	0.75	0.82	0.79	1252
Üzüntü	0.55	0.08	0.14	208
Şaşkınlık	0.50	0.74	0.59	281
Doğruluk			0.63	2606
Makro Ort	0.47	0.40	0.38	2606
Ağırlıklı Ort	0.64	0.63	0.60	2606

Test ile sınıflandırma raporu

Modelin genel performansı, test veri setinden elde edilen ortalama F1 skor ile değerlendirilmiştir: Test Veri Seti Genel F1 Skoru 0.60 bulunmuştur.


SInıfa Göre F1 Skorlar
Validation
Test
F1 Skor
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
anger
disgust
fear
joy
neutral
sadness
surprise
Sınıf


Confusion Matrix - Test Data
True
anger 249 6 1 11 37 0 41
disgust 39 1 0 3 21 3 1
fear 36 0 2 2 5 2 3
joy 84 0 1 144 96 0 77
neutral 107 1 2 30 1028 9 75
sadness 38 0 4 4 129 17 16
surprise 37 0 0 5 30 0 209
anger
disgust
fear
joy
neutral
sadness
surprise
Predicted
1000
800
600
400
200
0

4.2 Geç Füzyon Modeli

Diğer adıyla karar tabanlı füzyon için metin ve ses modalitesi için iki ayrı model kurulur. Ve çıktıları birleştirilerek karar modeli kullanılarak tahmin edilir.

4.2.1 Metin Modeli

Metin modelinin test üzerinde F1 skoru 0.77966158304014 bulunmuştur.

```
Model: "model_4"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_5 (InputLayer)        [(None, 5, 300)]          0

 bidirectional_6 (Bidirecti  (None, 5, 256)            439296
 onal)

 bidirectional_7 (Bidirecti  (None, 128)               164352
 onal)

 dense_9 (Dense)             (None, 256)               33024

 dropout_5 (Dropout)         (None, 256)               0

 dense_10 (Dense)            (None, 7)                 1799

=================================================================
Total params: 638471 (2.44 MB)
Trainable params: 638471 (2.44 MB)
Non-trainable params: 0 (0.00 Byte)
```

4.2.2 Ses Modeli

Ses modelinin test için F1 skoru 0.43 bulunmuştur.

```
Model: "model_2"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_3 (InputLayer)        [(None, 5, 1611)]         0

 bidirectional_4 (Bidirecti  (None, 5, 1024)           8699904
 onal)

 bidirectional_5 (Bidirecti  (None, 512)               2623488
 onal)

 dense_4 (Dense)             (None, 512)               262656

 dropout_2 (Dropout)         (None, 512)               0

 dense_5 (Dense)             (None, 7)                 3591

=================================================================
Total params: 11589639 (44.21 MB)
Trainable params: 11589639 (44.21 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

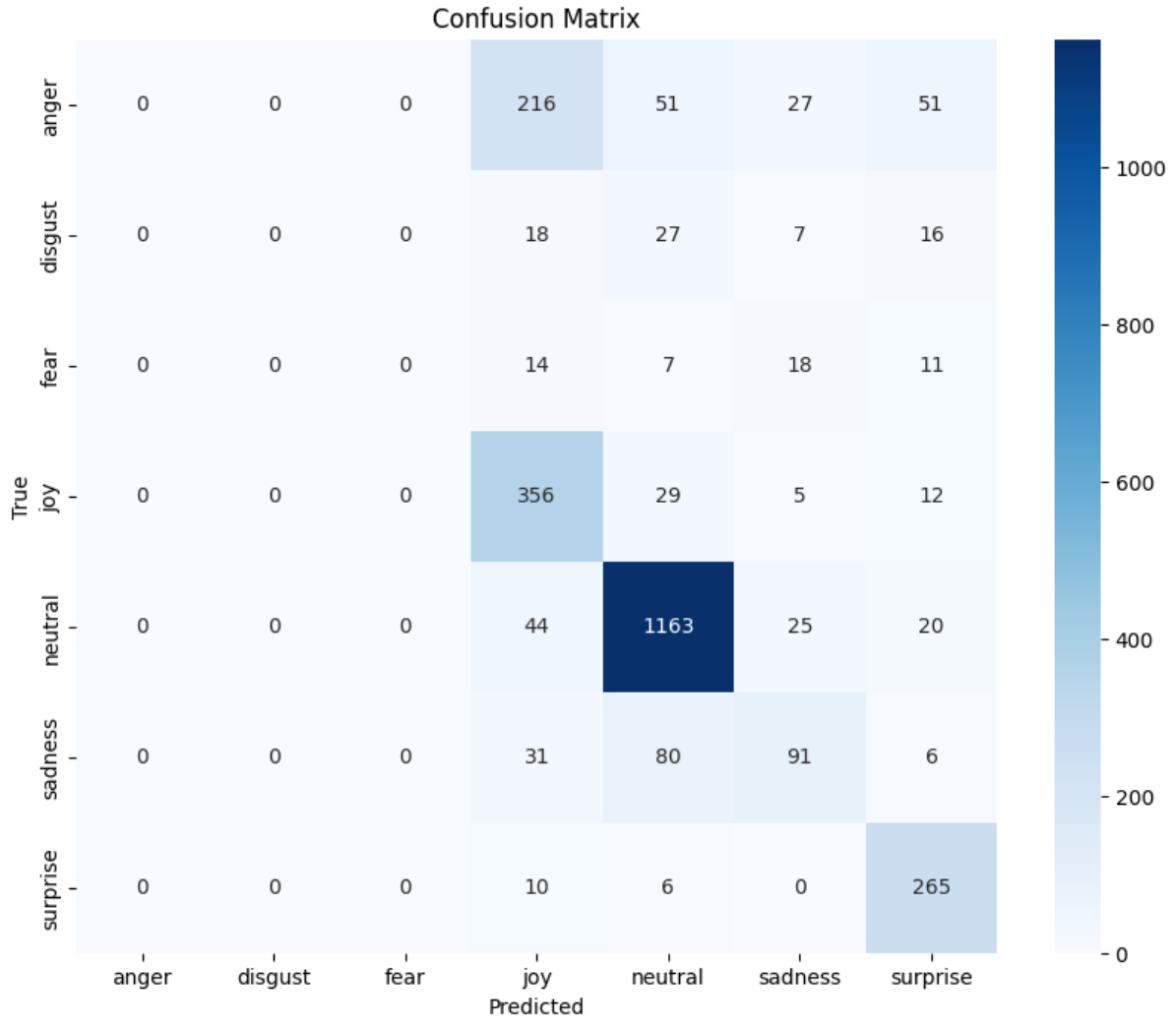
4.3 Geç füzyon (Nihai Karar)

Metin ve ses modellerinin kararları birleştirirek ek katmanlardan geçirilerek karar modeli oluşturulmuştur. Birbirleri arasındaki ilişkiyi yakalayamazlar. Ses modalitesi tek başına iyi performans veremediğinden ses ve metin model çıktıları birleştirerek kurduğumuz model performansı da metinin tek başına verdiği performanstan düşük olmuştur.

Test üzerinde F1 skoru 0.6525 bulunmuştur.

```
Model: "model_5"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_6 (InputLayer)        [(None, 14)]              0

 dense_11 (Dense)            (None, 512)               7680

 dropout_6 (Dropout)         (None, 512)               0

 dense_12 (Dense)            (None, 256)               131328

 dropout_7 (Dropout)         (None, 256)               0

 dense_13 (Dense)            (None, 7)                 1799

=================================================================
Total params: 140807 (550.03 KB)
Trainable params: 140807 (550.03 KB)
Non-trainable params: 0 (0.00 Byte)
```

Sınıf	Kesinlik	Duyarlılık	F1-Skoru	Destek
Öfke	0	0	0	345
İğrenme	0	0	0	68
Korku	0	0	0	50
Neşe	0.52	0.89	0.65	402
Nötr	0.85	0.93	0.89	1252
Üzüntü	0.53	0.44	0.48	208
Şaşkınlık	0.70	0.94	0.80	281
Doğruluk			0.72	2606
Makro Ort	0.37	0.46	0.40	2606
Ağırlıklı Ort	0.61	0.72	0.65	2606



4.3 Hibrit Füzyon

4.3.1 Ara Katmanlar Birleştirilerek Hibrit Füzyon

4.3.1.1 Metin Modeli

- Validation F1 Skoru: 0.759
- Test F1 Skoru: 0.782

```
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_1 (InputLayer)        [(None, 5, 300)]          0

 bidirectional (Bidirection  (None, 5, 512)            1140736
 al)

 bidirectional_1 (Bidirecti  (None, 256)               656384
 onal)

 dense (Dense)               (None, 256)               65792

 dropout (Dropout)           (None, 256)               0

 dense_1 (Dense)             (None, 7)                 1799

=================================================================
Total params: 1864711 (7.11 MB)
Trainable params: 1864711 (7.11 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

4.3.1.2 Ses Modalitesi Modeli ve Sonuçları

- Doğrualama F1 Skoru: 0.447
- Test F1 Skoru: 0.478

```
Layer (type)                 Output Shape              Param #
=================================================================
input_2 (InputLayer)         [(None, 5, 1611)]         0

bidirectional_2 (Bidirecti   (None, 5, 2048)           21594112
onal)

bidirectional_3 (Bidirecti   (None, 1024)              10489856
onal)

dense_2 (Dense)              (None, 512)               524800

dropout_1 (Dropout)          (None, 512)               0

dense_3 (Dense)              (None, 7)                 3591

=================================================================
Total params: 32612359 (124.41 MB)
Trainable params: 32612359 (124.41 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

4.3.1.3 Hibrit Füzyon

Metin ve ses modellerinden elde edilen ara katman çıktıları birleştirilir ve bu birleşik özellikler üzerinden nihai model tahminleri yapılır.Ara katmandaki çıktılar nihai modelin katmanına girdi olarak verilerek model beslenmiştir. Aslında biraz önce mimarisi kurulan metin ve ses modelleri burada tekrar işlemden geçmiştir. Geç füzyonda kararlar kapsamında birleştiriyorduk. Hibritte ara katmanlar birleştirildiğinden daha kompleks bir yapısı vardır.

1. **1. Giriş Katmanları (Input Layers)**
   Metin Girişi ,Ses Girişi
2. **2. Ara Katmanlar (Intermediate Layers)**
   Metin ve ses verisi işleme
3. **Özelliklerin Birleştirilmesi (Concatenation Layer)**
4. **Ek Katmanlar (Additional Layers)**
5. **Çıkış Katmanı (Output Layer)**

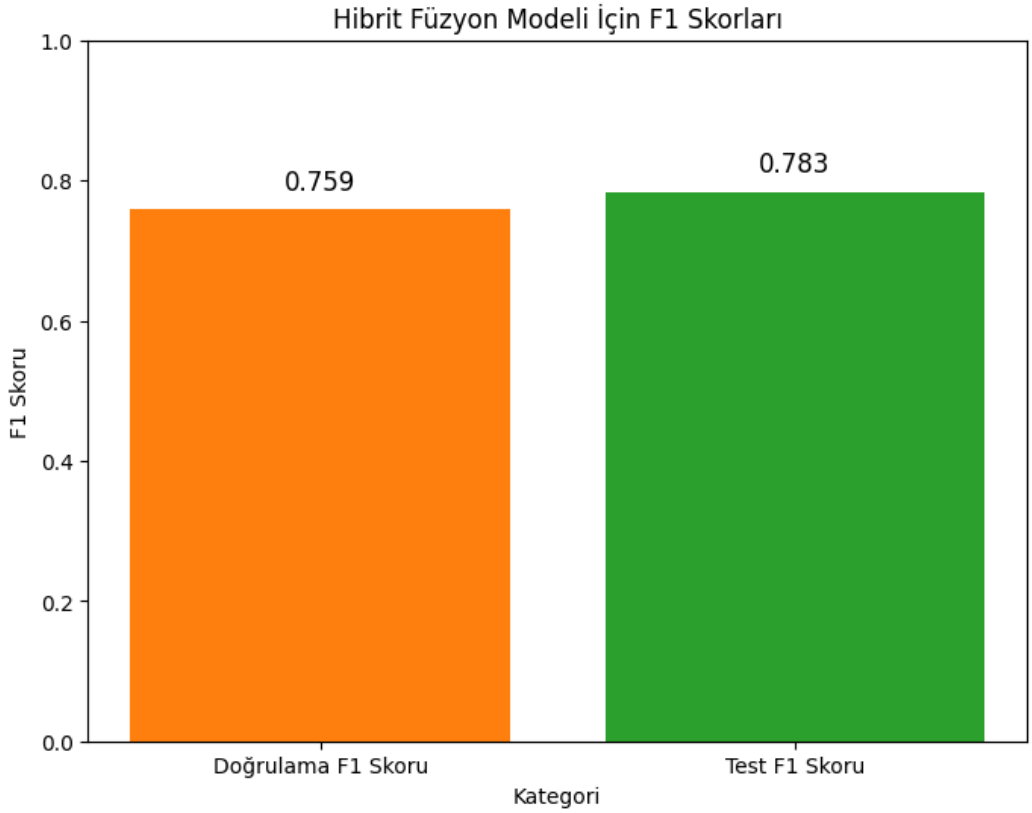
Ara katmanlardan birleştirme yapılarak kurulan hibrit füzyon performansı:

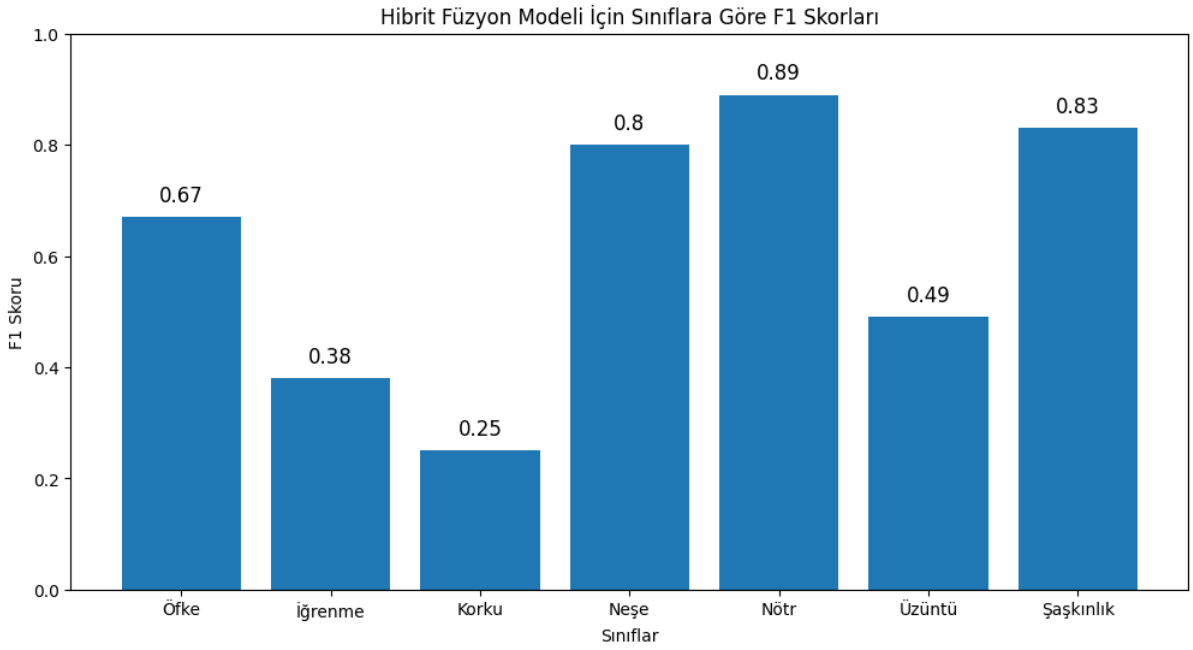
- Doğrulama F1 Skoru: 0.759
- Test F1 Skoru: 0.783

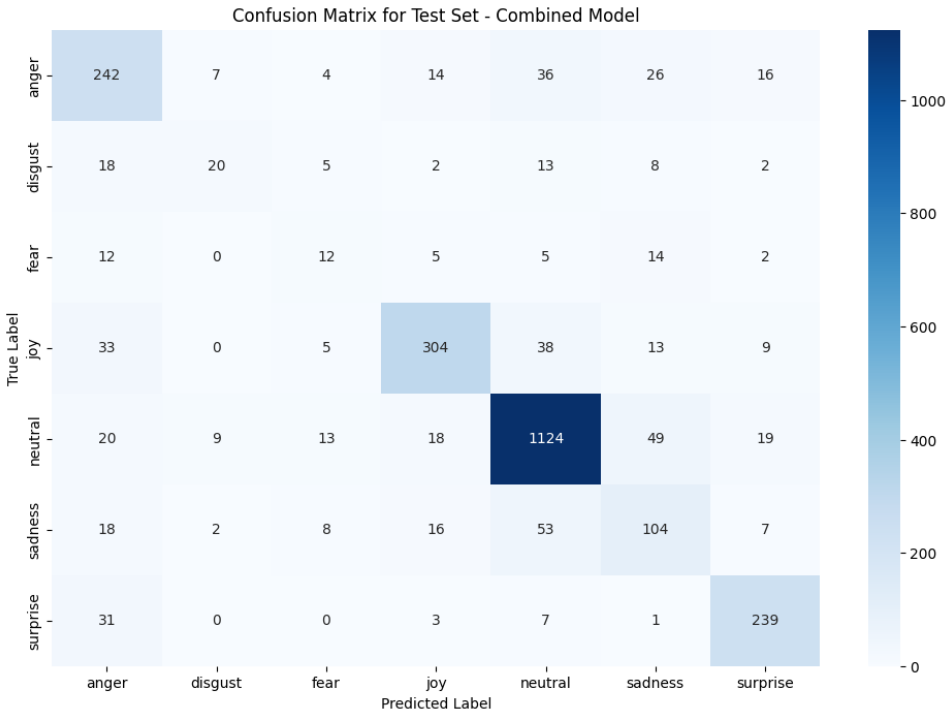
Layer (type)	Output Shape	Param #	Connected to
input_1 (InputLayer)	[(None, 5, 300)]	0	[]
input_2 (InputLayer)	[(None, 5, 1611)]	0	[]
bidirectional (Bidirectional)	(None, 5, 512)	1140736	['input_1[0][0]']
bidirectional_2 (Bidirectional)	(None, 5, 2048)	21594112	['input_2[0][0]']
bidirectional_1 (Bidirectional)	(None, 256)	656384	['bidirectional[0][0]']
bidirectional_3 (Bidirectional)	(None, 1024)	10489856	['bidirectional_2[0][0]']
dense (Dense)	(None, 256)	65792	['bidirectional_1[0][0]']
dense_2 (Dense)	(None, 512)	524800	['bidirectional_3[0][0]']
dropout (Dropout)	(None, 256)	0	['dense[0][0]']
dropout_1 (Dropout)	(None, 512)	0	['dense_2[0][0]']
dense_1 (Dense)	(None, 7)	1799	['dropout[0][0]']
dense_3 (Dense)	(None, 7)	3591	['dropout_1[0][0]']
concatenate (Concatenate)	(None, 14)	0	['dense_1[0][0]', 'dense_3[0][0]']
dense_4 (Dense)	(None, 512)	7680	['concatenate[0][0]']
batch_normalization (BatchNormalization)	(None, 512)	2048	['dense_4[0][0]']
dropout_2 (Dropout)	(None, 512)	0	['batch_normalization[0][0]']
dense_5 (Dense)	(None, 256)	131328	['dropout_2[0][0]']
batch_normalization_1 (BatchNormalization)	(None, 256)	1024	['dense_5[0][0]']
dropout_3 (Dropout)	(None, 256)	0	['batch_normalization_1[0][0]']
dense_6 (Dense)	(None, 128)	32896	['dropout_3[0][0]']
batch_normalization_2 (BatchNormalization)	(None, 128)	512	['dense_6[0][0]']
dropout_4 (Dropout)	(None, 128)	0	['batch_normalization_2[0][0]']
dense_7 (Dense)	(None, 7)	903	['dropout_4[0][0]']

Total params: 34653461 (132.19 MB)
Trainable params: 34651669 (132.19 MB)

Sınıf	Kesinlik	Duyarlılık	F1-Skoru	Destek
Öfke	0.65	0.70	0.67	345
İğrenme	0.53	0.29	0.38	68
Korku	0.26	0.24	0.25	50
Neşe	0.84	0.76	0.80	402
Nötr	0.88	0.90	0.89	1252
Üzüntü	0.48	0.50	0.49	208
Şaşkınlık	0.81	0.85	0.83	281
Doğruluk			0.78	2606
Makro Ort	0.64	0.61	0.62	2606
Ağırlıklı Ort	0.78	0.78	0.783	2606







4.3.2 Dikkat Mekanizmalarıyla Çapraz Füzyon

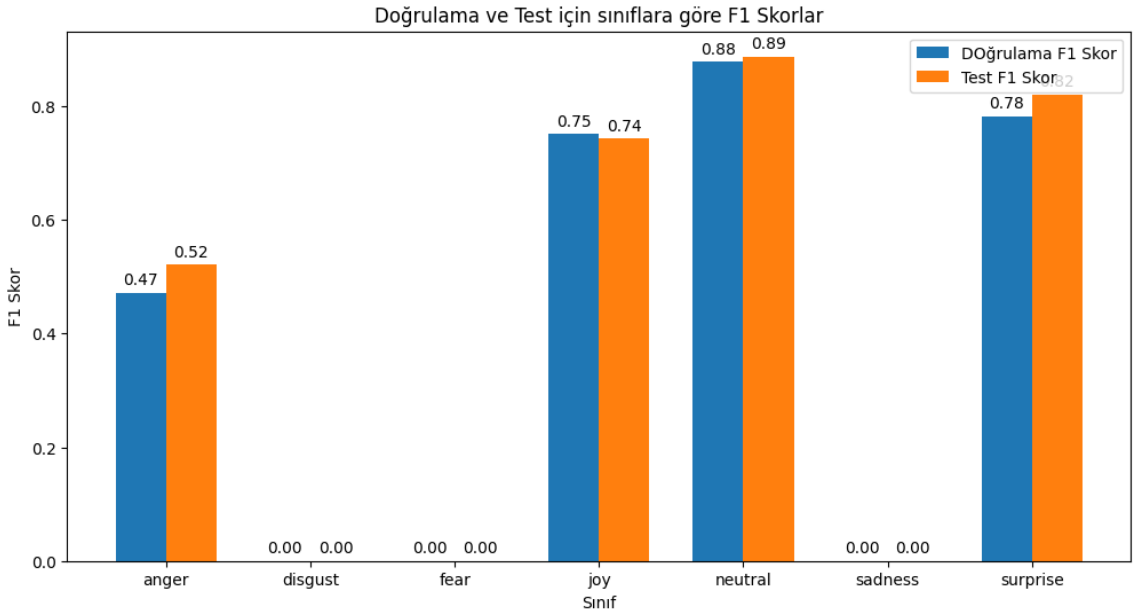
Bu çalışmada kullanılan çapraz füzyon modeli, ses ve metin modalitelerinden elde edilen özellikleri dikkate alarak duygusal durumları tahmin etmek amacıyla tasarlanmıştır. Modelin eğitim, doğrulama ve test aşamalarında verilerin kolayca yüklenip işlenmesini sağlamak için metin ve ses için pencerelenmiş verilerini tensorlere dönüştürmüştür. Model, her bir eğitim

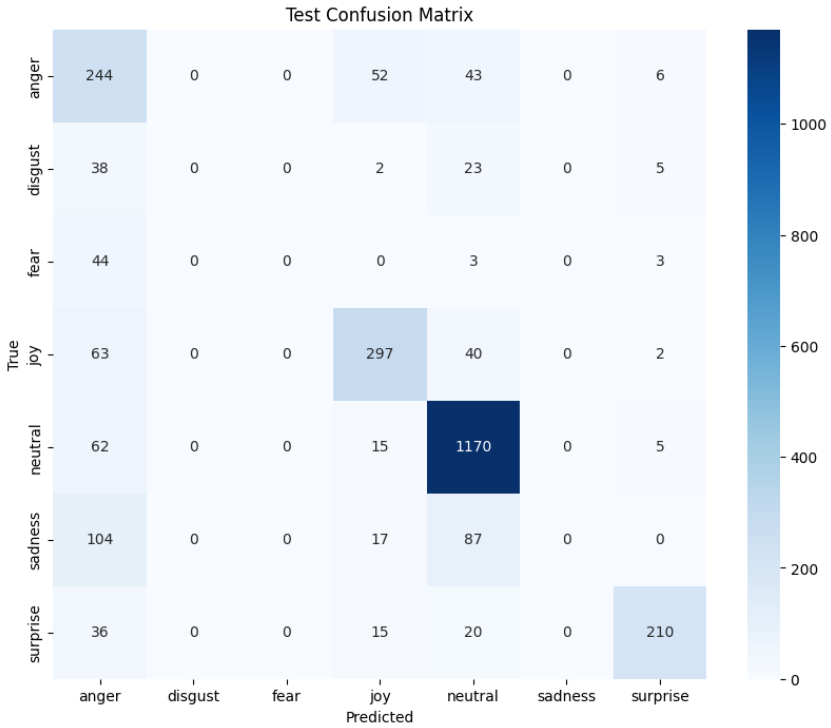
döngüsünde (epoch) metin ve ses özelliklerini projeksiyon katmanlarından geçirir. Çapraz dikkat ve kendi-kendine dikkat mekanizmaları ile etkileşimleri yakalayarak özellikleri birleştirir ve katman normalizasyonu ile çıktıları düzenler. Tam bağlantılı katmanlar, bu birleşik özellikleri işleyerek nihai sınıflandırmayı gerçekleştirir.

1. **Projeksiyon Katmanları (Linear-1 ve Linear-2):**
    - **Linear-1:** Metin özelliklerini 512 boyutlu bir gizli temsil vektörüne dönüştürür. Bu katman, metin modalitesi için 154,112 öğrenilebilir parametreye sahiptir.
    - **Linear-2:** Ses özelliklerini 512 boyutlu bir gizli temsil vektörüne dönüştürür. Bu katman, ses modalitesi için 825,344 öğrenilebilir parametreye sahiptir.
2. **Çok Başlı Dikkat Mekanizmaları (MultiheadAttention-3, 4, 5, 6):**
    - **MultiheadAttention-3 ve 4:** Metin ve ses özellikleri arasındaki çapraz dikkat mekanizmasını gerçekleştirir. Bu katmanlar, metin özelliklerine ses özelliklerine dikkat ederken, ses özellikleri de metin özelliklerine dikkat eder. Bu etkileşimler, modaliteler arası bilgi akışını sağlar.
    - **MultiheadAttention-5 ve 6:** Her bir modalite için kendi-kendine dikkat mekanizmasını gerçekleştirir. Metin ve ses özelliklerinin kendi içindeki önemli bilgileri yakalar.
3. **Katman Normalizasyonu (LayerNorm-7 ve LayerNorm-8):**
    - **LayerNorm-7:** Metin ve ses özelliklerinin birleşik çıktısını normalleştirir.
    - **LayerNorm-8:** Metin ve ses özelliklerinin birleşik çıktısını normalleştirir.
4. **CrossModalAttention-9:**
    - Çapraz dikkat ve kendi-kendine dikkat mekanizmalarını birleştirir. Bu katman, hem metin hem de ses özelliklerinin etkileşimlerini ve kendi içlerindeki bilgileri dikkate alarak nihai özellik vektörlerini oluşturur.
5. **Tam Bağlantılı Katmanlar (Linear-10, ReLU-11 ve Linear-12):**
    - **Linear-10:** Çapraz dikkat katmanından elde edilen birleşik özellikleri işleyerek 512 boyutlu bir vektör oluşturur. Bu katmanda 524,800 öğrenilebilir parametre bulunmaktadır.
    - **ReLU-11:** Aktivasyon fonksiyonu olarak ReLU kullanılır. Bu katman, doğrusal olmayan dönüşümler yaparak modelin öğrenme kapasitesini artırır.
    - **Linear-12:** Nihai sınıflandırma katmanı. Bu katman, modelin tahminlerini üretir ve 7 sınıflı duygu tanıma görevinde 3,591 öğrenilebilir parametreye sahiptir.

```
----------------------------------------------------------------
        Layer (type)               Output Shape         Param #
================================================================
            Linear-1               [-1, 5, 512]         154,112
            Linear-2               [-1, 5, 512]         825,344
MultiheadAttention-3  [[-1, 5, 512], [-1, 2, 2]]               0
MultiheadAttention-4  [[-1, 5, 512], [-1, 2, 2]]               0
MultiheadAttention-5  [[-1, 5, 512], [-1, 2, 2]]               0
MultiheadAttention-6  [[-1, 5, 512], [-1, 2, 2]]               0
         LayerNorm-7               [-1, 5, 512]           1,024
         LayerNorm-8               [-1, 5, 512]           1,024
CrossModalAttention-9     [[-1, 512], [-1, 512]]               0
           Linear-10                  [-1, 512]         524,800
             ReLU-11                  [-1, 512]               0
           Linear-12                    [-1, 7]           3,591
================================================================
Total params: 1,509,895
Trainable params: 1,509,895
Non-trainable params: 0
----------------------------------------------------------------
Input size (MB): 46.09
Forward/backward pass size (MB): 2.23
Params size (MB): 5.76
Estimated Total Size (MB): 54.08
----------------------------------------------------------------
```

Sınıf	Kesinlik	Duyarlılık	F1-Skoru	Destek
Öfke	0.41	0.71	0.52	345
İğrenme	0.0	0.0	0.0	68
Korku	0.00	0.0	0.0	50
Neşe	0.75	0.74	0.74	402
Nötr	0.84	0.94	0.89	1252
Üzüntü	0.0	0.0	0.0	208
Şaşkınlık	0.91	0.75	0.82	281
Doğruluk			0.74	2606
Makro Ort	0.42	0.45	0.42	2606
Ağırlıklı Ort	0.67	0.74	0.70	2606

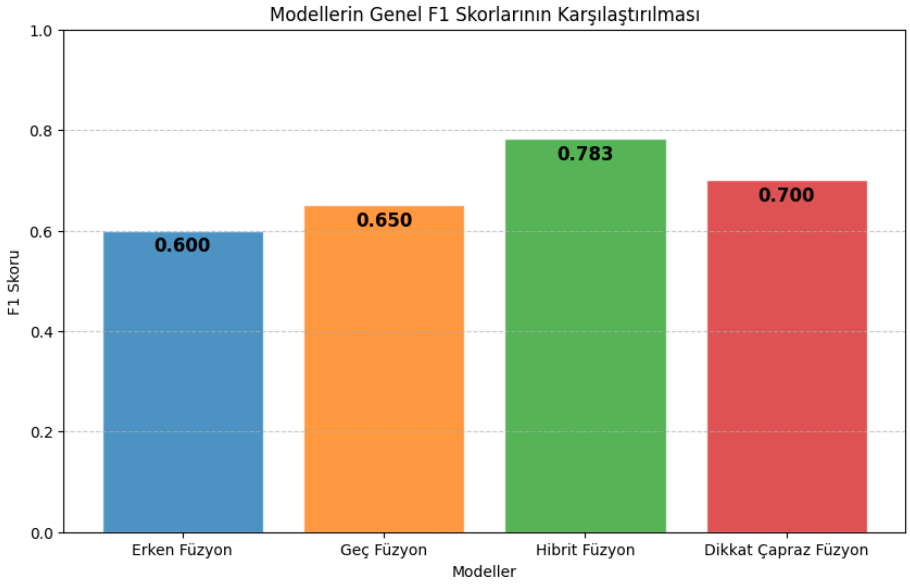


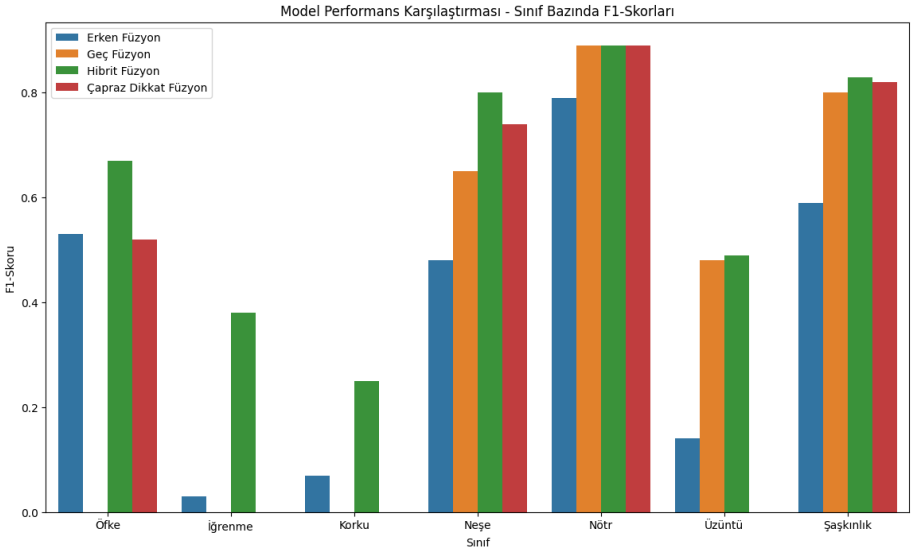


5. Modellerin Karşılaştırılması

Model karşılaştırmasında değerlendirme ölçütü olarak Test için F1 skoru olarak alınmıştır.

I. Erken Füzyon Modeli: 0.60
II. Geç Füzyon Modeli: 0.65
III. Ara Katmanalarda Birleştirerek Hibrit Füzyon Modeli: 0.78
IV. Dikkat Mekanizmaları Kullanıralarak Çapraz Füzyon Modeli: 0.70

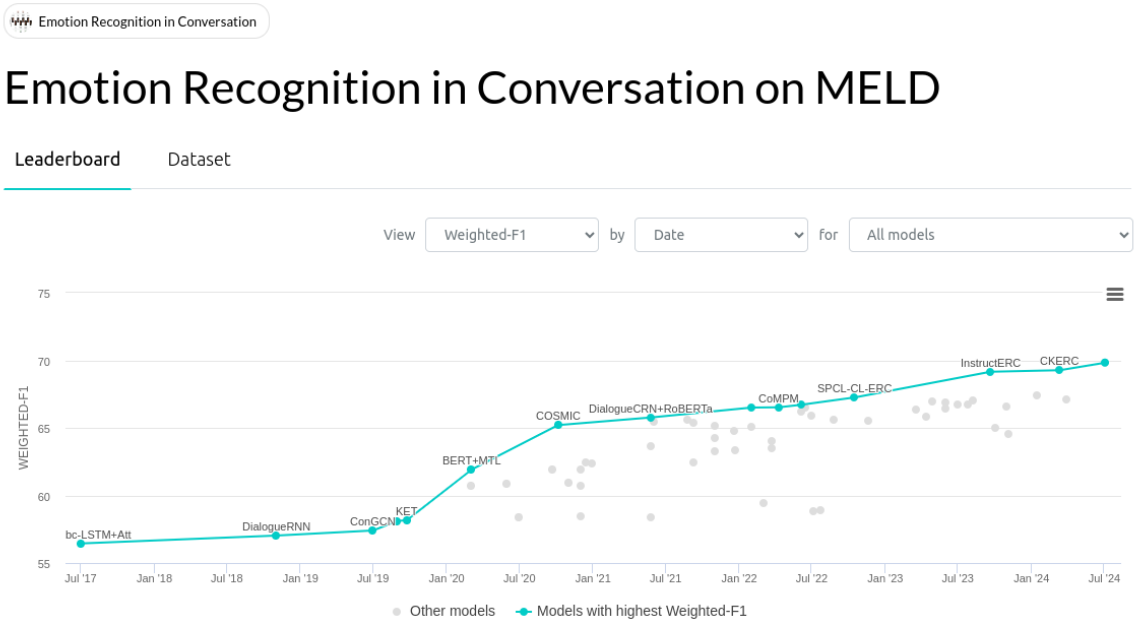




Modeller arasında en yüksek performansı Hibrit Füzyon Modeli göstermiş ve F1 skoru 0.78 olarak elde edilmiştir. Geç Füzyon Modeli, modaliteler arasındaki etkileşimleri tam olarak yakalayamadığı için performansı sınırlı kalmış ve F1 skoru 0.65 olarak bulunmuştur. Dikkat Mekanizmaları Kullanılarak Çapraz Füzyon Modeli ise 0.70 F1 skoru ile başarılı bir performans sergilemiştir. Erken Füzyon Modeli ise en düşük performansı göstermiştir. Pencereleme ve kaydırma tekniğinin başarıda önemli bir rolü vardır. Pencereleme yapıldıktan sonra bütün modeller için ortalama %10luk performans artışları olmuştur.

7. Diğer Çalışmalarla Karşılaştırma

MELD veri seti kullanılarak yapılan diğer duygu analizi performansları Papers With Code'da paylaşılmıştır. Aşağıdaki grafikte farklı çalışmaların F1 skorları gösterilmiştir. [20]



Çalışmamızda kullanılan Hibrit Füzyon ve Çapraz Dikkat Füzyon modelleri, Papers With Code'da yer alan diğer çalışmalara kıyasla daha yüksek F1 skorları elde etmiştir. Hibrit Füzyon modelimiz, genel F1 skoru 0.783 ile en yüksek performansı göstermiş, Çapraz Dikkat Füzyon modeli ise 0.70 F1 skoru ile rekabetçi bir performans sergilemiştir. Erken Füzyon modelimiz, 0.60 F1 skoru ile diğer çalışmalara kıyasla daha düşük performans göstermiştir. Diğer çalışmalarda kullanılan modellere kıyasla, erken füzyonun hızlı birleştirme avantajına rağmen her modalitenin özgün bilgilerini tam olarak yansıtamadığı gözlemlenmiştir.

8. Kaynaklar

[1] [3] Arianna DUlizia. Exploring multimodal input fusion strategies. In Multimodal Hu- man Computer Interaction and Pervasive Services, pages 34–57. IGI Global, 2009

[2] Baltrušaitis, T., Ahuja, C. & Morency, L.-P. (2017). Multimodal Machine Learning: A Survey and Taxonomy. *arXiv preprint arXiv:1705.09406*, .

[3] Blikstein, P. (2013). Multimodal learning analytics. Proceedings of the Third International Conference on Learning Analytics and Knowledge.

[4] Kaur, R., & Kautish, S. (2019). Multimodal Sentiment Analysis: A Survey and Comparison. *Int. J. Serv. Sci. Manag. Eng. Technol., 10*, 38-58.

[5] Lahat, D., Adalı, T., & Jutten, C. (2015). Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects. Proceedings of the IEEE, 103, 1449-1477.

[6] Luo, J., Phan, H., & Reiss, J. (2023, June). Cross-modal fusion techniques for utterance-level emotion recognition from text and speech. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* IEEE.

[7] Meng, Q., Qian, H., Liu, Y., Xu, Y., Shen, Z., & Cui, L. (2023). Unsupervised representation learning for time series: A review. *arXiv preprint arXiv:2308.01578*.

[8] Morency, L., Mihalcea, R., & Doshi, P. (2011). Towards multimodal sentiment analysis: harvesting opinions from the web. *International Conference on Multimodal Interaction*.

[9] Pan Q, Meng Z. Hybrid Uncertainty Calibration for Multimodal Sentiment Analysis. *Electronics*. 2024; 13(3):662.

[10] Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., & Mihalcea, R. (2018). MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. *ArXiv, abs/1810.02508*.).

[11] [22]Sahu, G. (2020). Adaptive Fusion Techniques for Effective Multimodal Deep Learning.

[12] Shen, W. F., Tang, H. W., Li, J. B., Li, X., & Chen, S. (2023). Multimodal data fusion for supervised learning-based identification of USP7 inhibitors: a systematic comparison. *Journal of cheminformatics*, *15*(1), 5. https://doi.org/10.1186/s13321-022-00675-8

[13] Y. Sun, D. Cheng, Y. Chen and Z. He, "DynamicMBFN: Dynamic Multimodal Bottleneck Fusion Network for Multimodal Emotion Recognition," *2023 3rd International*

[14] Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L. P. (2017). Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

[15] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2017.Emotional chatting machine: Emotional conversation generation with internal and external memory.*arXiv preprint arXiv:1704.01074*.

[16] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018.I know the feeling: Learning to converse with empathy.*arXiv preprint arXiv:1811.00207*

[17] Sheng-Yeh Chen, Chao-Chun Hsu, Chuan-Chun Kuo, Lun-Wei Ku, et al. 2018.Emotionlines: An emotion corpus of multi-party conversations.*arXiv preprint arXiv:1802.08379*.

[18] V7 Labs. (n.d.). *Multimodal Deep Learning Guide*. Retrieved July 19, 2024, from https://www.v7labs.com/blog/multimodal-deep-learning-guide#h6

[19] Karakaş, S. (2017). Prof. Dr. Sirel Karakaş Psikoloji Sözlüğü: Bilgisayar Programı ve Veritabanı - www.psikolojisozlugu.com (sürüm: 5.2.0/2022)

[98] A. Karpathy and L. Fei-Fei, “Deep visual-semantic alignments for generating image descriptions,” in CVPR, 2015.

[99] A. Karpathy, A. Joulin, and L. Fei-Fei, “Deep fragment embed- dings for bidirectional image sentence mapping,” in NIPS, 2014

[20] Papers with Code. (2024). Emotion Recognition in Conversation on MELD. Retrieved from https://paperswithcode.com/sota/emotion-recognition-in-conversation-on-meld

[131] J. Malmaud, J. Huang, V. Rathod, N. Johnston, A. Rabinovich, and K. Murphy, “What’s cookin’? interpreting cooking videos using text, speech and vision,” NAACL, 2015

252] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Tor-ralba, and S. Fidler, “Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books,” in ICCV, 2015.