T.C
BİTLİS EREN ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

FARKLI SOSYAL MEDYA PLATFORMLARINDA OTOMATİK HAKARET TESPİTİ

YÜKSEK LİSANS TEZİ

SEZER DÖYMAZ

DANIŞMAN

DOÇ. DR. VEDAT TÜMEN

İKİNCİ DANIŞMAN

DR. ÖĞR. ÜYESİ MEHMET EMİN BAKIR

AĞUSTOS 2025
BİTLİS

T.C
BİTLİS EREN ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

FARKLI SOSYAL MEDYA PLATFORMLARINDA OTOMATİK HAKARET TESPİTİ

YÜKSEK LİSANS TEZİ

SEZER DÖYMAZ
ORCID: 0009-0002-2799-2241

DANIŞMAN
DOÇ. DR. VEDAT TÜMEN

İKİNCİ DANIŞMAN
DR. ÖĞR. ÜYESİ MEHMET EMİN BAKIR

AĞUSTOS 2025
BİTLİS

T.C.

BİTLİS EREN ÜNİVERSİTESİ

LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

YÜKSEK LİSANS TEZ ÇALIŞMASI ETİK BEYANI

Lisansüstü Eğitim Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı Yüksek Lisans öğrencisiyim. Hazırlamış olduğum "**Farklı Sosyal Medya Platformlarında Otomatik Hakaret Tespiti**" başlıklı tez çalışmasında sunduğum veri, bilgi, analiz ve belgeleri akademik etik kurallar çerçevesinde elde ettiğimi; Tüm değerlendirme, analiz ve sonuçları bilimsel etik ve ahlak ilkelerine uygun şekilde sunduğumu; Tez çalışmasında yararlandığım tüm kaynaklara eksiksiz biçimde atıfta bulunduğumu ve kaynak gösterdiğimi; Kullanılan verilere hiçbir şekilde müdahale etmediğimi ve üzerinde herhangi bir değişiklik yapmadığımı; Bu tezde sunulan çalışmanın özgün olduğunu ve tamamen tarafımdan gerçekleştirildiğini; İleride aksi bir durumun tespit edilmesi halinde doğabilecek tüm hak kayıplarını kabul ettiğimi beyan ederim. 14/08/2025

SEZER DÖYMAZ

T.C.

BİTLİS EREN ÜNİVERSİTESİ

LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

TEZ YAZIM KILAVUZU UYGUNLUK BEYANI

"**Farklı Sosyal Medya Platformlarında Otomatik Hakaret Tespiti**" başlıklı bu yüksek lisans tezi, Bitlis Eren Üniversitesi Lisansüstü Eğitim Enstitüsü Tez Yazım Kılavuzuna uygun olarak hazırlanmıştır. **14/08/2025**

Tezi Hazırlayan	**Danışman**
İmza	İmza
Sezer DÖYMAZ	Doç. Dr. Vedat TÜMEN

Bilgisayar Mühendisliği Anabilim Dalı Başkanı

İmza

DOÇ. Dr. Musa ÇIBUK

T.C.

BİTLİS EREN ÜNİVERSİTESİ

LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

YÜKSEK LİSANS TEZ ÇALIŞMASI

Bitlis Eren Üniversitesi Lisansüstü Eğitim Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı öğrencisi Sezer DÖYMAZ tarafından hazırlanan "**Farklı Sosyal Medya Platformlarında Otomatik Hakaret Tespiti**" başlıklı yüksek lisans tezi ile ilgili tez savunma sınavı, 14/08/2025 tarihinde yapılmış ve tezin oybirliği ile kabul edilmesine karar verilmiştir.

JÜRİ:	**İMZA**
Danışman: Doç. Dr. Vedat TÜMEN (Bitlis Eren Üniversitesi)	……
Üye: Doç. Dr. Kubilay DEMİR (Bursa Teknik Üniversitesi)	…..……
Üye: Dr. Öğr. Üyesi İrfan ÖKTEN (Bitlis Eren Üniversitesi)	……..……

Bitlis Eren Üniversitesi Lisansüstü Eğitim Enstitüsü Yönetim Kurulu'nun…… tarih ve……sayılı kararıyla jüri tarafından kabul edilmiş bu çalışmanın yüksek lisans tezi olarak kabulü onaylanmıştır.

…. /… /2025

Prof. Dr. Mehmet Bakır ŞENGÜL

Enstitü Müdür V.

T.C

Bitlis Eren Üniversitesi Lisansüstü Eğitim Enstitüsü

Bilgisayar Mühendisliği Anabilim Dalı

FARKLI SOSYAL MEDYA PLATFORMLARINDA OTOMATİK HAKARET TESPİTİ

Yüksek Lisans Tezi

Sezer DÖYMAZ

Danışman: Doç. Dr. Vedat TÜMEN

II. Danışman: Dr. Öğr. Üyesi Mehmet Emin BAKIR

Ağustos 2025

ÖZET

Sosyal medya platformları, düşünce paylaşımının merkezi hâline gelirken, hakaret ve küfür içerikli söylemlerin yayılmasıyla birlikte otomatik hakaret tespiti giderek daha önemli bir araştırma alanı hâline gelmiştir. Bu bağlamda Facebook, Instagram, X ve Reddit'ten toplanan yorumlar üzerinden, derin öğrenme modelleriyle çok sınıflı ve ikili sınıflandırma çalışmalar yürütülmüştür. CNN, LSTM ve BERTurk modelleri kullanılmış; ayrıca GPT-4o tabanlı büyük dil modeli ile sıfır, bir ve üç örnekli etiketleme senaryoları gerçekleştirilmiştir. Birleşik veri ikili sınıflandırmasında BERTurk %90, LSTM %87, CNN %87 F1 skoruna ulaşmıştır. Birleşik veri çok sınıflı sınıflandırmada ise BERTurk %87 ile en yüksek başarıyı sağlamıştır. GPT-4o, en iyi sonucu ortalama %69 F1 skoru ile tek örnekli çalışmada göstermiştir. Çalışma, Türkçe sosyal medya verilerinde hakaret tespiti alanında dört farklı platformu bir arada ele alan ve derin öğrenme ile büyük dil modeli temelli analizleri yapan ilk kapsamlı araştırmalardan biri olarak literatüre özgün katkılar sağlamaktadır. Ayrıca çalışma, dört farklı platformdan derlenen ilk kapsamlı Türkçe hakaret veri setini sunmaktadır. Elde edilen bulgular, tüm platformların bir arada olduğu bir modelin çoğu durumda platforma özgü modellere benzer hatta daha yüksek başarı göstererek, platformlar arası genellemenin mümkün olduğunu göstermektedir.

Anahtar Kelimeler: Sosyal medya, Doğal dil işleme, Hakaret tespiti, Makine öğrenmesi

Republic Of Türkiye

Bitlis Eren University Graduate School

Department Of Computer Engineering

AUTOMATIC INSULT DETECTION ON DIFFERENT SOCIAL MEDIA PLATFORMS

Master's Thesis

Sezer DÖYMAZ

Advisor: Assoc. Prof. Dr. Vedat TÜMEN

Co-Advisor: Asst. Prof. Dr. Mehmet Emin BAKIR

August 2025

ABSTRACT

As social media platforms have become centers for sharing thoughts, the spread of abusive and profane expressions has made automatic insult detection an increasingly important area of research. In this context, multi-class and binary classification studies were conducted using deep learning models on comments collected from Facebook, Instagram, X, and Reddit. CNN, LSTM, and BERTurk models were employed, and zero-shot, one-shot, and three-shot labelling scenarios were implemented using the GPT-4o-based large language model. In binary classification with the combined dataset, BERTurk achieved an F1 score of 90%, while LSTM and CNN both reached 87%. In multi-class classification with the combined dataset, BERTurk again yielded the highest performance with an F1 score of 87%. GPT-4o produced its best result in the one-shot scenario with an average F1 score of 69%. This study provides an original contribution to the literature as one of the first comprehensive analyses examining four different platforms together and applying both deep learning and large language model-based approaches for insult detection in Turkish social media data. Furthermore, it presents the first extensive Turkish insult dataset compiled from four platforms. The findings show that a model trained on all platforms together often achieves similar or higher performance compared to platform-specific models, indicating that cross-platform generalisation is possible.

Keywords: Social media, Natural language processing, Insult detection, Machine learning

TEŞEKKÜR

Yüksek lisans eğitimim süresince bilgi birikimleri, rehberlikleri ve her daim hissettirdikleri destekleriyle bana yol gösteren değerli danışman hocalarım Doç. Dr. Vedat TÜMEN ve Dr. Öğr. Üyesi Mehmet Emin BAKIR'a en içten teşekkürlerimi sunarım.

Akademik yolculuğum boyunca bilgi ve tecrübeleriyle katkı sağlayan, desteklerini esirgemeyen Bitlis Eren Üniversitesi Bilgisayar Mühendisliği Bölümü'nün tüm akademik ve idari personeline teşekkür ederim.

Bugüne dek attığım her adımda yanımda olan, sabırları, sevgileri ve sonsuz inançlarıyla bana güç veren sevgili aileme, özellikle de babam ve anneme; zorlu anlarda desteğini esirgemeyen, moral ve motivasyon kaynağım olan arkadaşlarıma ve değerli dostlarıma gönülden teşekkür ederim.

Bu süreçte emeği geçen herkese sonsuz minnettarım.

2025

Sezer DÖYMAZ

İÇİNDEKİLER

SİMGELER

f	:Girdi matrisi
h	:Filtre, evrişim işlemini uygulamak için kullanılan ağırlık matrisi
m	:Çıkış matrisinin satır indeksi
n	:Çıkış matrisinin sütun indeksi
J	:Filtrenin matrisinin satır indeksi
k	:Filtrenin matrisinin sütun indeksi
$\sum$	:Toplama işlemi
$\mathrm{y}_r(x)$	:Girdi x için r sınıfa ait tahmin edilen olasılık
$a_r(x)$	:X girdisi için r. sınıfa karşılık gelen aktivasyon değeri
exp	:Üssel (exponential) fonksiyon
y_j	:j sınıfa ait tahmin edilen olasılık

KISALTMALAR

Tp	:Gerçek Pozitif (True Positive)
Tn	:Gerçek Negatif (True Negative)
Fp	:Yanlış Pozitif (False Positive)
Fn	:Yanlış Negatif (False Negative)
Cnn	:Evrişimsel Sinir Ağları (Convolutional Neural Networks)
Lstm	:Uzun Kısa Süreli Bellek (Long Short-Term Memory)
Bert	:Transformatörlerden Çift Yönlü Kodlayıcı Gösterimleri (Bidirectional Encoder Representations From Transformers)
Dna	:Deoksiribo Nükleik Asit
Rnn	:Tekrarlayan Sinir Ağı
Cls	:Sınıflandırma (Classification)
Sep	:Ayırıcı / Ayraç (Separator)
Nlp	:Doğal Dil İşleme (Natural Language Processing)
Llm	:Büyük Dil Modeli ((Large Language Models)

ÇİZELGELER DİZİNİ

ŞEKİLLER DİZİNİ

1. GİRİŞ

Dijitalleşen dünyada bireyler, gündelik yaşamlarının önemli bir bölümünü sosyal medya platformları üzerinde geçirmektedir [1]. Facebook, X (Twitter), Instagram ve Reddit gibi mecralar, kullanıcıların düşüncelerini serbestçe ifade edebildiği alanlara dönüşmüş; bu ifade özgürlüğü ise zaman zaman hakaret içerikli söylemlerin de yaygınlaşmasına neden olmuştur [2]. Bu tür söylemlerin bireyler üzerindeki psikolojik etkileri kadar, toplumsal düzeyde saygı ve güven duygusunu zedelemesi bakımından da dikkate alınması gereken bir problem olduğu açıktır.

Hakaret, genel anlamda bir bireyin şeref ve onuruna saldırı niteliği taşıyan söz veya eylemler olarak tanımlanmaktadır [3]. Küfür ise bu kapsamda değerlendirilen, çoğunlukla kaba ve aşağılayıcı kelimelerin kullanıldığı özel bir hakaret biçimidir. Gerek bireylerin dijital ortamdaki güvenlik duygusunu zedelemesi, gerekse ifade özgürlüğünün sınırlarını aşması nedeniyle, bu tür içeriklerin tespiti hem akademik hem de hukuki açıdan giderek önem kazanmaktadır.

Hakaret içerikli mesajların otomatik olarak tespit edilmesi üzerine yapılan çalışmalar son yıllarda artış göstermiştir. Özellikle Türkçe içeriklere yönelik yürütülen araştırmalarda, dilin yapısal özellikleri, kelime kökenleri ve deyimsel kullanımları nedeniyle bazı zorluklar yaşandığı vurgulanmaktadır [4]. Bu alandaki ilk çalışmalardan biri olan Çöltekin [4], X platformu üzerindeki Türkçe yorumlar içinde yer alan hakaret içeriklerini tespit edebilmek amacıyla kapsamlı bir veri kümesi oluşturmuştur. Benzer şekilde, Soykan ve ark. Türkçe küfür tespiti için çeşitli makine öğrenmesi ve derin öğrenme yöntemlerini karşılaştırmış ve Electra tabanlı modellerin üstün performans gösterdiğini rapor etmiştir [5].

Literatürdeki bu çalışmaların büyük kısmı ya belirli bir platforma odaklanmış ya da sınırlı sayıda veriyle gerçekleştirilmiştir. Karayiğit ve ark. [6] tarafından yürütülen araştırmada yalnızca Instagram verileri kullanılmış ve yaklaşık 30.000 yorum üzerinde değerlendirme yapılmıştır. Ancak bu çalışmaların çoğunda veri çeşitliliğinin yetersiz oluşu ve platformlar arası karşılaştırmalı analizlerin bulunmaması, kapsamlı sonuçlara ulaşmayı güçleştirmiştir [7].

Dört farklı sosyal medya platformundan (Facebook, Instagram, X ve Reddit) Türkçe veri toplanarak hem platform bazlı hem de tümleştirilmiş bir yapı üzerinden

değerlendirme yapılmıştır. Bu bağlamda üç farklı derin öğrenme tabanlı yöntem (BERTurk, CNN, LSTM) kullanılarak, hem her bir platformun kendine özgü dil yapısının sınıflandırma üzerindeki etkisi, hem de verilerin birleştirilmesinin model başarısına etkisi analiz edilmiştir.

Bu kapsamda çalışmada iki farklı sınıflandırma yaklaşımı benimsenmiştir: İlki, Çöltekin [4] tarafından önerilen yönergelere dayanan ve "hakaret içermeyen", "bireye yönelik", "gruba yönelik", "kuruma/olaya yönelik" ve "şaka/nesneye yönelik" gibi beş kategori içeren çok sınıflı etiketleme sistemidir. Ancak orijinal etiketleme sisteminde yer alan "X" (anlaşılmaz içerik) kategorisi çalışmadan çıkarılmıştır; böylece model performansının artırılması hedeflenmiştir. İkinci yaklaşım ise, aynı veri setinin yalnızca "hakaret içeriyor" ve "hakaret içermiyor" biçiminde yeniden etiketlenerek ikili sınıflandırma yapılmasıdır. Her iki yaklaşım için de modeller eğitilmiş ve karşılaştırmalı değerlendirmeler gerçekleştirilmiştir. Ayrıca tezde, her platform için ayrı modeller eğitmek yerine, tüm platformların birleşiminden oluşturulan tek bir modelin başarısı da test edilmiştir. Bu sayede, platformlara özel eğitilen modellerle karşılaştırıldığında birleşik modelin genellenebilirliği ve verimliliği incelenmiş; tek bir modelle çoklu platform performansı elde edilip edilemeyeceği değerlendirilmiştir.

Bunlara ek olarak, bu tezde yalnızca insan eliyle etiketlenmiş verilerle değil, aynı zamanda büyük dil modeli (LLM) tabanlı otomatik etiketleme yöntemleriyle de bir analiz gerçekleştirilmiştir. OpenAI'nin GPT-4o modeli aracılığıyla yürütülen bu süreçte, veri seti üzerinde sıfır örnekli (zero-shot) ve az örnekli (few-shot) olmak üzere üç farklı etiketleme deneyi uygulanmıştır. İlk aşamada, modele yalnızca beş kategoriye ait kısa tanımlar sunularak tüm veri kümesinin otomatik biçimde etiketlenmesi sağlanmıştır. İkinci aşamada, her kategoriye ait birer örnek metin modele verilerek etiketleme işlemi tekrarlanmıştır. Üçüncü ve son aşamada ise, her kategori için üçer örnek kullanılarak modelin karar alma süreçlerine daha fazla bağlamsal bilgi sunulmuştur. Her üç senaryo da OpenAI API üzerinden gerçekleştirilmiştir. Son olarak, GPT-4o modeli tarafından oluşturulan etiketler, insan uzmanlarca belirlenen gerçek etiketlerle karşılaştırılmış ve elde edilen hata matrisleri doğrultusunda model başarımı analiz edilmiştir. Bu yönüyle çalışma, büyük dil modellerinin Türkçe sosyal medya içeriklerinde çok sınıflı hakaret tespiti görevindeki etkinliğini değerlendiren ilk örneği olma niteliği taşımaktadır.

Bu tez çalışması, yalnızca derin öğrenme temelli hakaret tespiti yöntemlerinin karşılaştırılması açısından değil; aynı zamanda veri kaynağı, dil ve kapsam bakımından

da literatüre önemli katkılar sunmaktadır. Literatür taramaları sonucunda, Facebook, Instagram, X ve Reddit olmak üzere dört büyük sosyal medya platformundan Türkçe veri toplanarak yürütülmüş çok yönlü bir çalışmaya daha önce rastlanmamıştır. Bu yönüyle çalışma, dört platformu aynı çatı altında ele alan ilk Türkçe hakaret tespiti araştırması olma özelliği taşımaktadır. Ayrıca, platform bazlı olarak değerlendirildiğinde, Facebook ve Reddit üzerindeki Türkçe yorumlar için otomatik hakaret sınıflandırmasına yönelik daha önce herhangi bir bilimsel çalışmaya rastlanmamıştır. Bu çalışma yalnızca farklı model yapılarını karşılaştırmakla kalmamakta, aynı zamanda her bir platforma özgü etiketlenmiş Türkçe veri setini de literatüre kazandırmaktadır. Hem çok sınıflı hem ikili etiketleme senaryolarında derin öğrenme mimarileriyle elde edilen karşılaştırmalı sonuçlar, Türkçe sosyal medya verileriyle çalışan modellerin başarımını farklı platform düzeylerinde değerlendirme imkânı sunarak alana özgün ve bütüncül bir katkı sunması hedeflenmiştir.

1.1. Tez Organizasyonu

Bu tez beş ana bölümden oluşmaktadır. Birinci bölümde, çalışmanın amacı, kapsamı, problemi ve yöntemi açıklanmış; çalışmanın önemi vurgulanmıştır. İkinci bölümde, hakaret kavramı kuramsal ve hukuki boyutlarıyla ele alınmış; toplumsal etkileri değerlendirilmiş ve ilgili literatüre yer verilmiştir. Üçüncü bölümde, araştırmada kullanılan veri seti, etiketleme yapısı, ön işleme adımları ve makine öğrenmesi ile derin öğrenme temelli modeller tanıtılmış; ayrıca GPT-4o gibi büyük dil modeli (LLM) tabanlı otomatik sınıflandırma yaklaşımlarına da yer verilmiştir. Dördüncü bölümde, deneysel süreçler, modelleme sonuçları ve değerlendirme metrikleri detaylandırılmış; beşinci ve son bölümde ise elde edilen bulgulara dayalı genel sonuçlar sunulmuş ve gelecekteki çalışmalara yönelik önerilerde bulunulmuştur.

2. GENEL BİLGİLER

Bu bölümde, çalışmanın kavramsal, hukuki ve kuramsal temelleri ile birlikte ilgili literatürde yer alan mevcut araştırmalar sunulmaktadır. İlk olarak hakaretin tanımı, sınırları ve birey-toplum üzerindeki etkileri ele alınmakta; ardından Türkiye'deki hukuki düzenlemeler çerçevesinde hakaret suçunun cezai boyutu açıklanmaktadır. Bölümün devamında, sosyal medya platformlarında hakaret ve küfür içeriklerinin otomatik tespitine yönelik yapılmış ulusal ve uluslararası çalışmalar incelenerek kapsamlı bir literatür taraması sunulmakta ve bu tez çalışmasının akademik alana sağladığı katkılar değerlendirilmektedir.

2.1. Hakaretin Tanımı ve Sınırları

Hakaret, hukuk ve iletişim literatüründe bireyin kişilik haklarına doğrudan bir saldırı biçimi olarak tanımlanır. Kavramsal olarak "hakaret", bir kişiye onur kırıcı, küçük düşürücü veya alaycı biçimde yöneltilmiş söz, davranış ya da ima yoluyla yapılan bir eylemdir. Türk Ceza Kanunu'nun 125. maddesine göre, "bir kimseye onur, şeref ve saygınlığını rencide edebilecek nitelikte somut bir fiil veya olgu isnat eden ya da söven kişi" hakaret suçu işlemiş sayılır [8]. Dolayısıyla hakaret yalnızca açık biçimde edilen küfürle sınırlı kalmamakta; bağlam içinde onur kırıcı yorumlar, aşağılayıcı ifadeler ya da haysiyet zedeleyici ithamlar da bu kapsamda değerlendirilmektedir.

Geleneksel medya ortamlarında hakaret eylemleri daha sınırlı ve izlenebilir iken, günümüzün dijital medya ortamları özellikle sosyal medya platformları bu sınırları belirsizleştirmiştir. Sosyal medya üzerinden yapılan paylaşımlar hem hızla yayılmakta hem de geniş kitlelere ulaşabilmektedir. Bu durum, bireylerin çevrimiçi ortamda daha kolay şekilde hakarete maruz kalmalarına zemin hazırlamaktadır [9]. Aynı zamanda sosyal medya dilinin sıradanlaşan küfürlü yapısı, hangi ifadelerin hakaret sayılıp sayılmayacağı konusunda ciddi tartışmalar doğurmuştur.

Küfür, hakaretin yoğunlaştırılmış ve doğrudan biçimi olarak değerlendirilebilir. Küfürlü ifadeler, bireyin cinsiyetine, etnik kökenine, ailesine ya da kutsal değerlerine yöneldiğinde yalnızca sosyal bir sorun değil, aynı zamanda yasal bir suç teşkil eder. Sosyal medya platformları, anonimlik ve erişim kolaylığı sayesinde bu tür söylemlerin sıklıkla tekrarlandığı birer zemin haline gelmiştir. Örneğin, sosyal medya kullanıcıları çoğu zaman farkında olmadan bile hakaret sınırına yakın ifadeler kullanabilmektedir. Bu

da, dilin gündelik kullanımı ile hukuki çerçeve arasındaki mesafeyi belirginleştirmektedir [10] .

Zeybek'in yaptığı çalışmada, Türkçe sosyal medya içeriklerinde yer alan hakaret söylemleri hem geleneksel makine öğrenmesi yöntemleri hem de derin öğrenme modelleriyle analiz edilmiştir. Çalışmada, bağlamı doğru analiz edemeyen sınıflandırıcıların özellikle alay, ironi ya da dolaylı hakaret içeren cümlelerde başarısız olduğu gösterilmiştir [11]. Bu durum, küfür ya da hakaretin yalnızca kelime düzeyinde değil, anlamsal düzeyde de modellenmesi gerektiğini ortaya koymaktadır.

Benzer şekilde, Özar ve çalışma arkadaşları tarafından geliştirilen Türkçe küfür sözlüğü, sınıflandırıcı sistemlerin temelini oluşturmuştur. Ancak araştırmada, sabit sözlüklerin değişen sosyal medya diline uyum sağlayamadığı, bu nedenle sözlük temelli sistemlerin dinamik modellerle desteklenmesi gerektiği belirtilmiştir [12]

2.2. Toplumsal ve Bireysel Etkiler

Hakaret, yalnızca bireyler arasında gerçekleşen bir etik ihlal değil, aynı zamanda toplumun genel ruh hâlini ve dijital kültürünü etkileyen bir sorundur. Hakarete uğrayan bireylerde özgüven kaybı, dijital yalnızlık, psikolojik baskı ve sosyal geri çekilme gibi ciddi etkiler gözlemlenmektedir. Özellikle genç kullanıcılar arasında, sosyal medya üzerindeki hakaret ve küfür temelli etkileşimler, bireyin kendini ifade etme cesaretini kırmakta ve dijital ortama olan güvenini azaltmaktadır [13].

Çelen'in çalışması, dijital ortamda değişen ifade biçimlerinin geleneksel hakaret kavramını nasıl dönüştürdüğünü incelemektedir. Araştırmaya göre, günümüzde hakaret yalnızca bireye yönelik sözel saldırılarla sınırlı kalmamakta; görseller, emojiler, dolaylı göndermeler ve etiketleme gibi dijital pratikler de hakaret kapsamına alınabilmektedir [14]. Bu bulgular, hakaretin birey psikolojisi üzerindeki etkilerinin yanı sıra, iletişim teknolojileriyle dönüşen yeni biçimlerini anlamak açısından da önemlidir.

2.3. Türkiye'de Hakaretin Hukuki ve Cezai Boyutu

Türkiye'de hakaret suçu, 5237 sayılı Türk Ceza Kanunu'nun 125–131. maddeleri arasında düzenlenmiştir. Kanuna göre, bir kişiye yönelik alenen gerçekleştirilen hakaret eylemi, kamu davasına konu edilebilecek nitelikte olup, faile üç aydan iki yıla kadar hapis veya adli para cezası verilebilir. Suçun kamu görevlisine karşı, dini değerleri aşağılayacak şekilde ya da cinsiyet temelli şekilde işlenmesi hâlinde ise ceza artışı öngörülmüştür [8].

Ayrıca, sosyal medyada yapılan paylaşımlar genellikle “aleniyet” unsurunu taşıdığı için, bu tür hakaret vakalarında ceza süreci daha sıkı yürütülmektedir. Doğan’ın çalışmasına göre, sosyal medyada yapılan hakaretlerin, failin kimliğinin tespiti durumunda ciddi ceza süreçlerine yol açabileceği, anonimlik perdesinin delil olarak kullanılmasının mümkün olduğu belirtilmektedir [15].

Türk hukuk sisteminde hakaret suçunun cezalandırılması, mağdurun şikâyetine bağlı olarak yürütülmektedir. Ancak suçun alenen işlenmesi, kamu görevlilerine yönelik olması ya da kamu düzenini bozucu nitelik taşıması hâlinde, re’sen soruşturma açılması da mümkündür. Özellikle sosyal medya ortamlarında gerçekleşen hakaret vakalarında, cezai sorumluluk süreci teknik delil toplama, IP adresi tespiti ve içerik kayıtları gibi unsurlar üzerinden yürütülmektedir [16].

Bu süreçlerde çoğu zaman avukatlar veya bilişim uzmanlarının katkısıyla yürütülen teknik analizler, hakaretin hukuki delil niteliğini taşır hâle getirilmesini sağlar. Bu nedenle, sosyal medya paylaşımlarında kullanılan dilin yalnızca sosyal değil, aynı zamanda hukuki sonuçlar doğurabileceği göz önünde bulundurulmalıdır.

2.4. Literatür Taraması

Sosyal medya platformlarında yer alan Türkçe hakaret ve küfür içeriklerinin otomatik olarak tespitine yönelik yapılan çalışmalar hem dil işleme tekniklerinin gelişimi hem de toplumsal iletişimdeki dijital denetim ihtiyacı açısından önem arz etmektedir. Bu bölümde, tez çalışmasının kapsamı doğrultusunda yalnızca Türkçe dilinde gerçekleştirilmiş çalışmalar ele alınmıştır. Literatür taraması, dört temel sosyal medya platformu Facebook, Instagram, X ve Reddit özelinde gruplandırılmıştır. Her platforma ilişkin yapılan önceki araştırmaların veri seti büyüklüğü, etiketleme yöntemleri, kullanılan sınıflandırma modelleri ve başarı oranları gibi yönleri incelenmiştir. Ayrıca, platformlara özgü çalışmalardan sonra, Türkçe verilerle yapılan platformlar arası bütünleşik çalışmalar ile uluslararası İngilizce literatür kısaca özetlenmiş; son olarak bu tezin literatüre sağlayacağı katkılar değerlendirilmiştir.

2.4.1. Facebook Üzerine Türkçe Çalışmalar

Yapılan kapsamlı literatür taramasına göre, yalnızca Facebook platformundan elde edilen Türkçe kullanıcı yorumları üzerinde hakaret veya küfür tespiti amacıyla gerçekleştirilmiş herhangi bir akademik çalışmaya rastlanmamıştır. Mevcut veri

kaynaklarının ve yöntemsel yaklaşımların bu platformu doğrudan hedef almadığı, Facebook içeriklerinin sistematik biçimde incelenmediği görülmektedir.

Bu bağlamda, bu tez çalışması, tamamı Türkçe olan, platforma özgü yorumlardan oluşan bir veri setiyle, yalnızca hakaret ve küfür içeriklerinin tespiti üzerine odaklanarak literatürdeki belirgin bir boşluğu doldurmaktadır.

2.4.2. Instagram Üzerine Türkçe Çalışmalar

Instagram platformuna ait Türkçe kullanıcı yorumları üzerinden yürütülen hakaret ve küfür tespiti çalışmalarına yönelik literatür incelendiğinde, bu alanda yapılmış sınırlı sayıda akademik çalışmaya ulaşılmıştır. Yapılan taramada yalnızca Instagram verisi kullanılarak, doğrudan küfür ya da hakaret içeriğini sınıflandırmayı hedefleyen bir çalışma tespit edilmiştir.

Karayiğit ve ark. [6] tarafından gerçekleştirilen bu çalışmada 2017–2019 yılları arasında Instagram' daki popüler magazin ve spor içerikli hesaplardan toplanan toplam 30.084 Türkçe yorum incelenmiştir. Yorumlar, içerdikleri ifadeler doğrultusunda iki sınıfa ayrılmış; küfür veya hakaret içeren ve içermeyen olarak manuel biçimde etiketlenmiştir.

Çalışmada hem klasik makine öğrenmesi algoritmaları hem de derin öğrenmeye dayalı bir Evrişimsel Sinir Ağı (CNN) modeli kullanılmıştır. Kullanılan veri yalnızca Instagram yorumları ile sınırlı olup, başka bir sosyal medya platformu verisi içermemektedir. Aşırı örnekleme uygulanmış veri setiyle gerçekleştirilen deneylerde CNN modeliyle %97 F1 skoru elde edilmiştir [6].

Söz konusu çalışma yalnızca ikili sınıflandırma (saldırgan / saldırgan olmayan) temelinde geliştirilmiştir. Bu yönüyle saldırgan içeriklerin türü ya da derecesi gibi daha ayrıntılı sınıflandırmalara olanak tanımamaktadır.

2.4.3. Reddit Üzerine Türkçe Çalışmalar

Reddit platformuna yönelik literatür taraması sonucunda, Türkçe kullanıcı yorumlarında yer alan hakaret ve küfür ifadelerinin otomatik biçimde tespit edilmesine yönelik herhangi bir sistematik araştırma bulunmadığı belirlenmiştir. Reddit'in yapısı gereği farklı alt forumlarda (subreddit'lerde) kullanılan dil ve ifade biçimleri oldukça çeşitlilik göstermekle birlikte, bu çeşitlilik mevcut çalışmaların çoğunda dikkate alınmamıştır. Dolayısıyla Reddit'te kullanılan Türkçe dilinin belirli içerik türleri

bağlamında incelenmesi, literatürde hâlen açık bir alan olarak durmaktadır. Bu tezde yer alan analizler, Reddit'ten manuel olarak toplanmış Türkçe yorumlar üzerinden gerçekleştirilmiş olup, bu platform özelinde hakaret/küfür içeriklerinin otomatik sınıflandırmasına yönelik literatürdeki ilk sistematik değerlendirmeyi sunmaktadır..

2.4.4. X Platformu Üzerine Türkçe Çalışmalar

X platformu, Türkçe dilinde yapılmış sosyal medya analizlerinde en sık tercih edilen platformlardan biridir. Yapılan kapsamlı literatür taramasında, yalnızca hakaret veya küfür tespiti amacıyla, X platformundan toplanmış Türkçe veriler üzerinde gerçekleştirilen çeşitli akademik çalışmalara rastlanmıştır. Bu çalışmaların bir kısmı doğrudan küfür veya hakaret içeren söylemlerin tespitiyle ilgilenmekte, bazıları ise daha geniş kategoriler (örneğin saldırgan dil veya nefret söylemi) altında küfürlü içerikleri ayrı bir sınıf olarak ele almaktadır. Bu bağlamda, yalnızca hakaret veya küfür içeren verilerle çalışan ve X platformu verisini temel alan çalışmalar aşağıda özetlenmiştir.

Çöltekin [4] tarafından sunulan çalışma, Türkçe sosyal medya dilinde saldırgan içeriklerin sınıflandırılması amacıyla oluşturulmuş 35.282 tweetten oluşan geniş bir veri setini tanıtmaktadır. Tweet'ler, manuel olarak etiketlenmiş olup, küfür ve hakaret içerikleri "hedefsiz" (örneğin doğrudan bir kişiyi hedef almayan genel küfürler) ve "hedefli" (örneğin bireye veya gruba yönelik) biçiminde sınıflandırılmıştır. Etiketleme işlemi, iki insan etiketleyici ve kapsamlı yönergeler eşliğinde gerçekleştirilmiştir. Yazar, bu veri seti üzerinde temel metin sınıflandırma algoritmalarıyla ön deneyler de yürütmüştür; bu doğrultuda ikili sınıflandırmada F1 skoru %77,3, hedefli/hedefsiz ayrımında %77,9 ve alt kategori ayrımında ise %53,0 olarak raporlanmıştır. Ancak bu çalışma, daha çok kapsamlı bir veri seti sunmayı hedeflemiş; modelleme kısmı ikincil düzeyde kalmıştır. Söz konusu çalışma, küfür ve hakaretin sınıflandırılması konusunda önemli bir kaynak olmakla birlikte, bu tezin odaklandığı yalnızca küfür/hakaret içeren içeriklerin sınıflandırılması yönüyle kapsamdan kısmen ayrılmaktadır.

Mayda ve ark. [17] tarafından hazırlanan çalışma, 1000 Türkçe tweet üzerinde gerçekleştirilmiş olup, içerikler "nefret söylemi", "saldırgan ifade" ve "hiçbiri" olmak üzere üç sınıfa ayrılmıştır. Etiketleme işlemi iki uzman tarafından yürütülmüş, tutarsız durumlar üçüncü bir hakem aracılığıyla netleştirilmiştir. Tweetlerin bir bölümü doğrudan küfür içeren ifadeleri içermektedir. Kullanılan yöntemler arasında Naive Bayes, Karar Ağaçları, SVM ve Rastgele Orman gibi klasik makine öğrenmesi algoritmaları yer

almaktadır; bu yöntemler arasında en yüksek başarı, SVM tabanlı SMO algoritması ile elde edilen %79,9 F1 skoru olmuştur.

Yılmaz ve ark. [18] tarafından yürütülen başka bir çalışmada, yaklaşık 15.000 adet Türkçe tweet üzerinde çok düzeyli bir saldırgan dil sınıflandırması gerçekleştirilmiştir. Tweetler öncelikle saldırgan-saldırgan değil olarak ayrılmış, ardından saldırgan içerikler "hedefsiz", "birey hedefli", "grup hedefli" ve "diğer" olarak alt kategorilere ayrılmıştır. Etiketleme işlemi insan annotator'larca yürütülmüş ve içeriklerdeki şiddetli küfür veya hakaret söylemlerinin sınıflandırılması sağlanmıştır. Bu çalışmada Word2Vec temelli gömlemeler kullanılmış ve hem LSTM hem de GRU mimarileriyle sınıflandırıcılar eğitilmiştir. GRU modeliyle ikili sınıflandırmada %94,49, çoklu sınıflandırmada ise %71,97 ve %54,10 F1 skorları elde edilmiştir.

Canbay ve Ekinci [19] tarafından yürütülen çalışmada, Türkçe sosyal medya verileri üzerinde küfür ve nefret söylemi tespiti için farklı derin öğrenme modelleri karşılaştırılmıştır. Araştırmada CNN, LSTM, BiLSTM, GRU ve BiGRU modelleri kullanılmış; her bir model için embedding, dropout ve ilgili katman yapıları detaylandırılmıştır. Modellerin performansları incelendğinde en yüksek f1-skor başarı oranı %74 ile BİLSTM modelinde elde edilmiştir. Çalışma, Türkçe içeriklerde küfür ve nefret söylemi tespitinde farklı derin öğrenme mimarilerinin etkinliğini ortaya koyması açısından literatüre önemli bir katkı sağlamaktadır.

2.4.5. Platformlar Arası Türkçe Veri Birleştirme Üzerine Çalışmalar

Yürütülen literatür taramasında, Facebook, Instagram, X ve Reddit platformlarından elde edilen Türkçe yorumları birlikte ele alarak yalnızca hakaret ve küfür içeriklerinin sınıflandırılmasını hedefleyen çok yönlü bir akademik çalışmaya rastlanmamıştır. Çoğu araştırma yalnızca tek bir platform üzerinde çalışmış; bazı örneklerde ise platform birleştirmesi yapılsa bile iki platformla sınırlı kalınmış ve veri dengesi gözetilmemiştir. Bu tez çalışması, dört büyük sosyal medya platformundan eşit sayıda Türkçe yorum içeren dengeli bir veri kümesi oluşturmasıyla mevcut yaklaşımlardan ayrılmaktadır. Veriler anlam düzeyine dayalı beşli etiketleme sistemiyle yeniden sınıflandırılmış, ayrıca "hakaret içeriyor / içermiyor" biçiminde ikili yapı kullanılarak model performansları kapsamlı biçimde değerlendirilmiştir. Böylece hem platformlara özgü modellerin avantajları belirlenmiş hem de tüm platformları kapsayan tek bir modelin genellenebilirliği analiz edilmiştir. Bu yönüyle çalışma, Türkçe doğal dil

işleme alanında platformlar arası karşılaştırmalı analizlere olanak tanıyan özgün ve bütüncül bir yaklaşım ortaya koymaktadır.

2.4.6. Büyük Dil Modelleriyle Türkçe Etiketleme Üzerine Çalışmalar

Türkçe sosyal medya içeriklerinde, büyük dil modelleri kullanılarak otomatik ve çok sınıflı etiketleme gerçekleştiren akademik çalışmalara yapılan literatür taramasında rastlanılmadı. Özellikle beş kategoriye dayalı sınıflandırma senaryolarında, GPT tabanlı modellerin Türkçe dilinde uygulanmasına yönelik sistematik bir değerlendirme bulunmamaktadır. Bu tez kapsamında gerçekleştirilen deneysel uygulamada, GPT-4o modeli kullanılarak sosyal medya yorumları otomatik olarak beşli biçimde etiketlenmiş ve elde edilen sonuçlar, insan uzmanlarca belirlenen etiketlerle karşılaştırılmıştır. Böylece, büyük dil modellerinin Türkçe metinlerde kavramsal ayrım yapma yetkinliği, doğrudan veri üzerinden ölçülerek değerlendirilmiştir.

2.4.7. Uluslararası Literatürde Benzer Çalışmalar (İngilizce)

Uluslararası literatürde, sosyal medya platformlarında hakaret tespiti alanında yapılan çalışmaların çoğunluğu İngilizce veri setleri kullanılarak gerçekleştirilmiştir. Bu çalışmalar genellikle X platformu ve Reddit gibi platformlardan toplanan veriler üzerinde yoğunlaşmakta ve farklı makine öğrenmesi ile derin öğrenme modelleri uygulanmaktadır.

Ashok ark. [20], İngilizce tweetler üzerinde çeşitli modeller kullanarak derin öğrenme temelli yaklaşımların etkinliğini araştırmıştır. Ancak bu çalışmada doğrudan bir sınıflandırma başarım oranı (örneğin doğruluk ya da F1 skoru) raporlanmamış, bunun yerine açıklayıcı cümle üretimi üzerine BLEU, ROUGE gibi ölçütler kullanılmıştır. Bihari ark. [21] İngilizce Twitter verisi üzerinde CNN ve LSTM modellerini kullanmış; önerilen CNN+LSTM yapısı ile %92 doğruluk ve %92 F1 skoru elde etmişler. Hada ark. [22] Reddit yorumları üzerinde BERT ve HateBERT modellerini uygulayarak dil modeli tabanlı yaklaşımların saldırganlık tespiti için uygunluğunu göstermiştir; HateBERT modeliyle Pearson korelasyon katsayısı ~0.886 (MSE=0.025) olarak raporlanmıştır. Lee ark. [23], Twitter veri seti üzerinde tekrarlayan sinir ağlarını kullanmış; çift yönlü GRU temelli modelleriyle %80,5 F1 skoru elde etmişlerdir. Park ve Fung [24] ise İngilizce tweetler üzerinde çok aşamalı sınıflandırma yöntemleriyle İngilizce tweetler üzerinde çalışmış; Hybrid-CNN modeliyle %82,7, iki adımlı lojistik regresyon yaklaşımıyla ise %82,4 F1 skoruna ulaşmışlardır. Badjatiya ark. [25], İngilizce Twitter verisi üzerinde

farklı derin öğrenme teknikleri ile nefret söylemi ve hakaret tespiti yapmış; en başarılı modellerinde (LSTM+GBDT) F1 skorunu %93 olarak rapor etmişlerdir.

Sonuç olarak, yapılan kapsamlı literatür taraması neticesinde, sosyal medya platformlarında hakaret tespiti üzerine özellikle İngilizce dilinde yürütülmüş çalışmaların Facebook, Twitter, Instagram ve Reddit gibi farklı platformlara yönelik ayrı ayrı veya sınırlı biçimde mevcut olduğu tespit edilmiştir. Ancak, dört farklı sosyal medya platformunu bir arada kapsayan ve karşılaştırmalı analizler içeren kapsamlı bir çalışma bulunmamaktadır. Bu durum, çok platformlu ve çok dilli veri setleri kullanılarak gerçekleştirilecek yeni araştırmalar için önemli bir boşluk ve ihtiyaç olduğunu ortaya koymaktadır.

2.4.8. Literatürün Değerlendirilmesi ve Bu Tezin Katkıları

Yapılan kapsamlı literatür incelemesi sonucunda, Türkçe sosyal medya içeriklerinde hakaret ve küfür tespiti alanındaki çalışmaların büyük bölümünün X (Twitter) platformu ile sınırlı kaldığı görülmektedir. Facebook ve Reddit gibi geniş kullanıcı kitlesine sahip platformlar üzerinde Türkçe içerikleri esas alan sistematik bir hakaret tespiti çalışmasına rastlanmamıştır. Instagram özelinde yürütülen az sayıdaki araştırmalar ise çoğunlukla ikili sınıflandırma (saldırgan / saldırgan olmayan) düzeyinde kalmış, saldırgan dilin alt kategorilerine ilişkin detaylı ayrımlar yapılmamıştır. Bu durum, özellikle Türkçe dilinin yapısal özelliklerini ve platforma özgü dil kullanım biçimlerini dikkate alan çok yönlü karşılaştırmalı çalışmalara olan ihtiyacı ortaya koymaktadır.

Mevcut çalışmaların önemli bir kısmı sınırlı kapsamlı veri setleri üzerinde gerçekleştirilmiş olup çoğunlukla yalnızca tek bir platform verisine odaklanmıştır; platformlar arası veri birleştirme yaklaşımı ise ya hiç uygulanmamış ya da yalnızca iki platformla sınırlı tutulmuştur. Buna ek olarak, çoğu çalışma geleneksel makine öğrenmesi yöntemleriyle yürütülmüş, derin öğrenmeye dayalı mimariler ise yalnızca sınırlı örneklerde değerlendirilmiştir. Aynı şekilde, hakaret içeriklerinin beşli sınıflandırmaya dayalı detaylı biçimde etiketlenmesine yönelik çalışmaların oldukça sınırlı olduğu, bu nedenle anlam düzeyindeki içerik çeşitliliğinin büyük ölçüde göz ardı edildiği dikkat çekmektedir.

Bu tez çalışması tüm bu eksikliklere cevap verecek şekilde tasarlanmış olup, Facebook, Instagram, X ve Reddit platformlarından eşit sayıda yorum içeren dengeli bir veri seti oluşturmuş, hem çok sınıflı (beş kategoriye dayalı) hem de ikili (“hakaret var /

yok”) sınıflandırma senaryoları üzerinden üç farklı derin öğrenme mimarisini (BERTurk, CNN ve LSTM) kapsamlı biçimde değerlendirmiştir. Böylece hem her bir platformun kendine özgü dil yapısının model performansı üzerindeki etkisi incelenmiş hem de platformların birleştirilmesiyle eğitilmiş tek bir modelin genellenebilirliği test edilmiştir. Ayrıca, GPT-4o tabanlı otomatik etiketleme yöntemleri sıfır ve az örnekli senaryolar üzerinden uygulanarak, büyük dil modellerinin Türkçe sosyal medya verilerinde çok sınıflı hakaret tespitindeki etkinliği ilk kez sistematik olarak değerlendirilmiştir. Bu yönleriyle çalışma, yalnızca farklı mimarileri karşılaştırması bakımından değil, kapsamlı veri kaynağı ve metodolojik çeşitliliğiyle de Türkçe doğal dil işleme alanına özgün ve bütüncül bir katkı sunmaktadır.

3. MATERYAL VE YÖNTEM

Bu tez çalışmasında, Türkçe sosyal medya yorumlarında yer alan hakaret ve küfür içeriklerinin tespitine yönelik olarak dört farklı platformdan (X, Instagram, Facebook, Reddit) toplanan verilerle derin öğrenme modelleri eğitilmiştir. Veriler, Python programlama dili ile geliştirilen web kazıyıcı (scraper) yapılar aracılığıyla 2024 Eylül – 2025 Ocak tarihleri arasında farklı etkileşim düzeylerine sahip haber sayfalarından elde edilmiştir.

Etiketleme sürecinde, Çöltekin [4] tarafından tanımlanan yönergelere dayalı beş kategorili bir sistem (Non, Ind, Grp, Oth, Prof, X) benimsenmiş, yorumlar üç kişilik bir ekip tarafından uzlaşmalı şekilde etiketlenmiştir. Veri seti, hem her bir platformun kendi içinde hem de tüm platformların birleşimiyle oluşturulmuş tekil veri kümesi üzerinden analiz edilmiştir.

Bu kapsamda, yalnızca küfür ve hakaret tespiti hedeflenmiş; nefret söylemi, toksik dil, siber zorbalık gibi kapsam dışı konulara yer verilmemiştir. Yorumların sınıflandırılması amacıyla üç farklı derin öğrenme mimarisi kullanılmıştır: BERTurk (transformer tabanlı), CNN (evrişimsel sinir ağı) ve LSTM (uzun kısa süreli bellek ağı). Her model hem platform bazlı hem de birleşik veri ile ayrı ayrı eğitilmiş; model performansları doğruluk, F1 skoru, doğruluk ve hassasiyet gibi metriklerle değerlendirilmiştir.

Model eğitimi ve test işlemleri Google Colab üzerinde gerçekleştirilmiş; eğitim sürecinde GPU destekli donanım altyapısından faydalanılmıştır.

3.1. Veri Setinin Özellikleri ve Etiketleme Süreci

Bu çalışmada, sosyal medya platformlarında yapılan kullanıcı yorumları üzerinden Türkçe küfür ve hakaret içeriğinin tespiti amaçlanmıştır. Veri seti, dört farklı sosyal medya platformundan Facebook, Instagram, X platformu ve Reddit derlenen yorumlardan oluşmaktadır. Veriler, Eylül 2024-Ocak 2025 tarihleri arasında, her platformda yüksek etkileşim gösteren haber sayfalarından toplanmıştır. Böylece, kullanıcıların güncel olaylara yönelik tepkilerini içeren doğal yorumlar analiz kapsamına alınmıştır.

Toplam 43.582 adet yorum içeren veri seti, her platformdan benzer büyüklükte örnekleme yapılarak oluşturulmuş ve aşağıda sunulan tabloya göre dağılım göstermektedir: Platformlara göre toplanan yorum sayıları Çizelge 3.1'de gösterilmiştir.

Çizelge 3.1. Platform bazlı yorum sayıları

Platform	Yorum Sayısı
Facebook	11.716
Instagram	11.133
X Platformu	10.674
Reddit	10.059
Toplam	43.582

Verilerin içeriksel özellikleri incelendiğinde, platformlar arası dil ve söylem farklarının olduğu gözlemlenmiştir. Örneğin X platformunda kısa ve yoğun mesajlar ön plana çıkarken, Reddit'te uzun ve tartışmaya açık yorumlar göze çarpmaktadır. Facebook ve Instagram ise gündeme dayalı kısa, çoğunlukla tepkisel yorumlara sahiptir. Platform bazlı veri dağılımları Çizelge 3.2' de verilmiştir.

Veri seti, yalnızca küfür ve hakaret içeriklerini hedeflemiş; nefret söylemi, toksik dil veya siber zorbalık gibi diğer türler kapsam dışı bırakılmıştır. Etiketleme işlemi, Çöltekin [4] tarafından geliştirilen yönergelere uygun olarak beş kategoriden oluşan bir sistemle gerçekleştirilmiştir:

- **Non:** Hakaret içermeyen yorumlar
- **Ind:** Bireye yönelik hakaret
- **Grp:** Gruba yönelik hakaret
- **Oth:** Olay veya kuruma yönelik hakaret
- **Prof:** Nesneye yönelik ya da şaka yollu hakaret

Etiketleme süreci şu aşamalardan oluşmuştur:

1. Her yorum iki bağımsız etiketleyici tarafından işaretlenmiştir.
2. Etiket uyuşmazlığı olan yorumlar üçüncü bir hakem tarafından değerlendirilmiştir.
3. Nihai karar hakem görüşü doğrultusunda verilmiştir.

Çizelge 3.2. Platform bazlı kategorik sınıflandırma sayıları

Platform	Non	Ind	Prof	Grp	Oth	Toplam
Facebook	8.197	2.870	98	457	94	11.716
Instagram	10.284	584	180	47	38	11.133
X Platformu	7.823	2.231	375	93	152	10.674
Reddit	8.104	828	774	153	200	10.059
Toplam	34.408	6.513	1.427	750	484	43.582

Etiketleyicilerin değerlendirmeleri kimi zaman farklılık göstermiş, bu farklılıklar çoğunlukla hedef türünün yorumlanmasından, bağlam eksikliğinden ya da kelimelerin hakaret mi yoksa tanımlama mı sayılacağına ilişkin öznellikten kaynaklanmıştır. Aşağıda yer alan bazı örnekler, bu ayrışmaların hangi dilsel ve anlamsal belirsizliklerden doğduğunu somut biçimde ortaya koymaktadır. Üç farklı etiketleyici metin içinde A, B ve C olarak gösterilmiştir.

Örnek 1

"Kadınlar en çok kocaları tarafından dövülüp işkence edilip katlediliyorlar. Aile içi tecavüzler de ayrı konu İslami hassasiyetli kesim."

(A: non – B: grp – C: grp)

Bu ifadede A, saldırganlık unsuru görmeyerek non etiketini seçmiştir. Ancak B ve C, "İslami hassasiyetli kesim" ifadesinin doğrudan bir grubu hedef aldığını düşünerek grp kategorisine yönelmiştir. Ayrışmanın nedeni, ifadenin ilk bölümünde toplumsal bir olgu aktarılırken ikinci bölümde belirli bir grubun vurgulanmasıdır. Bu iki katmanlı yapı, bazı etiketleyicilerin bütüncül, bazılarının ise parça parça değerlendirme yapmasına yol açmıştır.

Örnek 2

*"Tamam hadi şimdi İslam hassasiyetiyle s**tir git!!"*

(A: ind – B: grp – C: ind)

Bu örnekte A ve C, ifadenin kişisel bir muhataba doğrudan hakaret içerdiğini düşünerek ind etiketini uygun görmüştür. Buna karşılık B, "İslam hassasiyeti" ifadesinin belirgin bir gruba gönderme yaptığı kanaatiyle grp seçmiştir. İfade hem bireye yönelmiş açık bir hakaret ("s**tir git"), hem de dini kimliği işaret eden grup hedeflemesi içerdiği için çifte hedef olasılığı ortaya çıkmakta ve etiketleyiciler arasında ayrışmaya neden olmaktadır.

Örnek 3

"Bu kadın narsist."

(A: non – B: ind – C: ind)

Bu ifadede B ve C, kişisel bir niteleme bulunduğunu ve bunun hakaret kapsamında değerlendirilebileceğini düşünerek ind seçmiştir. A ise "narsist" sözcüğünü psikolojik bir kişilik özelliği olarak yorumlayıp saldırganlık içermediğini varsayarak non etiketi vermiştir. Buradaki ayrışma, kelimenin hakaret mi yoksa betimleme mi sayılacağına ilişkin öznellikten kaynaklanmaktadır.

Örnek 4

"Beynine içine ot koysunlar geri zekalı mallar."

(A: ind – B: oth – C: ind)

Bu örnekte A ve C, ifadenin kişisel saldırı içerdiğini değerlendirerek ind seçmiştir. B ise doğrudan bir muhatap belirtilmediğini, ifadenin hedefsiz küfür barındırdığını düşünerek oth kategorisini tercih etmiştir. Ayrışma, hakaretin belirli bir kişiye mi yoksa belirsiz bir özneye mi yöneltildiği konusunda ortaya çıkmıştır.

Örnek 5

"Bunlar kadar alçak yalancı yok."

(A: ind – B: oth – C: ind)

A ve C, "bunlar" ifadesini doğrudan bir topluluğa ya da muhataba yönelik saldırı olarak değerlendirmiş ve ind seçmiştir. Buna karşılık B, muhatabın belirsizliği nedeniyle ifadenin kurumsal ya da hedefi belirsiz bir saldırı içerdiğini varsayarak oth kategorisine yönelmiştir. Buradaki ayrışma, "bunlar" zamirinin hangi bağlama işaret ettiğine dair belirsizlikten kaynaklanmaktadır.

Örnek 6

"Benzin döküp yakın, böylesi itleri içerde beslemeye değmez."

(A: ind – B: oth – C: ind)

A ve C, bu ifadeyi doğrudan kişilere veya belirli bir topluluğa yönelik ağır saldırı olarak yorumlamış ve ind etiketi vermiştir. B ise muhatabın açıkça belirtilmediği görüşüyle

ifadeyi oth kategorisine yerleştirmiştir. Belirsiz özne nedeniyle hedefli veya hedefsiz hakaret ayrımında farklılık oluşmuştur.

Örnek 7

"Kişi kendinden bilir işi. O sizin işiniz mhk ile yan yana giden şikeci tarikat."

(A: oth – B: grp – C: grp)

Bu örnekte B ve C, "tarikat" sözcüğünü doğrudan grup hedeflemesi olarak görmüş ve grp seçmiştir. A ise ifadenin kurum (MHK) üzerinden eleştiri içerdiğini düşünerek oth kategorisini tercih etmiştir. Buradaki ayrışma, ifadenin hedefinin dini grup mu yoksa kurumsal yapı mı olduğuna dair farklı yorumlardan kaynaklanmaktadır.

Örnek 8

"Niye herkes nevzat mağdur muş gibi konuşuyor nevzat bir canidir nevzat acımasızdır küçük bir kızı acımasızca katletmiş."

(A: ind – B: non – C: ind)

A ve C, "cani" ve "acımasız" ifadelerini doğrudan kişisel hakaret olarak yorumlayıp ind etiketini seçmiştir. B ise bu söylemi hakaret değil, suç isnadı ya da olay aktarımı olarak görmüş ve non kategorisine yönelmiştir. Ayrışmanın temelinde, kullanılan nitelemelerin hakaret mi yoksa betimleyici bir suç atfı mı olduğuna ilişkin yorum farkı bulunmaktadır.

Örnek 9

*"Bir kez daha siyasetin b**tan bir şey olduğunu bize gösterdiğiniz için teşekkür ederiz."*

(A: oth – B: prof – C: prof)

Bu ifadede B ve C, küfrün belirli bir hedef içermediğini düşünerek prof etiketini tercih etmiştir. A ise saldırının doğrudan siyasete yöneltildiğini değerlendirerek oth kategorisini seçmiştir. Ayrışma, küfrün hedefsiz mi yoksa bir olguya/kavrama mı yöneldiği konusunda ortaya çıkmıştır.

Örnek 10

"Ya siz artık bir gidin ya bunadınız başımıza bela oldunuz!"

(A: ind – B: non – C: ind)

A ve C, ifadenin doğrudan kişilere yönelik açık hakaret içerdiğini düşünerek ind etiketini uygun görmüştür. B ise bu söylemi şikâyet veya tepki ifadesi olarak değerlendirmiş ve

saldırganlık unsuru görmeyerek non seçmiştir. Ayrışma, “bunadınız” ifadesinin hakaret sayılıp sayılmayacağına ilişkin öznellikten kaynaklanmaktadır.

Bu örnekler göstermektedir ki, hakaret tespitinde en büyük zorluk hedefin niteliğinin (birey, grup, kurum veya hedefsiz küfür) doğru biçimde belirlenmesidir. Aynı ifade, bir etiketleyici tarafından kişisel saldırı olarak algılanırken bir diğeri tarafından grup ya da kuruma yönelik değerlendirilmekte; kimi zaman da suç isnadı ile hakaret arasındaki sınır bulanıklaşmaktadır. Dolayısıyla etiketleyiciler arasındaki uyumsuzluk, yalnızca teknik yönergelerden değil, dilin bağlamsal ve yorumsal doğasından da beslenmektedir.

3.2. Ön İşleme ve Temizleme Adımları

Bu çalışmada kullanılan sosyal medya verileri, ham haliyle çeşitli gürültüler içermekte olup doğrudan model eğitimine uygun değildir. Bu nedenle, Türkçe metinlere özgü dil yapısı dikkate alınarak kapsamlı bir ön işleme süreci gerçekleştirilmiştir. Ön işleme adımları, literatürdeki benzer çalışmalardan da yararlanılarak yapılandırılmıştır [26]

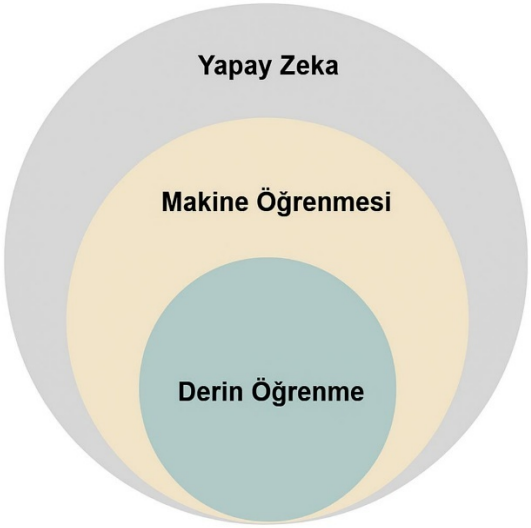
İlk aşamada tüm metinler küçük harflere dönüştürülmüştür. Bu işlem, Türkçe’nin büyük-küçük harf ayrımının model öğrenmesinde anlamlı bir fark yaratmaması ve sözcük çeşitliliğini azaltma amacıyla uygulanmıştır. Ardından, kullanıcı adları (@ile başlayan ifadeler), URL bağlantıları (http ile başlayan diziler), emojiler, noktalama işaretleri ve gereksiz semboller gibi gürültü öğeleri temizlenmiştir. Bu adımlar, özellikle sosyal medya içeriklerinin serbest biçimli ve yapısız doğası göz önüne alındığında, veri kalitesini artırmak açısından önemlidir [27].

Daha sonra, durak kelime (stop-word) temizliği yapılmıştır. Türkçe’de anlam bakımından katkısı düşük, ancak sıklıkla kullanılan bu kelimelerin çıkarılması, modelin daha anlamlı örüntüler öğrenmesini sağlamak açısından önem arz etmektedir [28]. Bu aşamada, güncel Türkçe durak kelime listeleri kullanılmıştır.

Son olarak, Türkçe’nin eklemeli yapısı nedeniyle kök bulma (lemmatization veya stemming) işlemi gerçekleştirilmiştir. Bu işlem için iki farklı yöntem uygulanmıştır: Zemberek kütüphanesi aracılığıyla sözcüklerin kökleri elde edilmiş; ayrıca NLTK kütüphanesinin SnowballStemmer bileşeni kullanılarak alternatif kök bulma süreci yürütülmüştür. Böylece, aynı anlama gelen ancak farklı ekler almış sözcükler tekil temsillere indirgenerek modelin genel performansına katkı sağlanmıştır [26].

3.3. Derin Öğrenme

Derin öğrenme, çok katmanlı yapay sinir ağları kullanarak veriden yüksek düzeyde özellikler öğrenebilen bir makine öğrenmesi yaklaşımıdır. Özellikle çok boyutlu ve karmaşık yapıdaki veriler üzerinde başarılı sonuçlar veren bu yöntem; görüntü işleme, ses tanıma, doğal dil işleme ve biyomedikal keşifler gibi pek çok farklı alanda yaygın biçimde kullanılmaktadır. Derin öğrenme mimarileri, parametrelerini optimize etmek ve hata oranlarını azaltmak amacıyla genellikle geri yayılım (backpropagation) algoritmasından yararlanır. Bu sayede model, verideki karmaşık örüntüleri katmanlar aracılığıyla öğrenerek temsil gücünü artırır [29].



Şekil 3.1. Yapay zeka şeması

Makine öğrenmesinin bir alt dalı olan derin öğrenme, özellikle büyük veri kümesine ve yüksek hesaplama gücüne sahip ortamlarda klasik yöntemlere kıyasla çok daha başarılı sonuçlar verebilmektedir. Derin öğrenme bu yönüyle; ilaç moleküllerinin etki düzeylerini tahmin etmeden, DNA mutasyonlarının hastalıklarla ilişkilendirilmesine kadar birçok kritik problemin çözümünde çığır açıcı bir rol üstlenmiştir. Sosyal medya verilerinin içerdiği dilsel çeşitlilik ve bağlamsal karmaşıklık düşünüldüğünde, bu tezde derin öğrenme modelleri (BERTurk, CNN, LSTM) kullanılmıştır.

3.3.1. Evrişimsel Sinir Ağı (CNN)

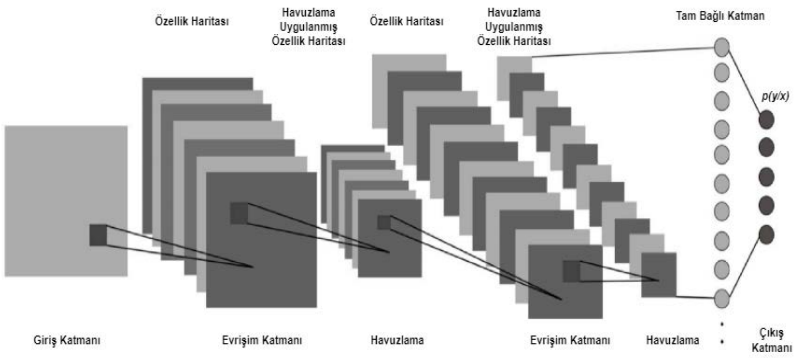
Evrişimsel Sinir Ağları (Convolutional Neural Networks – CNN), ilk olarak görüntü işleme alanında kullanılmak üzere geliştirilmiş olsa da, özellikle metin sınıflandırma gibi doğal dil işleme (NLP) problemlerinde de etkili sonuçlar vermektedir.

Metin tabanlı uygulamalarda CNN'ler, giriş cümlelerini kelime gömme (word embedding) matrisine çevirerek, bu matris üzerinde evrişim çekirdekleri (convolution filters) ile gezerek öne çıkan n-gram desenlerini yakalayabilir [30].

CNN mimarileri, metindeki lokal bağımlılıkları ve örüntüleri öğrenmekte başarılıdır. Bu sayede, belirli bir küfür veya hakaret ifadesi gibi kısa ama belirleyici yapılar, modelin erken evrelerinde tespit edilebilir. Özellikle sosyal medya gibi dilin oldukça informal ve yapısal olmayan şekilde kullanıldığı ortamlarda, CNN'in bu özelliği avantaj sağlar. Ayrıca RNN tabanlı modellere göre eğitim süresi daha kısa ve paralel işlemeye daha uygun olması sayesinde, büyük veri kümeleriyle daha verimli çalışabilir.

Tipik bir metin sınıflandırma CNN mimarisi; gömme (embedding) katmanıyla başlayan, ardından farklı boyutlarda evrişim (convolution) ve havuzlama (pooling) işlemleri içeren, nihayetinde tam bağlantılı (fully connected) sınıflayıcı bir yapıya sahiptir. Bu yapı, küfür veya hakaret içerikli ifadelerin yer aldığı metinleri, örüntüsel yapıları üzerinden ayrıştırabilmektedir [31].

Bu tez çalışmasında, Türkçe sosyal medya yorumları ön işlemden geçirilmiş ve Tokenizer yardımıyla dizilere dönüştürülmüştür. CNN modeli, bu diziler üzerinden eğitim almıştır. CNN modeli genel yapısı şekil 3. 2'de gösterilmektedir.



Şekil 3.2. CNN mimarisinde bulunan katmanlar [32]

Evrişimsel Sinir Ağları, geleneksel makine öğrenmesi sınıflandırıcılarına kıyasla daha başarılı sonuçlar üretmesinin temel nedenlerinden biri, veriden doğrudan anlamlı temsiller elde edebilen öznitelik çıkarımı ve örnekleme süreçlerini içeren katmanlara sahip olmalarıdır [33].

3.3.1.1. Giriş Katmanı (Input Layer)

Evrişimsel Sinir Ağlarında, giriş katmanı modelin ilk temas noktasıdır ve ham verilerin (örneğin metinlerin) sayısal temsilleri burada alınır. Giriş verisinin boyutu doğrudan hesaplama maliyetini etkiler. Büyük boyutlu veriler daha fazla ayrıntı içerirken işlem süresini uzatır; küçük boyutlu veriler ise hızlı işlenir ancak bilgi kaybına neden olabilir [34].

3.3.1.2. Evrişim Katmanı (Convolution Layer)

Bu katman, CNN'nin temel bileşenlerinden biridir. Filtreler (kernel) ve özellik haritaları kullanılarak girdi matrisi üzerinde doğrusal olmayan örüntüler yakalanır. Filtre, görüntü veya metin matrisine soldan sağa ve yukarıdan aşağıya kaydırılarak uygulanır. Her adımda yapılan işlem çapraz korelasyon işlemidir [33].

Evrişim katmanında uygulanan işlemin matematiksel temeli eşitlik 3.1'de gösterilmiştir.

$$(f * h)[m,n] = \sum_{\mathbf{j}} \sum_{k} h[j,k], f[\mathrm{m} - j, n - k] \tag{3.1}$$

Burada:

- f: Giriş verisi (örneğin kelime gömme matrisi),
- h: Filtre (kernel),
- m, n: Çıkış matrisinin satır ve sütun indeksleri,
- j, k: Elde edilen yeni matrisin satır ve sütun indeksleri

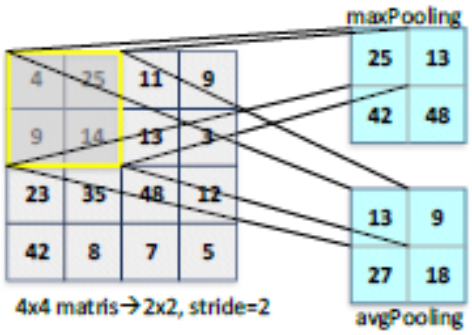
Bu formül, her filtre konumunda yapılan çarpım-toplam işlemini gösterir. Filtre, giriş verisi üzerinde konumlandırıldığında, örtüştüğü bölgedeki değerlerle çarpılır ve elde edilen sonuçlar toplanarak çıktı matrisinin ilgili hücresine aktarılır. Bu işlem, tüm giriş matrisi üzerinde sistematik biçimde uygulanarak, girişin belirli özelliklerini öne çıkaran yeni bir özellik haritası oluşturur [34].

3.3.1.3. Havuzlama Katmanı (Pooling Layer)

Evrişim işleminden sonra elde edilen özellik haritalarındaki boyutları azaltmak ve gereksiz bilgileri ortadan kaldırmak için havuzlama işlemi uygulanır. Bu işlem alt

örnekleme (subsampling) olarak da bilinir. Bu katmanda maksimum havuzlama (Max Pooling) ve ortalama havuzlama (Average Pooling) gibi teknikler kullanılmaktadır.

Filtre ve adım sayısı(stride) parametreleri ile yapılan bu işlemler hesaplama maliyetini azaltır ve modelin genelleme yeteneğini artırır [33], [34].



Şekil 3.3. Havuzlama katmanına ait işlem adımları [33]

3.3.1.4. Tam Bağlı Katman (Fully Connected Layer)

Havuzlama katmanından gelen veriler düzleştirilerek (flatten) bir vektöre dönüştürülür. Bu katman, bu vektör üzerinden sınıflandırma işlemi yapar. Örneğin 4x4x35 boyutunda bir özellik haritası, bu katmanda 560 boyutlu bir vektöre çevrilir ve çıkışa aktarılır [34].

3.3.1.5. Softmax Katmanı (Softmax Layer)

Softmax katmanı, lojistik regresyonun çok sınıflı problemler için genelleştirilmiş halidir ve olasılık temelli bir sınıflandırıcı olarak çalışır. Genellikle bir sinir ağının çıkış katmanı olarak kullanılır ve iki ya da daha fazla sınıfa ait olasılıkların hesaplanmasında tercih edilir. Bu katman, her sınıfa ait olasılığı [0, 1] aralığında bir değer olarak üretir ve tüm sınıfların olasılıklarının toplamı 1 olacak şekilde normalize eder.

Softmax fonksiyonu aşağıdaki matematiksel ifasi eşitlik 3.2'de verilmiştir.

$$y_r(x) = \frac{exp(a_r(x))}{\sum_{j=1}^{k} exp(a_j(x)}; 0 \leq \mathrm{y_r} \leq \sum_{j=1}^{k} y_j = 1 \qquad (3.2)$$

Bu formülde her bir sınıfın aktivasyon değeri önce üssel fonksiyonla dönüştürülür, ardından tüm sınıfların üssel değerlerinin toplamına bölünerek normalize edilir [33], [34].

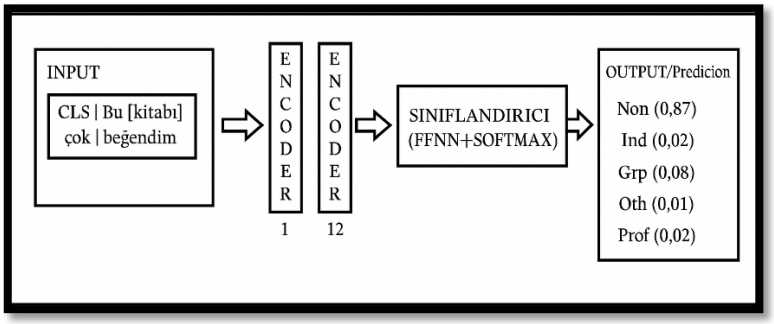
3.3.2. BERTurk (Transformer Tabanlı Dil Modeli)

BERTurk, Google tarafından önerilen BERT (Bidirectional Encoder Representations from Transformers) mimarisine dayanan ve yalnızca Türkçe metinler üzerinde önceden eğitilmiş bir dil modelidir. Modelin temelini oluşturan Transformer mimarisi, geleneksel sıralı ağlardan farklı olarak her bir girdinin tüm diğer girdilerle olan ilişkisini hesaplayan self-attention mekanizmasına dayanır [35]. Bu yapı sayesinde, kelimeler bağlam içinde çift yönlü (hem soldan sağa hem sağdan sola) olarak analiz edilebilir.

Transformer mimarisi katmanlı yapıya sahiptir ve her katmanda çok başlı dikkat (multi-head attention), normalize edilmiş besleme (feed-forward) katmanları ve kalıntı bağlantılar (residual connections) yer alır. Bu yapı, metinler üzerinde derin ve bağlamsal temsiller öğrenilmesini sağlar. BERT modelleri, girdi metinlerinde maskeleme (masked language modeling) ve cümle ilişkisi tahmini (next sentence prediction) gibi görevlerle önceden eğitilir. Böylece model, dilin yapısal özelliklerini öğrenmiş olur [36].

BERTurk modeli ise, bu mimarinin Türkçe diline özgü bir versiyonudur. Model, yaklaşık 35 GB'lık Türkçe Wikipedia, haber siteleri ve sosyal medya verilerinden oluşan veri kümesi üzerinde eğitilmiştir [37].

Bert model sınıflandırma genel işleyişi, şekil 4.4'te gösterilmiştir. Girdi metni, WordPiece tokenizasyonu ile alt bileşenlerine ayrılarak CLS ve SEP gibi özel simgelerle birlikte modele aktarılmaktadır. Token dizisi, 12 katmanlı encoder bloklarından geçirilerek bağlamsal temsillere dönüştürülmekte; son olarak, [CLS] tokenın çıktısı sınıflandırma katmanına verilerek yorumun ait olduğu sınıf belirlenmektedir [38].

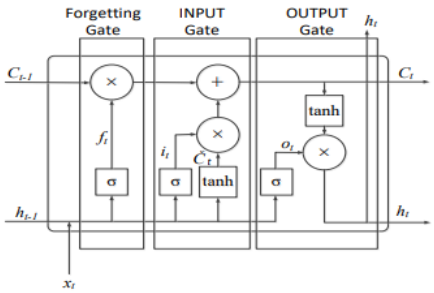


Şekil 3.4. Bert model sınıflandırma [38]

3.3.3. LSTM (Uzun Kısa Süreli Bellek Ağı)

LSTM (Long Short-Term Memory), özellikle sıralı verilerle çalışmak üzere tasarlanmış bir tür tekrarlayan sinir ağıdır (Recurrent Neural Network – RNN). Geleneksel RNN'ler, zaman içinde bilgi aktarımı yaparak ardışık veri analizinde başarılı olsa da, uzun süreli bağımlılıkları öğrenmede zorluk yaşamaktadır. Bu durum, gradyanların kaybolması (vanishing gradient) problemi olarak literatürde geniş yer bulmuştur. LSTM mimarisi, bu sınırlamayı aşmak üzere geliştirilmiştir [39].

LSTM'nin temel yapısı, giriş kapısı (input gate), unutma kapısı (forget gate) ve çıkış kapısı (output gate) olmak üzere üç ana bileşene dayanır. Bu kapılar sayesinde, hangi bilginin hücrede tutulacağına, hangisinin güncelleneceğine veya tamamen unutulacağına karar verilir. Bu yapı, modelin daha önceki zaman adımlarından gelen bilgileri etkili biçimde taşımasına olanak sağlar. Metin sınıflandırma görevlerinde, özellikle bağlamın ve kelime sırasının önemli olduğu durumlarda LSTM ağları tercih edilmektedir [40], [41]. LSTM yapısı Şekil 3.5' te gösterilmiştir.



Şekil 3.5. LSTM Yapısı [42]

Sosyal medya içerikleri genellikle bağlama bağlı, kısa ve dağınık ifadelerden oluşur. Bu tür yapıların işlenmesinde, LSTM'nin bağlamı taşıyabilen hafızası oldukça avantaj sağlar. LSTM modelleri, küfür ve hakaret gibi bağlamdan bağımsız değerlendirildiğinde anlamını kaybedebilecek ifadeleri doğru şekilde sınıflandırma potansiyeline sahiptir [43].

Bu tez çalışmasında kullanılan LSTM modeli, ön işlemden geçirilmiş Türkçe metin dizileri üzerine uygulanmıştır. Metinler, Tokenizer ile sayısal dizilere dönüştürülmüş, ardından sabit uzunlukta pad_sequences ile hizalanmıştır. Modelin yapısı şu katmanlardan oluşmaktadır:

- Gömme katmanı (Embedding)
- LSTM katmanı

- Yoğun (Dense) katmanlar ve Softmax çıkış

Model, hem platformlara özgü veri setleriyle hem de birleşik veriyle eğitilmiş ve karşılaştırmalı analizlerde CNN ve BERTurk modelleriyle birlikte değerlendirilmiştir. LSTM'nin bağlama duyarlı yapısı sayesinde özellikle bireysel hakaret gibi nüanslı ifadelerin sınıflandırılmasında güçlü sonuçlar verdiği gözlemlenmiştir.

3.4. Performans Ölçütleri

Confusion Matrix (Karmaşıklık Matrisi): Confusion matrix (karmaşıklık matrisi), bir sınıflandırma modelinin tahmin performansını değerlendirmek amacıyla, modelin doğru ve yanlış sınıflandırmalarını tablo biçiminde gösteren yapıdır [44].

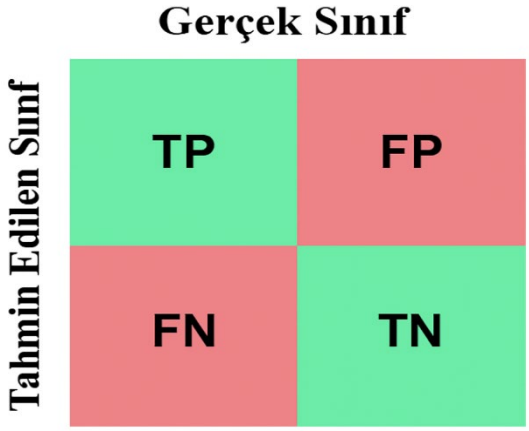
Gerçek etiketlerle modelin tahmin ettiği etiketlerin karşılaştırılması sonucunda aşağıdaki dört temel durum ortaya çıkar:

Gerçek Pozitif (True Positive – TP): Model, olumlu bir durumu doğru şekilde olumlu olarak tahmin etmiştir.

Gerçek Negatif (True Negative – TN): Model, olumsuz bir durumu doğru şekilde olumsuz olarak tahmin etmiştir.

Yanlış Pozitif (False Positive – FP): Model, olumsuz bir durumu yanlışlıkla olumlu olarak tahmin etmiştir.

Yanlış Negatif (False Negative – FN): Model, olumlu bir durumu yanlışlıkla olumsuz olarak tahmin etmiştir [44], [45].



Şekil 3.6. Karmaşıklık matrisi [46]

Model performansı, true positive, false positive, false negative ve true negative gibi dört temel ölçüte göre değerlendirilmekte olup, bu durum Şekil 3.6'da gösterilmektedir.

Doğruluk (Accuracy): Modelin tüm tahminleri içerisinde, doğru sınıflandırdığı örneklerin oranını gösteren temel başarı ölçütüdür. Genellikle genel performansı özetlemek için kullanılır. Matematiksel tanımı Eşitlik 3.3'te verilmiştir [44], [45].

$$Doğruluk = \frac{TP+TN}{TP+TN+FP+FN} \tag{3.3}$$

Kesinlik (Precision): Pozitif olarak tahmin edilen örneklerin ne kadarının gerçekten pozitif olduğunu gösteren metriktir. Bu ölçüt, modelin yanlış pozitif oranını düşürme becerisini değerlendirir. Matematiksel ifadesi Eşitlik 3.4'te sunulmuştur [46].

$$Kesinlik = \frac{TP}{TP+FP} \tag{3.4}$$

Duyarlılık (Recall): Gerçek pozitif örneklerin ne kadarının model tarafından doğru şekilde tahmin edildiğini gösterir. Bu metrik, özellikle pozitif sınıfların belirlenmesinde modelin başarısını değerlendirmek amacıyla kullanılır. Matematiksel tanımı Eşitlik 3.5'te verilmiştir.

$$\text{Duyarlılık} = \frac{TP}{TP+FN} \tag{3.5}$$

F1 Skoru: Bu metrik, kesinlik (precision) ve duyarlılığın (recall) harmonik ortalamasını temsil eder. İki ölçüt arasında denge sağlayarak sınıflar arası dengesizlikten etkilenmeyi azaltır. Matematiksel ifadesi Eşitlik 3.6'da sunulmuştur [46], [47].

$$\text{F} - 1 = \frac{2*Kesinlik*Duyarlılık}{Kesinlik+Duyarlılık} \tag{3.6}$$

4. BULGULAR VE TARTIŞMA

Bu bölümde, farklı sosyal medya platformlarından (Facebook, Instagram, X, Reddit) elde edilen Türkçe yorumlar üzerinde uygulanan derin öğrenme tabanlı üç farklı modelin (CNN, LSTM, BERTurk) elde ettiği sınıflandırma sonuçları sunulmaktadır. Her bir model, hem beş kategorili çoklu sınıflandırma (non, ind, grp, oth, prof) hem de ikili sınıflandırma (hakaret içeriyor / içermiyor) senaryolarında değerlendirilmiştir. Modeller, her bir platformun verisiyle ayrı ayrı eğitilip test edilmenin yanı sıra, tüm verilerin birleşimiyle oluşturulan birleşik veri seti üzerinde de eğitilmiştir. Tüm deneylerde veri seti %80 eğitim ve %20 test olarak bölünmüştür.

Her ne kadar doğruluk (accuracy), hassasiyet (precision), duyarlılık (recall) ve F1 skoru gibi birçok metrik analiz edilmiş olsa da tez kapsamında temel değerlendirme ölçütü olarak F1 skoru esas alınmıştır. Çünkü F1 skoru, kesinlik ve duyarlılık arasındaki dengeyi yansıtarak modelin genel performansı hakkında daha güvenilir bir gösterge sunar. Ayrıca, hata matrisleri aracılığıyla her bir modelin hangi sınıflarda ne tür hatalar yaptığı ayrıntılı biçimde analiz edilmiştir. Bu analizler, hem platformların kendine özgü dilsel yapısının model performansına etkisini hem de derin öğrenme mimarilerinin genelleme kapasitesini ortaya koymayı hedeflemektedir. Böylece hem çoklu hem de ikili sınıflandırma senaryolarında, tek bir birleşik modelin platform bazlı modellere göre ne ölçüde başarılı olduğu da karşılaştırmalı olarak değerlendirilmiştir.

4.1. Kategorik Sınıflandırma Sonuçları

Bu bölümde, her bir derin öğrenme modeli için beşli kategori yapısına (non, ind, grp, oth, prof) dayalı olarak gerçekleştirilen çoklu sınıflandırma sonuçları sunulmakta; platform bazlı ve birleşik veri seti üzerinden elde edilen performans metrikleri detaylı biçimde değerlendirilmektedir.

4.1.1. CNN Modeli ile Kategorik Sınıflandırma Sonuçları

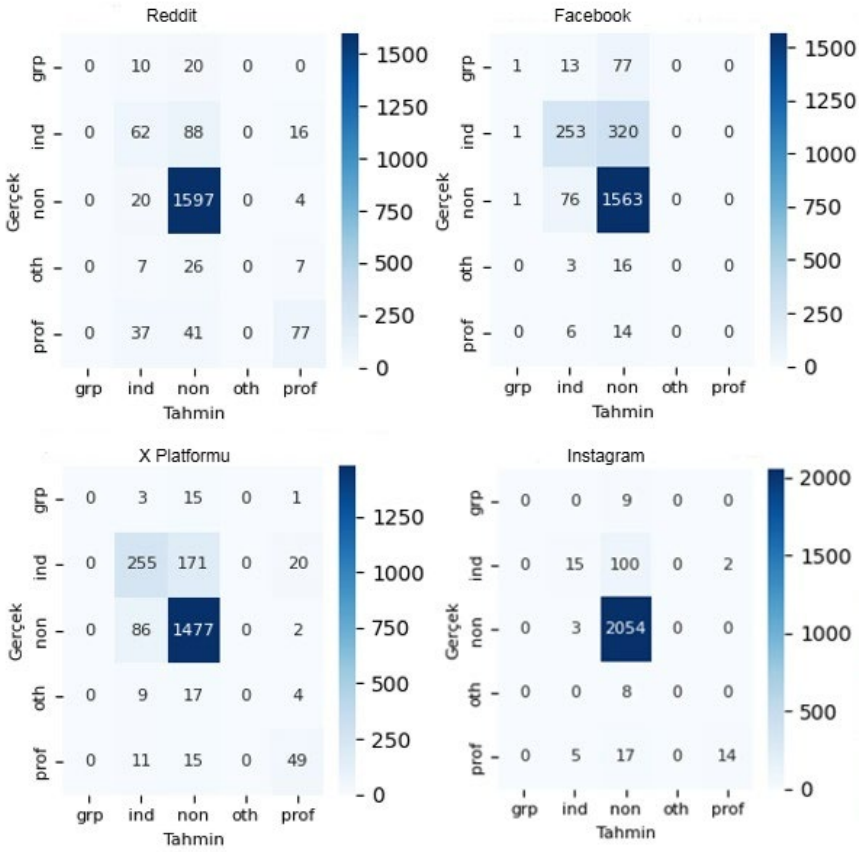
CNN modeli, platform bazlı çok sınıflı (beş kategorili) sınıflandırma görevinde sınırlı bir performans sergilemiştir. En yüksek F1 skoru, Instagram platformunda %91 olarak elde edilirken, Reddit platformu %83 ile onu izlemiştir. X platformunda %81, Facebook'ta ise ancak %73'lük bir F1 skoru sağlanabilmiştir. Bu verilere göre ortalama F1 skoru yaklaşık %82 düzeyinde kalmıştır. CNN mimarisinin çoklu sınıflandırmadaki bu başarımı, yalnızca modelin yapısal özelliklerinden değil, aynı zamanda veri setindeki

sınıf dengesizliklerinden ve her bir sınıfa düşen örnek sayısından kaynaklanmaktadır. Özellikle bazı sınıfların örnek sayılarının düşük olması, modelin bu sınıfları etkili biçimde öğrenmesini zorlaştırarak genel başarıyı düşürmektedir. Bunun yanında, bazı sınıflar içerik bakımından birbirine oldukça benzer ifadeler barındırdığından, etiketleme sürecinde yorumcular arası tutarsızlıklara yol açabilmekte ve modelin sınıflar arasındaki farkları net biçimde öğrenmesini engellemektedir. Sonuç olarak CNN modeli, beşli kategori yapısında platform bazında ortalama %82 F1 skoru ile sınırlı kalmış ve çok sınıflı etikette beklentinin altında bir performans göstermiştir. Çizelge 4.1'de platform bazlı skorlar verilmiştir.

Çizelge 4.1. CNN modeli platform bazlı kategorik sınıflandırma sonuçları

Platform	Doğruluk	Hassasiyet	Duyarlılık	F1-Skor
Reddit	0.86	0.82	0.86	0.83
Facebook	0.77	0.73	0.77	0.73
Instagram	0.93	0.91	0.93	0.91
X Platformu	0.83	0.80	0.83	0.81
Ortalama	0.84	0.81	0.84	0.82

CNN modelinin beşli kategorik yapı üzerindeki sınıflandırma performansı, her bir sosyal medya platformu özelinde incelenmiş ve elde edilen sonuçlar karmaşıklık matrisleri üzerinden değerlendirilmiş, karmaşıklık matrisleri 4.1' de gösterilmiştir



Şekil 4.1. CNN modeli kategorik sınıflandırma karmaşıklık matrisleri

CNN modeli baskın sınıf olan “non” (hakaret içermeyen) kategorisinde yüksek doğruluk gösterdiği görülmektedir. Gerçek “non” etiketli örnekler, Reddit’te 1597, Facebook’ta 1563, X Platformu’nda 1477 ve Instagram’ da 2054 kez doğru tahmin edilmiştir. Bu eğilim, veri setindeki örnek dağılımının “non” sınıfı lehine dengesiz olmasıyla doğrudan ilişkilidir.

Modelin “ind” (bireye yönelik) sınıfına yönelik başarısı platformdan platforma değişmekle birlikte genel olarak sınırlı düzeydedir. Facebook’ta 253, X’te 255, Reddit’te 62 ve Instagram’ da 100 doğru “ind” tahmini yapılmıştır. Bu sınıfa ait birçok örnek model tarafından “non” ya da kısmen “prof” veya “oth” sınıflarıyla karıştırılmıştır. Bu karışıklıklar, dilsel benzerliklerin ve bağlam eksikliklerinin etkisini göstermektedir. Öte yandan “grp” gruba yönelik) ve “oth” (olaya/kuruma yönelik) sınıflarında ise doğru tahmin oranı son derece düşüktür. Facebook’ta “grp” sınıfına ait 13, Reddit’te 10, X Platformu’nda 3 ve Instagram ’da 0 doğru tahmin yapılmıştır. Benzer şekilde, “oth” sınıfında da düşük doğruluk değerleri elde edilmiş, model genellikle bu sınıfları “non” veya “ind” olarak etiketlemiştir. Bu sınıflardaki zayıf başarı, örnek sayısının düşüklüğü ve anlamsal ayrımların yeterince keskin olmamasıyla ilişkilidir. Diğer kategori “prof”

(şaka/nesne) sınıfında ise Reddit'te 77, Facebook'ta 0, X Platformu'nda 49 ve Instagram' da 14 doğru tahmin yapılabilmiştir. Bu sınıfta da doğruluk oranı oldukça sınırlı kalmış, model sıklıkla bu ifadeleri "non" veya "ind" sınıfı altında değerlendirmiştir. Özellikle bağlamdan kopuk ya da imalı söylemler, bu kategorideki karışıklıkların temel nedenlerinden biri olarak değerlendirilmektedir.

Genel olarak bakıldığında CNN modeli, özellikle "non" sınıfında oldukça istikrarlı sonuçlar üretmiş, buna karşılık azınlık sınıflarda düşük başarı göstermiştir. Sınıf dengesizliği, modelin öğrenme sürecini yönlendirmiş; örnek sayısı az olan kategorilerde hem öğrenme hem de genelleme gücü sınırlı kalmıştır. Ayrıca bazı sınıfların semantik olarak birbirine yakın olması, sınıf ayrımını daha da zorlaştırmıştır. Modelin sınıflandırma eğilimleri, her platform özelinde oluşturulan hata matrislerinde ayrıntılı biçimde görülmektedir. Bu bulgular, Şekil 4.1'de sunulmuştur.

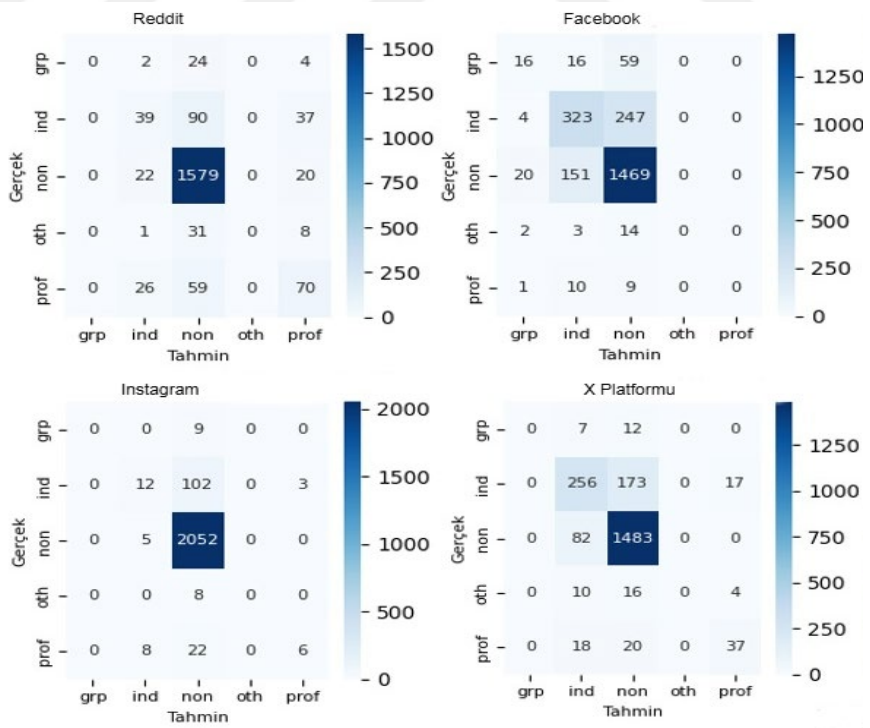
4.1.2. LSTM Modeli ile Kategorik Sınıflandırma Sonuçları

LSTM modeli, platform bazlı çoklu sınıflandırma da CNN'ye kıyasla kısmen iyileşmiş bir performans ortaya koymuştur. En yüksek F1 skoru Instagram platformunda %90 ile elde edilirken, X platformunda bu değer %81 düzeyinde kalmıştır. Reddit'te ise LSTM'nin F1 başarımı %80 olarak gerçekleşmiştir. Facebook'ta ise F1 skoru %75 ile en düşük seviyede gerçekleşmiştir. Bu sonuçlar, LSTM modelinin beş sınıflı etiketleme görevinde platformlar arasında değişken fakat genel olarak yeterli bir performans sunduğunu göstermektedir. Performansın platformlar arasında farklılık göstermesinin başlıca nedenleri arasında, veri setindeki sınıf dengesizlikleri ve azınlık sınıfların yetersiz temsili bulunmaktadır. Bu koşullar, özellikle nadir görülen sınıflarda LSTM modelinin öğrenme kapasitesini sınırlandırmış ve genel başarıyı etkilemiştir. Ayrıca, çok sınıflı yapıdaki kategoriler arasında bulunan anlamsal benzerlikler ve yorum farklılıkları da modelin sınıfları birbirinden ayırmasını zorlaştıran önemli faktörlerdir. Bu zorluklar, özellikle Facebook platformundaki düşük F1 skoru ile net biçimde ortaya çıkmaktadır. Platformlar genelinde LSTM'nin ortalama F1 skoru %81 olarak hesaplanmıştır ki bu, modelin çoklu sınıflandırmada genel başarısının makul düzeyde olduğuna işaret etmektedir. Çizelge 4.2'de platform bazlı skorlar verilmiştir.

Çizelge 4.2. LSTM modeli platform bazlı kategorik sınıflandırma sonuçları

Platform	Doğruluk	Hassasiyet	Duyarlılık	F1-Skor
Reddit	0.83	0.78	0.83	0.80
Facebook	0.77	0.74	0.77	0.75
Instagram	0.92	0.90	0.92	0.90
X Platformu	0.83	0.80	0.83	0.81
Ortalama	0.83	0.80	0.83	0.81

LSTM modelinin dört farklı sosyal medya platformunda gerçekleştirdiği çok sınıflı sınıflandırma performansı, hata (karmaşıklık) matrisleri üzerinden ayrıntılı biçimde analiz edilmiştir. Beşli etiketleme sistemine göre yapılan bu değerlendirmede, her bir platform için sınıf bazlı başarı durumları, sınıflar arası karışmalar ve genel eğilimler sistematik olarak ele alınmıştır. Sunulan matrisler, modelin farklı veri kümeleri üzerinde nasıl çalıştığını karşılaştırmalı biçimde inceleyebilmek açısından önemli bir referans niteliği taşımaktadır. Ayrıca, sınıflar arası ayrımların sayısal temsilleri üzerinden yapılan analizlere temel oluşturmaktadır. Şekil 4.2'de, her bir sosyal medya platformu için oluşturulan LSTM tabanlı karmaşıklık matrisleri ayrıntılı olarak sunulmuştur.



Şekil 4.2. LSTM modeli kategorik sınıflandırma karmaşıklık matrisleri

LSTM modeliyle yürütülen çok sınıflı sınıflandırma görevinde, dört farklı sosyal medya platformunda elde edilen sonuçlar benzer eğilimler göstermiştir. Tüm

platformlarda “non” sınıfı, modelin en başarılı tahmin ettiği kategori olmuştur. Gerçek “non” etiketli örneklerin büyük çoğunluğu doğru şekilde sınıflandırılmıştır; bu eğilim, “non” sınıfının veri setlerindeki baskın yapısından kaynaklanmaktadır. Örneğin, gerçek “non” örneklerinden Reddit’te 1579, Facebook’ta 1523, X Platformu’nda 1483 ve Instagram’ da 2052 adedi doğru tahmin edilmiştir. Modelin “ind” sınıfındaki başarısı ise sınırlı kalmıştır. Facebook’ta 323, X Platformu’nda 256 ve Reddit’te 128 adet gerçek “ind” örneği doğru tahmin edilirken, Instagram’ da bu sınıfa ait hiçbir örnek doğru sınıflandırılamamıştır. Bu sınıfa ait birçok örnek model tarafından “non” veya “oth” sınıfı olarak etiketlenmiştir; ifadelerin anlamsal yakınlığı, modelin “ind” sınıfını ayırt etmesini zorlaştırmıştır. Buna karşılık, “grp”, “oth” ve “prof” sınıflarında doğru tahmin oranı son derece düşüktür. Facebook’ta “grp” sınıfı için 16, diğer platformlarda ise 0’a yakın doğru tahmin yapılmıştır. Olay/kurum sınıfında da her platformda öğrenme yetersizliği gözlemlenmiştir. Öte yandan “prof” sınıfında ise Reddit’te 111, X Platformu’nda 53 ve Instagram’ da 6 örnek doğru tahmin edilirken; Facebook’ta bu sınıfa ilişkin doğru tahmin yapılmamıştır. Bu sınıflardaki düşük başarı, azınlık sınıflara ait örnek sayısının çok az olmasından ve ifadeler arasındaki anlamsal benzerliklerin sınıf ayırımı zorlaştırmasından kaynaklanmaktadır.

Genel olarak değerlendirildiğinde, LSTM modeli dört platformda da benzer sınırlılıklar göstermiştir. Başarı büyük oranda “non” sınıfına odaklanmış; nadir görülen sınıflar ya yeterince öğrenilememiş ya da benzer yapılar nedeniyle birbirine karıştırılmıştır. Bu durum hem veri setindeki dengesiz dağılımdan hem de kategoriler arası anlamsal örtüşmeden kaynaklanmaktadır. Özellikle “grp” ve “oth” sınıfları, tüm platformlarda en zor öğrenilen ve en çok karıştırılan kategoriler olmuştur. Diğer yandan “ind” ve “prof” ise birbirinden ayrıştırılması güç etiketler olarak görünmektedir.

4.1.3. BERTurk Modeli ile Kategorik Sınıflandırma Sonuçları

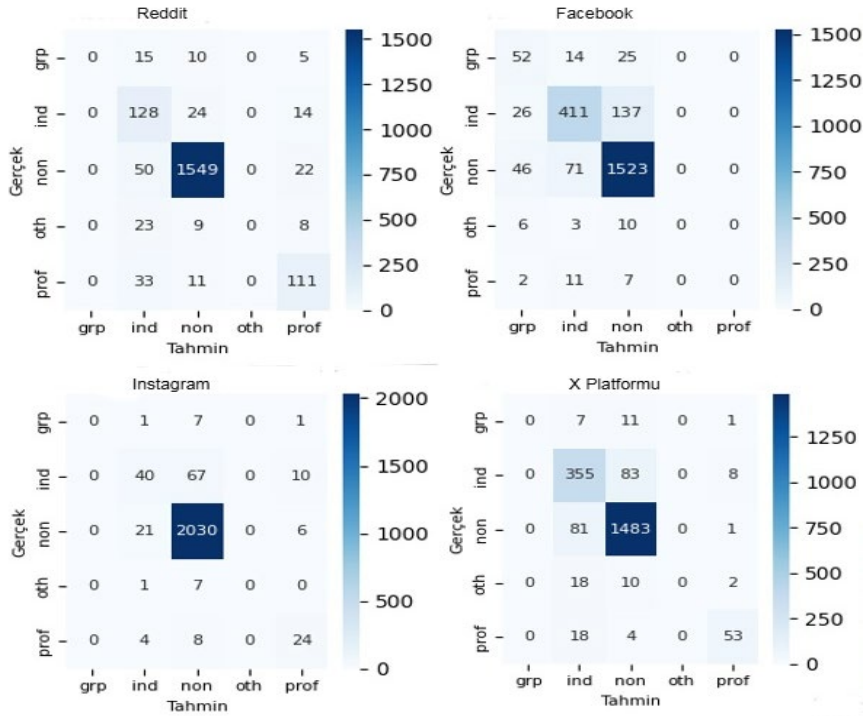
BERTurk modeli, platform bazlı çoklu sınıflandırmada CNN ve LSTM modellerine kıyasla genel olarak daha yüksek F1 skorları elde etmiştir. En iyi performansı Instagram platformunda gösteren BERTurk, Instagram verisinde %93 F1 skoru ile ilk sırada yer almıştır. Reddit ve X platformlarında %87, Facebook’ta ise %84’lük F1 değerleri gözlenmiştir. Platformlar genelinde ortalama F1 skoru da yaklaşık %87 olarak hesaplanmıştır. Bu sonuçlar, BERTurk modelinin bağlamsal ve dilbilgisel özellikleri klasik modellerden daha iyi yakalayarak çok sınıflı hakaret tespitinde üstünlük sağladığını ortaya koymaktadır. Öte yandan, veri setindeki azınlık sınıfların dengesiz

dağılımı ve sınıflar arası anlamsal benzerlik sorunları, BERTurk'un performansını da tavana vurmasından alıkoymuştur. Model her ne kadar genel olarak daha başarılı olsa da bu tür zorluklar nedeniyle elde edilen skorların daha da yükselmesi mümkün olamamıştır. Çizelge 4.3'de platform bazlı skorlar verilmiştir.

Çizelge 4.3. BERTurk modeli platform bazlı kategorik sınıflandırma sonuçları

Platform	Doğruluk	Hassasiyet	Duyarlılık	F1-Skor
Reddit	0.88	0.87	0.88	0.87
Facebook	0.84	0.83	0.84	0.84
Instagram	0.94	0.92	0.94	0.93
X Platformu	0.88	0.86	0.88	0.87
Ortalama	0.88	0.87	0.88	0.87

BERTurk modeli ile yürütülen çok sınıflı sınıflandırma deneylerinde, dört farklı sosyal medya platformunda elde edilen karmaşıklık matrisleri üzerinden modelin sınıf bazlı başarı durumu, sınıflar arası karışmalar ve genel eğilimler değerlendirilmiştir. Bu görsel analizler, Şekil 4.3'te sunulmuştur.



Şekil 4.3. BERTurk modeli kategorik sınıflandırma karmaşıklık matrisleri

BERTurk modeli ile gerçekleştirilen çok sınıflı sınıflandırma deneylerinde, dört sosyal medya platformunda elde edilen sonuçlar, her platforma ait hata matrisleri üzerinden analiz edilmiştir. Modelin, özellikle baskın sınıf olan "non" kategorisinde

yüksek doğruluk oranlarına ulaştığı görülmektedir. Gerçek “non” örnekleri büyük ölçüde doğru tahmin edilmiş olup, Reddit’te 1549, Facebook’ta 1523, X platformunda 1483 ve Instagram’ da 2030 örnek doğru sınıflandırılmıştır. Bu durum, “non” sınıfının veri setlerinde sayıca baskın olmasından kaynaklanmaktadır.

Modelin “ind” sınıfındaki başarısı, anlamlı sayıda doğru tahmin ile desteklenmiştir. Facebook’ta 411, X Platformu’nda 355, Reddit’te 128 ve Instagram’ da 67 adet doğru “ind” tahmini yapılmıştır. Ancak bu sınıf zaman zaman “non” veya “oth” sınıfları ile karıştırılmış; sınıflar arası anlamsal benzerlik nedeniyle sınırlı hata oluşmuştur. Diğer yandan “grp”, “oth” ve “prof” sınıflarında ise sınırlı düzeyde başarı elde edilmiştir. Reddit’te “grp” sınıfı için 10, Facebook’ta 52, X Platformu’nda 7 ve Instagram’ da yalnızca 1 doğru tahmin yapılmıştır. Benzer şekilde, “oth” sınıfı için doğru tahmin sayıları oldukça düşük seviyede kalmış; örneklerin çoğu model tarafından “non” veya “ind” olarak etiketlenmiştir. Söz konusu sınıflar arasında 'prof' etiketi ise Reddit’te 111, Facebook’ta 7, X Platformu’nda 53 ve Instagram’ da 24 doğru tahmin ile temsil edilmiştir. Sınıf dağılımının dengesiz olması, modelin azınlık sınıflara ait örüntüleri öğrenmesini zorlaştırmıştır. Ayrıca bazı ifadelerin birden fazla kategoriye ait olabilecek şekilde yorumlanabilmesi, modelin sınıf ayrımını net biçimde yapamamasına neden olmuştur.

BERTurk modelinin genel sınıflandırma eğilimleri incelendiğinde, özellikle “non” ve kısmen “ind” sınıflarında istikrarlı bir başarı elde ettiği, diğer üç kategoride ise sınırlı da olsa öğrenme işaretleri gözlendiği söylenebilir. Şekil 4.3’te ilgili hata dağılımları ayrıntılı olarak sunulmaktadır.

4.1.4. Tüm Platform Verisi ile Kategorik Sınıflandırma Sonuçları

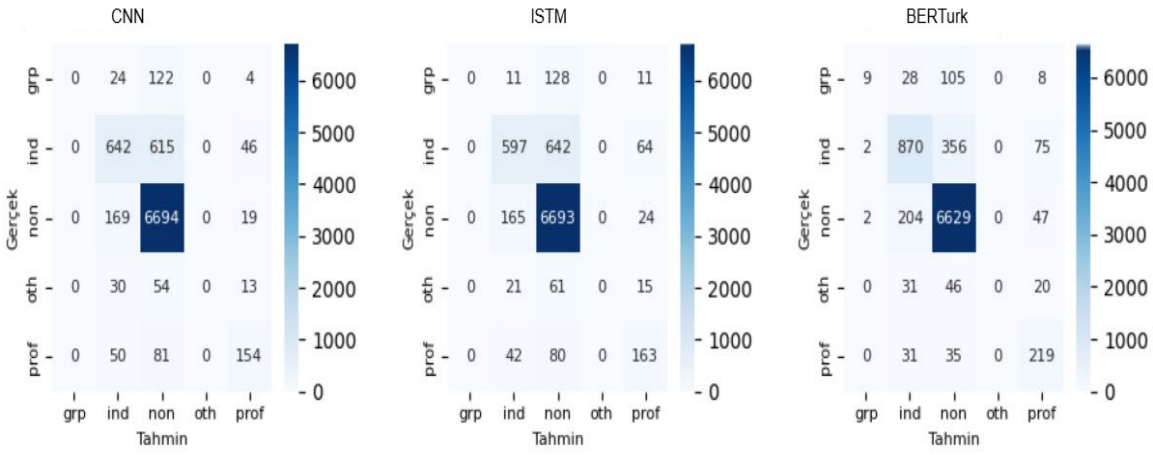
Birleşik (tüm platformları içeren) veri seti üzerinde gerçekleştirilen çok sınıflı hakaret tespiti deneyinde, BERTurk modeli %87 F1 skoru ile en yüksek performansı sergilemiştir. CNN ve LSTM modellerinin F1 başarımları ise sırasıyla %83 ve %83 düzeyinde kalmıştır. Üç modelin F1 skorlarının ortalaması da %84 seviyesinde gerçekleşmiştir. Bu sonuçlar, BERTurk modelinin birleşik veri kümesinde bağlamsal dil özelliklerini diğer modellere kıyasla daha etkili yakalayarak sınıfları birbirinden daha iyi ayırt edebildiğini göstermektedir. Öte yandan, nadir görülen (azınlık) sınıflardaki örnek yetersizliği ve bazı kategoriler arasındaki anlamsal örtüşmeler, tüm modellerin performansını sınırlayan önemli faktörler olarak gözlenmiştir. Bu sebeple, modellerin F1

skorları ancak orta-üst düzeyde kalabilmiş; daha yüksek değerlere ulaşılamamıştır. İlgili tüm F1 skorları Çizelge 4.4'te ayrıntılı olarak sunulmuştur.

Çizelge 4.4. Karma veri kategorik sınıflandırma sonuçları

Derin Öğrenme Modelleri	Doğruluk	Hassasiyet	Duyarlılık	F1-Skor
CNN	0.85	0.82	0.85	0.83
BERTurk	0.88	0.87	0.88	0.87
LSTM	0.85	0.82	0.85	0.83
Ortalama	0.86	0.83	0.86	0.84

CNN, BERTurk ve LSTM modellerinin tüm sosyal medya platformları birleştirilmiş veri seti üzerinden gerçekleştirdiği çok sınıflı sınıflandırma performansı, karmaşıklık matrisleri aracılığıyla incelenmiştir. Şekil 4.4'te CNN, BERTurk ve LSTM modellerine ait sonuçlar görselleştirilmiştir.



Şekil 4.4. Tüm platformlar kategorik karmaşıklık matrisleri

Birleşik veri seti üzerinde gerçekleştirilen çok sınıflı hakaret tespiti deneylerinde, CNN, LSTM ve BERTurk modellerinin sınıflandırma performansları karşılaştırılmıştır. Bu çalışmada CNN modeli elde ettiği makro F1 skoruyla "non" sınıfında en yüksek doğruluğu sergilemiştir; gerçek "non" örneklerinden 6694'ü doğru tahmin edilmiştir. Öte yandan "ind" sınıfında ise 615 doğru tahmin yapılırken, bu sınıfa ait örneklerin önemli bir kısmı "non" ve "prof" sınıflarıyla karıştırılmıştır. Buna karşın "grp", "oth" ve "prof"

kategorilerinde başarı görece daha düşük kalmış, özellikle “grp” sınıfı büyük ölçüde “non” veya “ind” sınıflarına yanlış etiketlenmiştir.

LSTM modeliyle elde edilen birleşik veri sonuçlarında da “non” sınıfı 6629 doğru tahminle en başarılı kategori olarak öne çıkmıştır. Buna karşın “ind” sınıfında 870 doğru tahmine karşılık 356 örnek model tarafından “non” olarak etiketlenmiş; “ind” ifadelerini kısmen öğrenebildiği ancak diğer sınıflarla sınırlarının netleşmediği görülmüştür. Modelin en çok zorlandığı kategoriler arasında “grp” sınıfında 28, “oth” sınıfında 46 ve “prof” sınıfında 219 doğru tahmin yapılmıştır.

BERTurk modelinde de “non” sınıfı 6693 doğru tahminle yüksek başarı göstermiştir. Ayrıca “ind” sınıfında 642 doğru tahmin yapılmış ve dilin bağlamsal yapısı sayesinde anlamlı bir ayrım sağlanabilmiştir. Diğer kategorilerde ise “grp” sınıfı için 24, “oth” sınıfı için 54 ve “prof” sınıfı için 154 doğru tahmin yapılmıştır. BERTurk’ün bağlamı kavrama yeteneği, özellikle “prof” gibi nüans içeren kategorilerde belirgin bir öğrenme eğilimi göstermiştir.

Üç modelin tüm platformları kapsayan verilerle yürüttüğü sınıflandırma görevinde sergilediği performans eğilimleri, Şekil 5.4’te ayrıntılı olarak sunulmuştur. Değerlendirme sonuçları, BERTurk modelinin genel olarak en yüksek makro F1 skorlarına ulaştığını göstermiştir. Platform bazında bakıldığında, X platformunda tüm modeller için en yüksek başarımlar elde edilirken Instagram ve Facebook verilerinde göreli olarak daha düşük sonuçlar kaydedilmiştir. CNN ve LSTM modelleri, baskın sınıf dışındaki azınlık kategorilerde belirgin şekilde zorlanmış; veri setindeki sınıf dengesizliği ve kategoriler arası anlamsal benzerlikler bu iki modelin performansını olumsuz etkilemiştir. BERTurk modeli ise dilin bağlamını anlama konusundaki üstünlüğü sayesinde hem ayrı platformlarda hem de birleşik veri üzerinde daha yüksek doğruluklar elde etmiş, ancak azınlık sınıflardaki karışıklıklardan tamamen muaf olamamıştır. Bu bulgular, Türkçe sosyal medya yorumlarının kategorik hakaret tespitinde BERTurk modelinin en başarılı sonucu verdiğini, sınıf dengesizliği ve kategoriler arası anlamsal örtüşme gibi sorunların giderilmesinin gelecekte genel performansı yükseltebilecek unsurlar olduğunu ortaya koymaktadır.

4.1.5. Kategorik Sınıflandırma Modeller Değerlendirmeleri

Üç derin öğrenme modeli (CNN, LSTM ve BERTurk) ile gerçekleştirilen kategorik hakaret sınıflandırma sonuçları karşılaştırmalı olarak incelendiğinde, BERTurk

modelinin genel olarak en yüksek F1 skorlarına ulaştığı görülmüştür. Özellikle platform bazında bakıldığında, Instagram platformunda BERTurk %93 gibi yüksek bir başarı elde ederken, X platformunda %87, Reddit'te %87 ve Facebook'ta %84 F1 skorları kaydedilmiştir. CNN ve LSTM modellerinde ise en yüksek başarılar yine Instagram'da %91 (CNN) ve %90 (LSTM) olarak gözlenmiş, diğer platformlarda %73 ile %83 arasında sonuçlar elde edilmiştir. CNN ve LSTM modelleri, baskın sınıf dışındaki kategorilerde (örneğin grup veya diğer hakaret türleri gibi az örnekli sınıflarda) belirgin şekilde zorlanmış; veri setindeki sınıf dengesizliği ve kategoriler arası anlamsal benzerlikler bu iki modelin performansını olumsuz yönde etkilemiştir. BERTurk modeli ise dilin bağlamını anlama konusundaki üstünlüğü sayesinde hem ayrı platformlarda hem de birleşik veri üzerinde %87'ye varan F1 skorları ile daha yüksek doğruluklar elde etmiş, ancak onun da azınlık sınıflardaki karışıklıklardan tamamen muaf olmadığı görülmüştür. Nitekim, bütün platformların birleştirildiği kapsamlı veri setiyle modellerin eğitilmesi genel performans seviyelerini artırmış (%84 ortalama F1) ve daha genellenebilir sonuçlar elde edilmesini sağlamıştır. Yine de tüm modeller için az temsil edilen sınıflardaki öğrenme zorlukları F1 skorlarının beklenenden düşük kalmasına yol açmıştır. Sonuç olarak, BERTurk modeli Türkçe sosyal medya yorumlarının kategorik hakaret tespitinde en başarılı sonuçları verirken, sınıf dengesizliği ve sınıflar arası anlamsal örtüşme gibi sorunların giderilmesi gelecekte genel performansı yükseltebilecek unsurlar olarak belirlenmiştir.

4.2. Büyük Dil Modeli Kategorik Sınıflandırma Sonuçları

Bu bölümde, OpenAI tarafından geliştirilen büyük dil modeli GPT'nin farklı örnekleme (prompting) stratejileri altında gerçekleştirdiği çok sınıflı sınıflandırma sonuçları sunulmaktadır. Çalışmada OpenAI API'nin 1.8 sürümü kullanılmış; sıfır örnek (zero-shot), tek örnek (one-shot) ve üç örnekli (three-shot) prompt türleriyle sınıflandırma yetenekleri test edilmiştir. Model herhangi bir yeniden eğitme (fine-tuning) işlemine tabi tutulmamış, yalnızca doğal dil girdileri üzerinden yorumları beşli etiketleme sistemine göre sınıflandırması beklenmiştir. Bu bağlamda, modelin sıfırdan bağlam anlama ve örüntü oluşturma gücü değerlendirilmektedir. Her bir strateji ayrı başlıklar altında sunulmuş, hata matrisleri ve sayısal başarı metrikleri üzerinden analiz gerçekleştirilmiştir.

4.2.1. Sıfır Örnekle (Zero-Shot) Kategorik Sınıflandırma Sonuçları

GPT-4o modeliyle gerçekleştirilen sıfır örnekli sınıflandırma çalışmasında, modele yalnızca sınıf isimleri değil, her bir etiketin kısa açıklayıcı tanımları da prompta dâhil edilmiştir. Bu yapı, modelin yorumları hangi kategoriye dâhil etmesi gerektiğini açıklamalar üzerinden değerlendirmesine imkân tanımıştır. Modelin, herhangi bir örnek görmeden yalnızca kavramsal yönlendirmelerle sınıflandırma yapması, dilsel ayrıştırma yeteneğinin test edilmesini sağlamıştır. Çizelge 4.5'te görüldüğü üzere, en yüksek F1 skoru %66 ile Reddit platformunda elde edilmiştir. Facebook ve X platformları sırasıyla %56 ve %61 oranlarıyla orta düzeyde performans göstermiştir. Instagram verilerinde ise %59 ile görece daha düşük bir başarı kaydedilmiştir. Tüm platformlar dikkate alındığında ortalama F1 skoru %60 olarak hesaplanmıştır. Bu oran, modelin sıfır örnekli koşullarda sınıflar arası ayırt etme yeteneğinin sınırlı olduğunu ve özellikle içerik bakımından örtüşen ya da bağlamdan bağımsız yorumlarda hataya açık hâle geldiğini göstermektedir. Örneğin, bazı ifadelerin hem bireye hem gruba yönelik algılanabilmesi ya da şaka ile hakaret arasındaki anlam farklarının yakın olması, modelin kararlarını güçleştiren başlıca etkenler arasında yer almaktadır.

Çizelge 4.5. Büyük dil modeli sıfır örnekle kategorik sınıflandırma sonuçları

Platform	Doğruluk	Hassasiyet	Duyarlılık	F1-Skor
Reddit	0.56	0.84	0.56	0.66
Facebook	0.47	0.82	0.47	0.56
Instagram	0.45	0.92	0.45	0.59
X Platformu	0.54	0.81	0.54	0.61
Ortalama	0.50	0.84	0.50	0.60

Sıfır örnekli etiketleme sürecinde model, bağlam bilgisi verilmeden sınıflandırma yaptığı için özellikle tepkisel ya da kalıp ifadeleri yanlış kategorilere yönlendirmiştir. Bu duruma ilişkin bazı örnekler aşağıdaki gibidir

Örnek 1

"Yaa bi git."

LLM bu ifadeyi doğrudan kişiye yönelen saldırı olarak görmüş ve ind etiketlemiştir. Oysa ifade kaba bir söylem değil, basit bir gönderme/tepki niteliği taşımaktadır; bağlamda

hakaret unsuru bulunmadığından doğru etiket non olmalıdır. Burada model, sıradan bir tepkisel söylemi yanlış biçimde saldırı kategorisine kaydırmıştır.

Örnek 2

"Allah belanızı versin."

Model bu ifadeyi bir kuruma/olguya yönelmiş saldırı gibi değerlendirmiş ve oth seçmiştir. Ancak aslında ifade, dini bir kalıp üzerinden dile getirilen temennidir; bağlamda hakaret ya da hedef belirleme olmadığı için doğru etiket non'dur. Yanlışlık, "beddua" kalıbının saldırganlıkla karıştırılmasından kaynaklanmaktadır.

Örnek 3

"71 yaşında seçim 3,5 yıl sonra yaş olacak 74 neden bu koltuk sevdası bırakın artık alttan genç dinamik insanlar gelsin sözüm bütün yaşlı siyasetçilere."

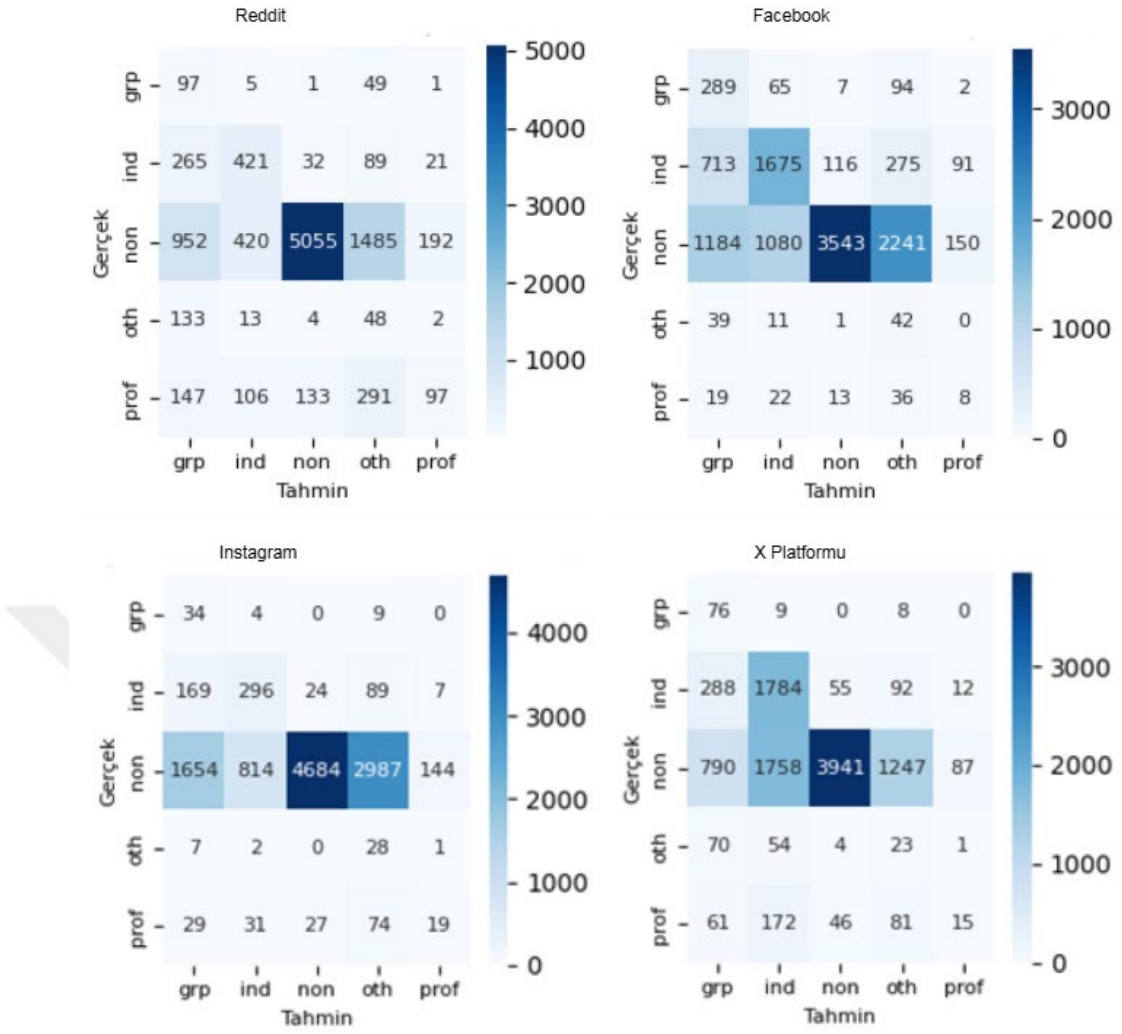
Model, "yaşlı siyasetçiler" ifadesini grup hedeflemesi saymış ve grp kategorisine yerleştirmiştir. Oysa söylem yaşa dayalı bir eleştiri niteliğindedir, doğrudan hakaret barındırmaz. Bu nedenle doğru etiket non olmalıdır. Model, topluluk referansı içeren her ifadeyi saldırı kategorisine alma eğilimi göstermiştir.

Örnek 4

"İyi ki Müslüman değilim."

Model, dini kimliği işaret ettiği için grp etiketini seçmiştir. Oysa bu ifade başkalarını küçümsemez; yalnızca kişinin kendi inancı üzerine bir beyanıdır. Dolayısıyla saldırganlık unsuru yoktur, doğru etiket non'dur. Burada modelin hata nedeni, din temalı her ifadenin saldırı olarak yorumlanmasıdır.

GPT-4o modeliyle yürütülen sıfır örnekli çok sınıflı sınıflandırma çalışmaları, modelin sınıflar arası ayırt etme başarısını daha ayrıntılı değerlendirebilmek amacıyla platform bazlı hata matrisleri üzerinden analiz edilmiştir. Her bir platform için elde edilen sonuçlar, Şekil 4.5'te görsel olarak sunulmuş ve modelin hangi sınıflarda tutarlı, hangilerinde hataya açık tahminler yaptığı somut biçimde ortaya konmuştur.



Şekil 4.5. Sıfır örnekle kategorik sınıflandırma karmaşıklık matrisleri

Sıfır örnekli sınıflandırma sürecinde, GPT-4o modelinin hata matrisleri incelendiğinde tüm platformlarda en yüksek doğru sınıflandırmanın "non" sınıfında gerçekleştiği gözlemlenmiştir. Reddit'te 5055, Instagram' da 4684, X platformunda 3941 ve Facebook'ta 3543 "non" etiketi doğru biçimde tahmin edilmiştir. Bu durum, modelin veride baskın olan sınıfları daha kolay ayırt edebildiğini göstermektedir. Buna karşın "ind" ve "prof" gibi içeriksel açıdan daha yorum gerektiren sınıflarda karışıklık oldukça yüksektir. Örneğin, Instagram' da 296 "ind" etiketi doğru tahmin edilmesine rağmen 244'ü "non" olarak sınıflandırılmıştır. Benzer biçimde, Facebook'ta 275 "ind" örneği yanlışlıkla "oth" olarak etiketlenmiştir. Diğer yandan "prof" sınıfında da sıklıkla "non" ve "oth" ile karıştırmalar yaşanmıştır; örneğin Reddit'te bu sınıfa ait 291 yorum "oth", 133'ü ise "non" olarak tahmin edilmiştir. Ayrıca "grp" ve "oth" kategorilerinde modelin performansı genellikle zayıf kalmıştır. Bu sınıflarda doğru sınıflandırma sayıları oldukça

düşük olup çoğu zaman “non” veya “ind” sınıfı ile karıştırılmıştır. Örneğin, Reddit’te “grp” sınıfındaki 97 yorum doğru tahmin edilirken 49’u “oth”, 5’i “ind”, 952’si ise “non” olarak etiketlenmiştir.

Bu sonuçlar, modelin yalnızca etiket açıklamalarına dayalı olarak sınıflar arasında belirgin farklar kurmakta zorlandığını göstermektedir. Özellikle sınıflar arasında anlam geçişliliği olan içeriklerde modelin ayrım gücünün zayıfladığı görülmektedir. Hata matrislerinde gözlemlenen karışıklıklar, modelin bazı içerikleri çoklu kategorilere yakın görerek kesin bir sınıfa atama konusunda kararsız kaldığını göstermektedir.

4.2.2. Tek Örnekle (One-Shot) Kategorik Sınıflandırma Sonuçları

GPT-4o modeliyle gerçekleştirilen tek örnekli sınıflandırma deneylerinde, modele her kategori için açıklayıcı tanımların yanı sıra her sınıfa ait birer örnek de sunulmuştur. Bu yaklaşım, modelin kavramsal ayrımları daha net biçimde öğrenmesini ve yorumları sınıflandırırken örnek temelli karşılaştırma yapabilmesini amaçlamaktadır. Çizelge 4.6’da sunulan F1 skorlarına göre en yüksek başarı %73 ile Reddit platformunda elde edilmiştir. Instagram %71, X platformu %67 ve Facebook %65 ile onu takip etmiştir. Tüm platformlar genelinde ortalama F1 skoru %69 olarak hesaplanmıştır. Zero-shot senaryoya kıyasla genel bir iyileşme olduğu gözlemlenmiş, özellikle azınlık sınıflarda yapılan doğru tahminlerin sayısında artış kaydedilmiştir. Bu artış, tek bir örnek sunmanın modelin karar süreçlerini yönlendirmede etkili olduğunu ortaya koymaktadır. Hata matrisleriyle birlikte ayrıntılı biçimde değerlendirilmektedir.

Çizelge 4.6. Büyük dil modeli tek örnekle kategorik sınıflandırma sonuçları

Platform	Doğruluk	Hassasiyet	Duyarlılık	F1-Skor
Reddit	0.65	0.87	0.65	0.73
Facebook	0.56	0.83	0.56	0.65
Instagram	0.58	0.94	0.58	0.71
X Platformu	0.59	0.84	0.59	0.67
Ortalama	0.59	0.87	0.59	0.69

Tek örnekli etiketlemede modele her sınıftan bir örnek verilmesine rağmen, bazı ifadelerde yanlış genellemeler gözlemlenmiştir. Bu duruma ilişkin bazı örnekler aşağıdaki gibidir.

Örnek 1

"Tahtası eksik falan ama hiç tahtası olmayan bidondan iyidir."

Model bu ifadeyi hedefsiz küfür kategorisine (prof) almıştır. Oysa burada belirli bir kişiye yönelik "tahtası eksik" gibi doğrudan hakaret söz konusudur. Dolayısıyla doğru etiket ind olmalıdır. Hata, hedefin kişisel olduğunun gözden kaçırılmasından doğmaktadır.

Örnek 2

"Seni istemiyoruz. Açlık, işsizlik, sefalet, adaletsizlikten başka bir şey getirmedin."

Model, bu ifadeyi kişisel hakaret gibi algılayıp ind seçmiştir. Oysa içerikte bir aşağılama yoktur; siyasal/ekonomik bağlamda eleştiri niteliğindedir. Dolayısıyla doğru etiket non'dur. Yanlışlık, sert eleştiriyi hakaretle karıştırmasından kaynaklanmıştır.

Örnek 3

"Sen bu ücrete gel çalış bakalım yetişebiliyor musun."

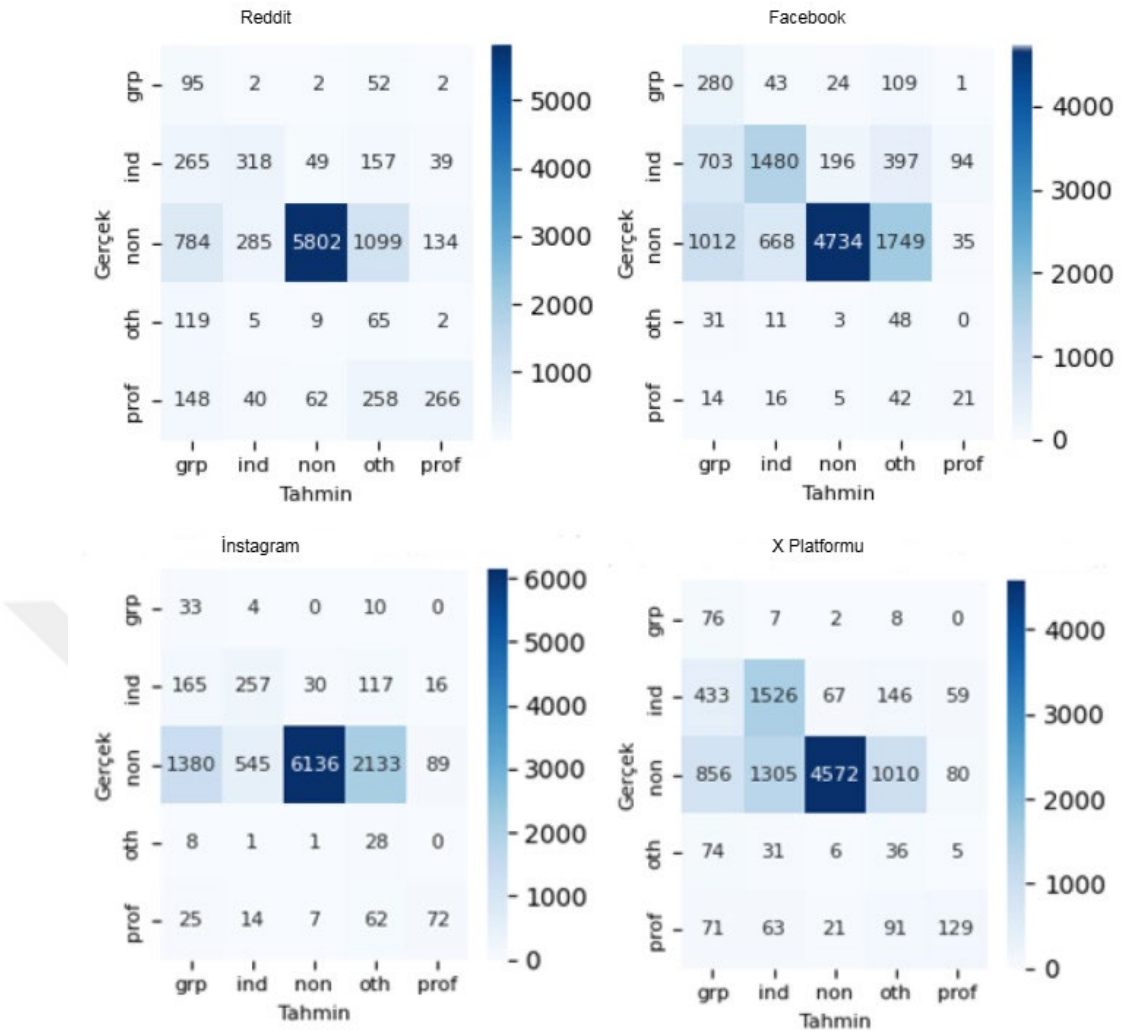
Burada model, doğrudan muhatap olduğu için ind seçmiştir. Oysa bu cümle yalnızca bir önerme/itiraz içermektedir, hakaret barındırmaz. Dolayısıyla doğru etiket non olmalıdır. Model, ikinci şahıs kullanımını saldırı olarak görme eğilimi göstermektedir.

Örnek 4

"Türk milletinin sonunu getirmeyin de."

Model, "Türk milleti" ifadesini grup referansı sayarak grp etiketini seçmiştir. Oysa bu ifade bir hakaret değil, kaygı bildiren bir uyarıdır. Doğru etiket non'dur. Buradaki hata, ulusal kimlik içeren ifadelerin yanlış biçimde saldırı kategorisine kaydırılmasıdır.

GPT-4o modelinin one-shot yaklaşımıyla gerçekleştirdiği sınıflandırmalara ilişkin hata matrisleri, her platform özelinde elde edilen tahmin performansını görsel olarak ortaya koymaktadır. İlgili matrisler Şekil 4.6'da sunulmuştur.



Şekil 4.6. Tek örnekle kategorik sınıflandırma karmaşıklık matrisleri

One-shot sınıflandırmada GPT-4o modeline her bir kategori için yalnızca bir örnek sunulmuş; böylece modelin hem tanım hem de örnek üzerinden karar verme kapasitesi değerlendirilmiştir. Hata matrislerine bakıldığında, tüm platformlarda modelin "non" sınıfında oldukça yüksek doğrulukla çalıştığı görülmektedir. Reddit'te 5802, Facebook'ta 4734, Instagram' da 6136 ve X platformunda 4572 adet "non" etiketi doğru biçimde tahmin edilmiştir. Bu sonuç, modelin çoğunluk sınıfı daha net kavrayabildiğini ve örnek destekli bağlamlarda güvenli tahminler üretebildiğini göstermektedir. Ancak "ind" sınıfında da önceki senaryoya göre anlamlı bir gelişme kaydedildiği görülmektedir. X platformunda 1526, Facebook'ta 1480, Reddit'te 318 ve Instagram' da 257 adet "ind" örneği doğru sınıflandırılmıştır. Bu ilerleme, örnek destekli yönlendirmenin modelin karar verme kabiliyetini güçlendirdiğini göstermektedir.

Buna karşın “grp”, “oth” ve “prof” sınıflarında başarı hâlâ düşüktür. Reddit’te “prof” sınıfında 266 doğru tahmine karşılık 258 “oth” ve 133 “non” ile karıştırma yapılmıştır. Benzer şekilde, Facebook’ta “oth” sınıfında yalnızca 48 doğru tahmin yapılırken, bu sınıfa ait yorumların büyük kısmı “non” ve “ind” olarak yanlış sınıflanmıştır. Instagram’ da ise “prof” sınıfına ait 72 doğru tahmin bulunmasına rağmen 62 örnek “oth” ve 74’ü “non” olarak tahmin edilmiştir. Çoğu yorumun ‘grp’ ve ‘oth’ sınıflarında örtüşen ifadelere sahip olması, bu kategorilerin model için halen belirsiz alanlar olduğunu göstermektedir.

Sonuç olarak, one-shot yönlendirme ile modelin bazı kategorilerde belirgin şekilde daha iyi sonuçlar verdiği, özellikle “ind” sınıfında anlamlı bir ilerleme sağlandığı görülmektedir. Ancak azınlık sınıflarda hâlâ karışıklıklar devam etmekte ve bazı içeriklerin birden fazla sınıfla ilişkilendirilebilir olması modelin kesin sınıflandırma yapmasını zorlaştırmaktadır. Bu bulgular, Şekil 4.6’da görsel olarak sunulmuştur.

4.2.3. Üç Örnekle (Three-Shot) Kategorik Sınıflandırma Sonuçları

Bu bölümde, GPT-4o modeline her bir kategori için üç örnek gösterilerek yürütülen üç örnekli sınıflandırma deneylerinin sonuçları sunulmaktadır. Bu strateji, modele sadece kategori tanımları değil, aynı zamanda her sınıfa ait çok sayıda bağlamsal örnek de sağlayarak, daha güçlü bir yönlendirme oluşturmayı hedeflemiştir. Üç örnekli sınıflandırma stratejisinde GPT-4o modelinin gösterdiği performansa ilişkin F1 skorları aşağıda Çizelge 4.7’de sunulmuştur. F1 skorlarına göre en yüksek başarı %70 ile Reddit platformunda kaydedilirken, X platformu %62 ile onu takip etmiştir. Facebook %58 ve Instagram ise %63’lük skorlarla daha düşük seviyede kalmıştır. Tüm platformların ortalaması alındığında elde edilen %63’lük F1 skoru, modelin özellikle baskın sınıfı doğru tahmin etme eğilimini sürdürdüğünü, ancak azınlık sınıflarda hâlâ ayırt edici bir başarı düzeyine ulaşamadığını göstermektedir.

Çizelge 4.7. Büyük dil modeli üç örnekle kategorik sınıflandırma sonuçları

Platform	Doğruluk	Hassasiyet	Duyarlılık	F1-Skor
Reddit	0.61	0.85	0.61	0.70
Facebook	0.48	0.81	0.48	0.58
Instagram	0.49	0.93	0.49	0.63
X Platformu	0.54	0.82	0.54	0.62
Ortalama	0.53	0.85	0.53	0.63

Üç örnekli senaryoda modele daha fazla bağlamsal bilgi verilmiş olsa da bir örnekli modelden daha fazla yaptığı görülmüştür. Bu duruma ilişkin bazı örnekler aşağıdaki gibidir.

Örnek 1

"Yeter artık düş şu milletin yakasından."

Model, "milletin yakasından" ifadesini kurumsal/olguya saldırı şeklinde yorumlayarak oth etiketini seçmiştir. Oysa bu ifade bir tepki ya da şikâyettir, hakaret yoktur. Dolayısıyla doğru etiket non olmalıdır.

Örnek 2

"Sen bi daha varsayımda bulunma Ebubekir."

Model, doğrudan isim verilmesi ve emredici ton nedeniyle ind seçmiştir. Oysa ifade kişisel saldırı içermez; yalnızca uyarı niteliğindedir. Dolayısıyla doğru etiket non'dur. Model, özel isim + ikinci şahıs kipini çoğu zaman saldırı olarak algılamaktadır.

Örnek 3

"Azınla kuş tutsa artık bitti kimse güvenmiyor. Herkes yaka silkeliyor beddua ediyor durum bu."

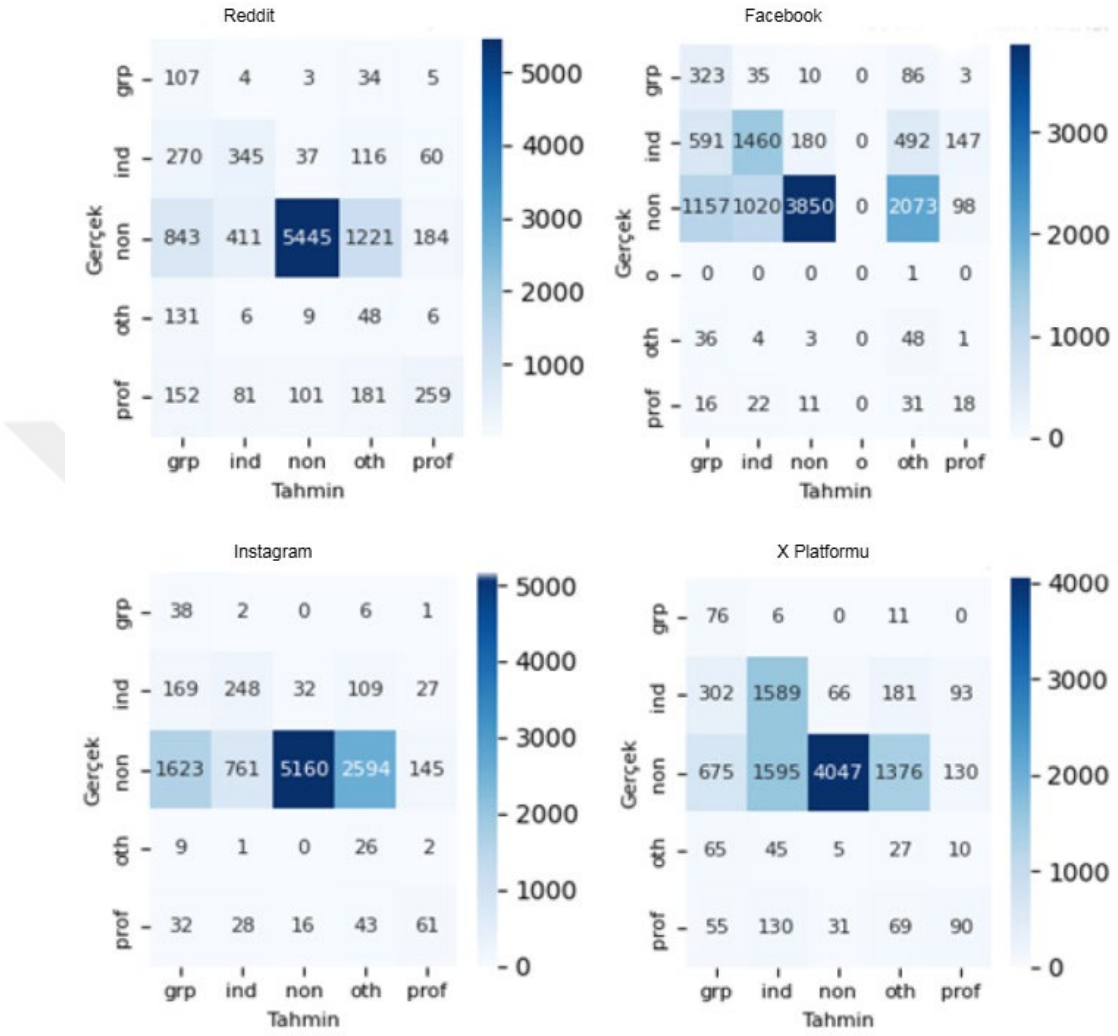
Model bu ifadeyi kuruma/olguya yönelik saldırı olarak görmüş ve oth seçmiştir. Oysa söylem yalnızca güvensizlik ve hayal kırıklığı bildirmektedir. Hakaret içermediğinden doğru etiket non'dur. Yanlışlık, olumsuz duygu yüklü cümlelerin saldırı gibi algılanmasından kaynaklanmıştır.

Örnek 4

"Son dakika alkışlayın arkadaşlar. Ulan dünyaya zulüm eden bir itin övgüsünün haberini mi yapıyorsunuz."

Burada model, söylemi genel bir olguya yönelmiş saldırı gibi yorumlamış ve oth etiketini vermiştir. Oysa cümlede "bir it" ifadesi doğrudan hakaret içermektedir ve belirli bir hedefi işaret etmektedir. Dolayısıyla doğru etiket ind'dir. Hata, kişisel hakaretin bağlamdaki öfke söylemiyle gölgelenmesinden kaynaklanmaktadır.

GPT-4o modelinin three-shot yaklaşımıyla gerçekleştirdiği sınıflandırmalara ilişkin hata matrisleri, her platform özelinde elde edilen tahmin performansını görsel olarak ortaya koymaktadır. İlgili matrisler Şekil 4.7'de sunulmuştur.



Şekil 4.7. Üç örnekle kategorik sınıflandırma karmaşıklık matrisleri

Üç örnekli karmaşıklık matrisleri incelendiğinde, "non" sınıfında yine en yüksek doğruluk oranları elde edilmiştir. Instagram' da 5160, Reddit'te 5445, X platformunda 4047 ve Facebook'ta 3850 örnek doğru biçimde sınıflandırılmıştır. Bu sonuçlar, üç örnekle sağlanan bağlamsal desteğin özellikle baskın sınıflarda modelin güvenini artırdığını göstermektedir. Önceki senaryolarla karşılaştırıldığında "ind" sınıfı belirgin bir gelişme kaydedilmiştir. X platformunda 1589, Facebook'ta 1460, Reddit'te 345 ve Instagram' da 248 örnek doğru sınıflandırılmıştır. Ancak hâlen bu sınıf "non", "oth" ve "prof" ile karıştırılabilmektedir; örneğin Reddit'te 270 "ind" örneği "grp" olarak tahmin edilmiştir. Özellikle "ind" ve "prof" sınıflarındaki anlam farklarının bağlama bağlı

biçimde anlaşılması gerektiği için sınırlı düzeyde hata devam etmektedir. Buna karşın "grp", "oth" ve "prof" sınıflarında modelin genel başarımı hâlâ düşüktür. Reddit'te "prof" sınıfında 259 doğru tahmin yapılırken, 181 örnek "oth" ve 133'ü "non" olarak sınıflandırılmıştır. Facebook'ta "grp" sınıfına ait 323 örnek doğru tahmin edilirken, 86'sı "oth" ile karıştırılmıştır. Instagram' da ise "grp" ve "oth" gibi daha az temsil edilen sınıflar için doğruluk oldukça sınırlı kalmıştır. Bu sınıflarda hem veri yoğunluğu düşük hem de dilsel sınırlar daha esnek olduğu için modelin karar verme süreci belirsizleşmektedir.

Sonuçlar genel olarak değerlendirildiğinde, üç örnekle yapılan yönlendirmenin modelin karar doğruluğunu olumlu yönde etkilediği görülmektedir. Özellikle "non" ve "ind" sınıflarında anlamlı artışlar elde edilmiş; ancak örtüşen veya nüans içeren kategorilerde modelin yanılma eğilimi sürmüştür. Şekil 4.7'de bu sınıflandırma sonuçlarına ait hata matrisleri görsel olarak sunulmuştur.

4.2.4. LLM Kategorik Sınıflandırma Sonuçlarının Değerlendirilmesi

GPT-4o modeli aracılığıyla yürütülen sıfır ve az örnekli etiketleme çalışmaları, beşli kategorik sınıflandırma bağlamında genel eğilimler açısından benzerlik göstermektedir. Zero-shot çalışmasında modele yalnızca kategori tanımları sunulmuş; one-shot ve three-shot çalışmalarında ise her kategoriye ait bir veya üç örnek cümle ile bağlamsal destek sağlanmıştır. Üç etiketleme türünde de modelin özellikle "non" sınıfını doğru biçimde tanımlayabildiği, buna karşılık "grp", "oth" ve "prof" gibi sınıflarda belirgin karışıklıklar yaşandığı gözlemlenmiştir. Bu durum, kategoriler arası anlamsal sınırların yakınlığı ile birlikte, bazı sınıfların içerik açısından birbiriyle örtüşmesinden kaynaklanmaktadır.

Zero-shot uygulamasında genel başarı sınırlı kalmış, dört platformun ortalama F1 skoru %60 olarak ölçülmüştür. One-shot senaryosunda bu oran %69'a yükselmiş ve modelin bağlamsal örneklerle daha sağlıklı bir sınıflama yapabildiği görülmüştür. Three-shot senaryosunda ise F1 skoru %63 olarak gerçekleşmiştir. Bu sonuçlar, modele daha fazla örnek sunulmasının her zaman daha yüksek doğruluk sağlamadığını, hatta sınıflar arası örneklerin çeşitlenmesinin karar ağacında belirsizlik yaratabileceğini göstermektedir. One-shot ve three-shot sonuçlarının birbirine yakın çıkması, modelin sınıflar arasındaki ayrımları tam anlamıyla kavrayamadığını ve dilsel benzerliklerin karar sürecini zorlaştırdığını ortaya koymaktadır.

Öte yandan, manuel olarak gerçekleştirilen etiket kontrol sürecinde büyük dil modelinin bazı yapısal eğilimleri dikkat çekmiştir. Özellikle eleştirel, iğneleyici veya alaycı ton taşıyan yorumlar, açık bir hakaret içermese de GPT-4o tarafından sıklıkla "hakaret" olarak yorumlanmış ve çoğunlukla "ind" ya da "grp" kategorilerine atanmıştır. Bu durum, modelin biçimsel sertliği anlam düzeyinde saldırganlıkla eşleştirme eğiliminde olduğunu ve bağlamı yeterince ayırt edemediğini göstermektedir.

Genel olarak değerlendirildiğinde, GPT-4o modelinin az sayıda örnekle dahi sınıflandırma eğilimini kısmen geliştirebildiği, ancak özellikle Türkçe dilinde ve çok sınıflı yapılarda ayrım gücünün hâlâ sınırlı olduğu görülmektedir. Sınıf içi dengesizlik, semantik benzerlik ve bazı kategori sınırlarının net olmaması, modelin karar süreçlerini zorlaştırmakta; bu da özellikle "grp", "oth" ve "prof" sınıflarında yüksek hata oranlarına yol açmaktadır. Yine de one-shot ve three-shot senaryolarında gözlemlenen kısmi performans artışı, büyük dil modellerinin sınırlı bağlamsal veriyle daha kararlı sonuçlar üretebildiğini göstermektedir.

4.3. Derin Öğrenme Modelleri ile Büyük Dil Modelinin Karşılaştırması

Bu bölümde, üç farklı derin öğrenme modeli (CNN, LSTM ve BERTurk) ile büyük dil modeli (GPT-4o) arasında, çok sınıflı hakaret tespiti görevinde gözlemlenen performans eğilimleri karşılaştırmalı olarak ele alınmıştır. Her iki yaklaşım farklı mimari temellere dayansa da model çıktıları arasında anlamlı benzerlikler ve ayrışmalar gözlenmiştir.

Derin öğrenme modelleri, her biri eğitim verisiyle optimize edilmiş ve hem platform bazlı hem de birleşik veri setleriyle değerlendirilmiştir. Bu modeller arasında BERTurk, genellikle en yüksek başarıyı gösterirken; CNN ve LSTM, özellikle dengesiz sınıf dağılımlarında daha fazla hata üretmiştir. Örneğin, birleşik veri seti üzerinde BERTurk modeli %87'ye kadar ulaşan F1 skorları üretmiştir. Buna karşın, CNN ve LSTM modellerinde bu değer %83 düzeyinde kalmıştır.

Öte yandan, GPT-4o modeli sıfır ve az örnekli yönlendirme senaryolarıyla (zero-shot, one-shot, three-shot) değerlendirilmiş; bu bağlamda eğitim süreci geçirmeden doğrudan etiketleme görevine tabi tutulmuştur. Bu yaklaşımda en yüksek F1 skoru %69 ile one-shot senaryosunda elde edilmiş, three-shot denemede bu değer %63'e gerilemiştir. Zero-shot senaryosunda ise başarı %60 olarak ölçülmüştür. Bu değerler, derin öğrenme

modellerine göre genel olarak daha düşüktür; ancak GPT-4o'nun herhangi bir eğitim süreci olmaksızın yalnızca yönerge ve örneklerle bu sonuçlara ulaşması dikkate değerdir.

Tüm modellerde ortak bir zorluk olarak, "grp", "oth" ve "prof" sınıflarında düşük başarı oranları dikkat çekmektedir. Bu sınıflar arasında semantik örtüşmelerin bulunması ve sınıf içi örnek sayılarının azlığı hem geleneksel derin öğrenme hem de büyük dil modeli mimarileri için önemli bir sınırlayıcı unsur olmuştur. Özellikle GPT-4o'nun, açıkça hakaret içermeyen ama eleştirel ya da alaycı ifadeleri "ind" veya "grp" olarak sınıflandırması, yorum düzeyinde bağlamı tam olarak kavrayamadığını ortaya koymaktadır. LSTM ve CNN gibi modellerde ise bu sınıflar genellikle "non" etiketiyle karıştırılmıştır.

Sonuç olarak, derin öğrenme modelleri veriyle öğrenme gücü sayesinde genel olarak daha yüksek doğruluk sağlamış; GPT-4o ise dilsel modelleme kapasitesiyle, özellikle "non" ve "ind" gibi yaygın sınıflarda tatmin edici düzeyde performans sergilemiştir. Ancak çok sınıflı yapılarda yüksek doğruluk ve ayrım gücü elde edebilmek için, veriyle doğrudan öğrenen ve her sınıfın dilsel özelliklerini kapsamlı biçimde modelleyebilen derin öğrenme mimarileri hâlen daha avantajlıdır.

4.4. İkili Sınıflandırma Sonuçları

Bu bölümde, sosyal medya platformlarından toplanan Türkçe yorumların "hakaret içeriyor" veya "hakaret içermiyor" olarak iki sınıfa ayrıldığı ikili sınıflandırma deneylerinin sonuçları incelenmektedir. Performans değerlendirmelerinde doğruluk, hassasiyet, duyarlılık ve F1 skoru gibi temel metrikler birlikte kullanılmıştır. Bunun sebebi, veri setinde sınıflar arasında ciddi dengesizliklerin bulunması ve modellerin hem genel doğruluk hem de sınıflar arası ayırt etme gücünün kapsamlı biçimde değerlendirilmek istenmesidir. F1 skoru ise doğruluk ve duyarlılığı birlikte ele alarak modellerin genel performansını daha dengeli şekilde yansıtmaktadır. CNN, LSTM ve BERTurk modelleri hem platform bazlı hem de birleşik veri seti üzerinde test edilerek karşılaştırılmış; bu sayede Türkçe sosyal medya metinlerinde hakaret tespiti için en uygun ikili sınıflandırma yaklaşımları kapsamlı biçimde ortaya konmuştur.

4.4.1. CNN Modeli ile İkili Sınıflandırma Sonuçları

CNN modeli, platform bazlı ikili sınıflandırma görevinde genel olarak başarılı bir performans sergilemiştir. En yüksek F1 skoru, %92 ile Instagram platformunda elde edilmiş; bu platformu %85'lik skorlarla Reddit ve X Platformu takip etmiştir. Buna

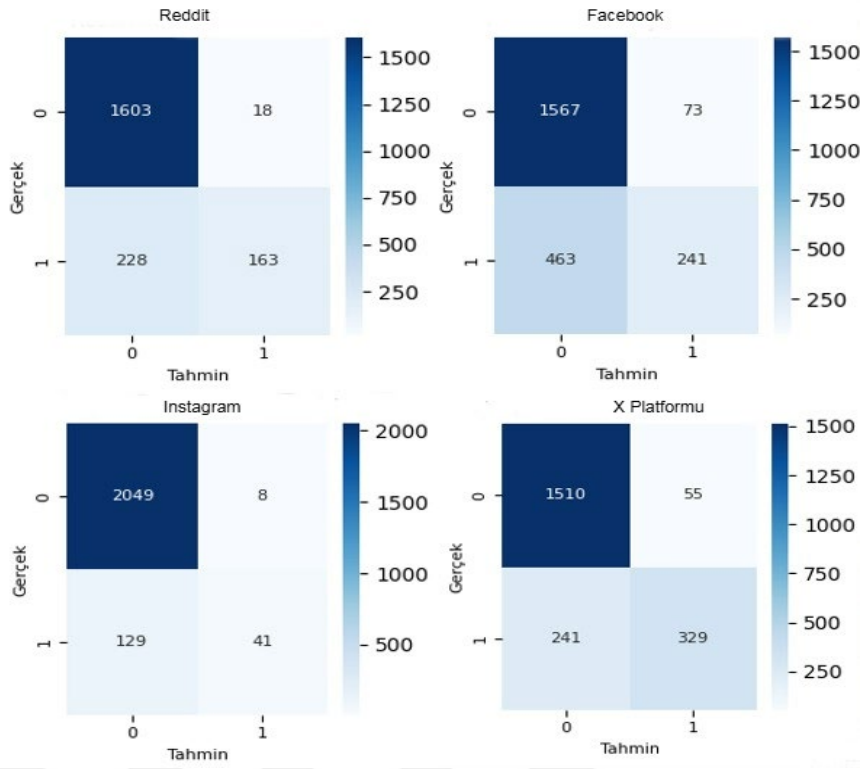
karşılık, Facebook verileriyle yapılan sınıflandırmada modelin F1 skoru %73'te kalmış ve bu platform en düşük başarı oranına sahip olmuştur.

Genel ortalamalara bakıldığında, CNN modelinin dört platformda da ortalama %83 F1 skoru elde ettiği görülmektedir. Bu sonuçlar, CNN modelinin farklı sosyal medya platformlarında Türkçe içerikli hakaret tespitinde iki sınıflı ayrımda oldukça etkili bir yöntem olabileceğini ortaya koymaktadır. Ancak, Facebook'taki görece düşük performans, bu platformdaki dil yapısının, kullanıcı davranışlarının ve muhtemelen sınıf dengesizliğinin model başarımını olumsuz etkilediğine işaret etmektedir. Çizelge 4.8'de, CNN modelinin her platformda elde ettiği doğruluk, hassasiyet, duyarlılık ve F1-Skor değerleri ayrıntılı şekilde sunulmuştur.

Çizelge 4.8. CNN modeli platform bazlı ikili sınıflandırma sonuçları

Platform	Doğruluk	Hassasiyet	Duyarlılık	F1-Skor
Reddit	0.87	0.88	0.87	0.85
Facebook	0.77	0.77	0.77	0.73
Instagram	0.93	0.93	0.93	0.92
X Platformu	0.86	0.86	0.86	0.85
Ortalama	0.85	0.86	0.85	0.83

CNN modeli ile yürütülen ikili sınıflandırmada, dört sosyal medya platformuna ait yorumlar "0" (hakaret içermeyen) ve "1" (hakaret içeren) olarak iki gruba ayrılmış ve modelin sınıflandırma performansı hata matrisleri üzerinden analiz edilmiştir. Elde edilen sonuçlar Şekil 4.8'de görsel olarak sunulmuştur.



Şekil 4.8. CNN modeli ikili sınıflandırma karmaşıklık matrisleri

CNN modeli ile yürütülen ikili sınıflandırma görevi, dört sosyal medya platformunda ("hakaret içeren" / "içermeyen") değerlendirilmiştir. Model, genel olarak "0" (hakaret içermeyen) sınıfında yüksek doğruluk sergilemiş; Reddit'te 1603, Facebook'ta 1567, Instagram' da 2049 ve X Platformu'nda 1510 örnek doğru sınıflandırılmıştır. Bu sınıfa ait yanlış tahmin sayıları oldukça düşüktür. Öte yandan "1" (hakaret içeren) sınıfındaki başarı daha sınırlı kalmıştır. Facebook ve Reddit'te "1" sınıfı büyük oranda "0" olarak etiketlenmiş; sırasıyla 463 ve 228 yanlış tahmin yapılmıştır. Instagram' da bu sayı 129, X Platformu'nda 241 olarak kaydedilmiştir. Doğru "1" sınıflandırmaları ise Facebook'ta 241, Reddit'te 163, Instagram' da 41 ve X Platformu'nda 329 şeklindedir. Bu sonuçlar, modelin "0" sınıfına karşı yüksek bir hassasiyete sahip olduğunu; ancak azınlık konumundaki "1" sınıfını yeterince öğrenemediğini göstermektedir. Veri dengesizliği ve bağlamsal belirsizlik, özellikle "1" sınıfındaki hataların temel nedenleri arasında öne çıkmaktadır. Şekil 4.8'de sunulan hata matrisleri, bu sınıflandırma eğilimlerini görsel olarak desteklemektedir.

4.4.2. LSTM Modeli ile İkili Sınıflandırma Sonuçları

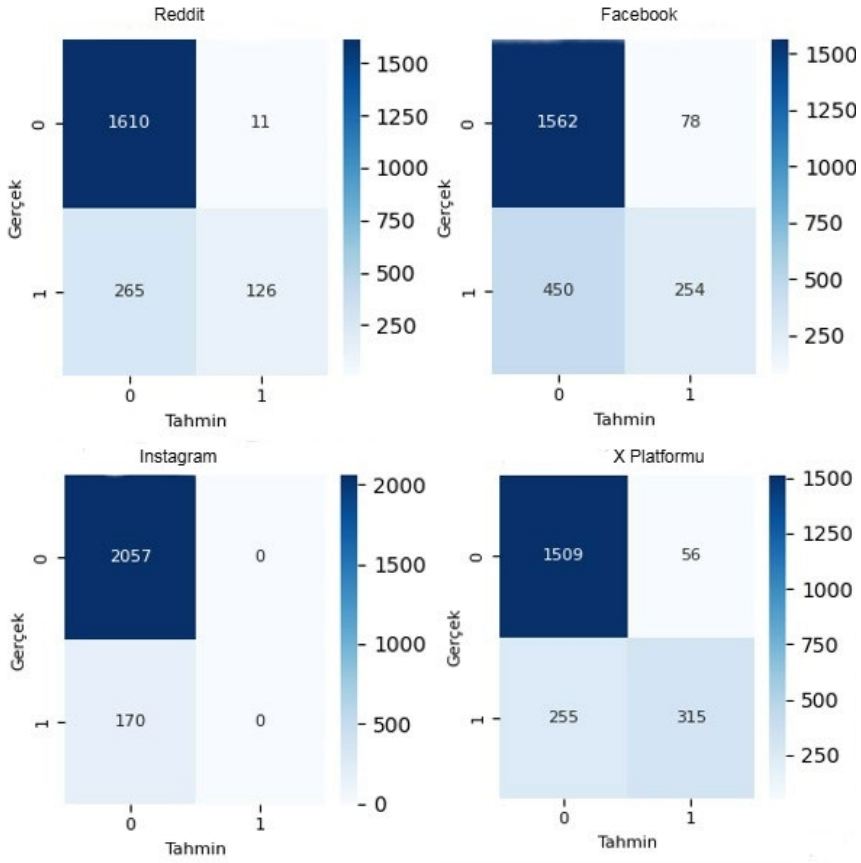
LSTM modeli, platform bazlı ikili sınıflandırma görevinde değişken performanslar sergilemiştir. En yüksek F1 skoru Instagram platformunda %88 ile elde

edilirken, X platformunda %84, Reddit'te %83 ve Facebook platformunda %74 olarak gerçekleşmiştir. Platformlar arasındaki bu farklılıklar, özellikle Facebook'taki veri yapısının ve içerik özelliklerinin modelin öğrenme sürecini zorlaştırdığını göstermektedir. Genel olarak LSTM modeli, özellikle Instagram ve X platformları gibi ortamlarda makul sonuçlar verebilirken, Facebook'taki göreli düşük skorlar veri çeşitliliği ve dengesizliğinin model başarımını etkilediğini ortaya koymaktadır. Ortalama F1 skoru %82 olan LSTM, bu ikili sınıflandırma senaryosunda belirli durumlarda etkili, ancak veri dengesizliğine karşı hassas bir performans sergilemiştir. Çizelge 4.9'de platform bazlı skorlar sunulmuştur.

Çizelge 4.9. LSTM modeli platform bazlı ikili sınıflandırma sonuçları

Platform	Doğruluk	Hassasiyet	Duyarlılık	F1-Skor
Reddit	0.86	0.87	0.86	0.83
Facebook	0.78	0.77	0.78	0.74
Instagram	0.92	0.85	0.92	0.88
X Platformu	0.85	0.85	0.85	0.84
Ortalama	0.85	0.83	0.85	0.82

LSTM modeli ile gerçekleştirilen ikili sınıflandırma görevinde, dört sosyal medya platformuna ait yorumlar "hakaret içeren" ve "hakaret içermeyen" olmak üzere iki temel sınıfa ayrılmış ve modelin tahmin başarımı hata matrisleri üzerinden değerlendirilmiştir. Her bir platform için elde edilen sonuçlar, sınıf bazlı doğruluk oranları ve hata eğilimleri açısından incelenmiş; özellikle azınlık sınıf olan "hakaret içeren" kategorisinde modelin gösterdiği performans ayrıntılı olarak ele alınmıştır. LSTM mimarisinin dilsel örüntüleri ne ölçüde yakalayabildiği ve veri dengesizliği karşısındaki duyarlılığı, Şekil 4.9'da sunulan hata matrisleri aracılığıyla analiz edilmiştir.



Şekil 4.9. LSTM modeli ikili sınıflandırma karmaşıklık matrisleri

LSTM modeli, platform bazlı ikili sınıflandırma görevinde değişken performanslar sergilemiştir. En yüksek makro F1 skoru X platformunda %79 ile elde edilirken, Reddit ve Facebook'ta bu değerler %70 ve %67 olarak gerçekleşmiştir. Öte yandan Instagram' da makro F1 skoru sadece %48 seviyesinde ölçülmüştür. Bu farklılıklar, özellikle Instagram' daki veri yapısı ve içerik özelliklerinin modelin öğrenme sürecini zorlaştırdığını göstermektedir. Elde edilen hata matrislerine göre, modelin tüm platformlarda "0" sınıfında yüksek doğruluk oranlarına ulaştığı görülmektedir. Gerçek "0" sınıfındaki yorumlar Reddit'te 1610, Facebook'ta 1562, Instagram' da 2057 ve X Platformu'nda 1509 kez doğru şekilde sınıflandırılmıştır. Bu sınıfa ait yanlış tahminler ise oldukça sınırlıdır. Öte yandan "1" sınıfında modelin başarısı daha zayıf bir profil çizmektedir. Facebook'ta 254, X Platformu'nda 315 ve Reddit'te 126 örnek doğru "1" tahmini yapılmasına karşın, Instagram' da modele "1" olarak işaretli hiçbir örneği doğru olarak tanıyamamıştır. Tüm "1" sınıfı örneklerin çoğunluğunun "0" olarak etiketlenmesi, modelin azınlık sınıfı ayırt etmekte zorlandığını göstermektedir. Bu durum, sınıf dengesizliği, bağlamsal varyasyonlar ve LSTM mimarisinin karmaşık örüntüleri sınırlı

biçimde temsil edebilmesiyle ilişkilendirilebilir. Platform bazlı sınıflandırma eğilimlerini gösteren hata matrisleri Şekil 4.9'da sunulmuştur.

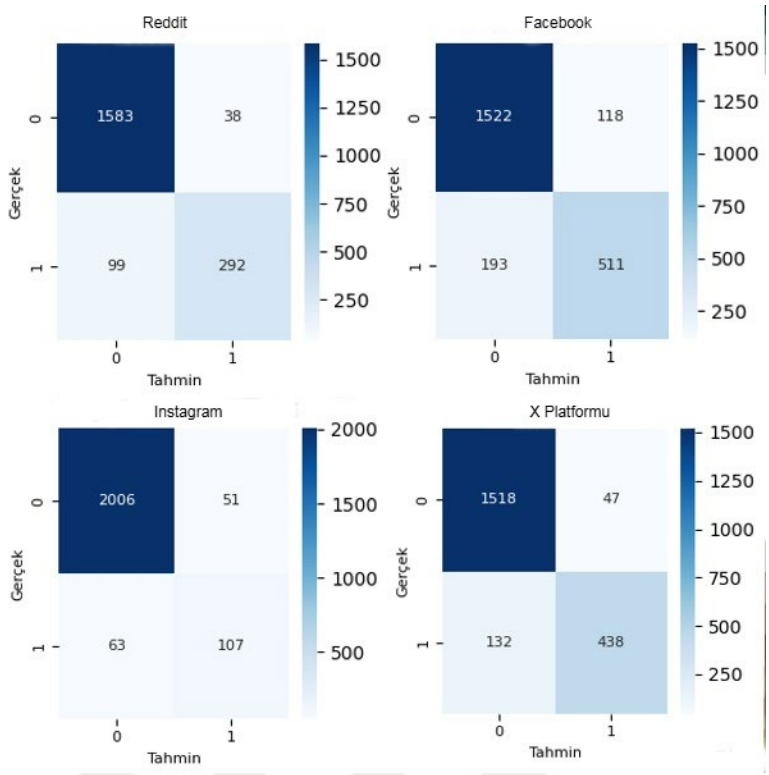
4.4.3. BERTurk Modeli ile İkili Sınıflandırma Sonuçları

BERTurk modeli, platform bazlı ikili sınıflandırma görevinde oldukça yüksek ve dengeli performanslar sergilemiştir. Çizelge 4.10'da görüldüğü üzere, en yüksek F1 skoru %95 ile Instagram platformunda elde edilmiştir. Reddit'te %92, X platformunda %91 ve Facebook'ta %87 F1 skoru kaydedilmiştir. Bu sonuçlar, BERTurk modelinin Türkçe sosyal medya yorumlarında hakaret içerikli ifadeleri iki sınıf arasında başarılı şekilde ayırt edebildiğini göstermektedir. Özellikle Instagram ve Reddit platformlarındaki yüksek F1 skorları, modelin bağlamsal dil çözümleme kapasitesinin ve önceden eğitilmiş derin Transformer mimarisinin gücünü ortaya koymaktadır. BERTurk'un, dilin bağlamsal ilişkilerini CNN ve LSTM gibi klasik modellerden daha iyi öğrenebilmesi, özellikle karmaşık ya da dolaylı ifadelerin sınıflandırılmasında önemli avantaj sağlamıştır. Genel ortalama F1 skoru %91 olan BERTurk, ikili sınıflandırma görevinde hem doğruluk hem de istikrar açısından diğer modellere kıyasla üstün bir performans sergilemiştir.

Çizelge 4.10. BERTurk modeli platform bazlı ikili sınıflandırma sonuçları

Platform	Doğruluk	Hassasiyet	Duyarlılık	F1-Skor
Reddit	0.93	0.93	0.93	0.92
Facebook	0.86	0.87	0.86	0.87
Instagram	0.94	0.95	0.94	0.95
X Platformu	0.91	0.91	0.91	0.91
Ortalama	0.91	0.91	0.91	0.91

BERTurk modeli ile yürütülen ikili sınıflandırma görevinde, dört sosyal medya platformuna ait yorumlar "hakaret içeren" ve "hakaret içermeyen" olmak üzere iki temel kategori altında değerlendirilmiştir. Modelin sınıflandırma performansı, her platform için oluşturulan hata matrisleri üzerinden incelenmiş ve sınıf bazlı doğruluk durumu, yanlış sınıflandırma eğilimleri ile genel başarı düzeyi analiz edilmiştir. Bağlamsal dil modellemesi yeteneği yüksek olan BERTurk, özellikle her iki sınıfı da ayırt etme konusunda önceki modellerden farklı bir eğilim göstermiştir. Modelin pozitif sınıfa dair örüntüleri ne ölçüde tanıyabildiği ve olası karışıklık alanları, Şekil 4.10'da sunulan hata matrisleri üzerinden değerlendirilmiştir.



Şekil 4.10. BERTurk modeli ikili sınıflandırma karmaşıklık matrisleri

BERTurk modeli ile yürütülen ikili sınıflandırma görevinde, dört sosyal medya platformundaki yorumlar "hakaret içeren" ve "hakaret içermeyen" olmak üzere iki kategoriye ayrılmıştır. Modelin sınıflandırma performansı, her platform için oluşturulan hata matrisleri üzerinden incelenmiştir. Gerçek "0" etiketli yorumlar Reddit'te 1583, Facebook'ta 1522, Instagram' da 2006 ve X Platformu'nda 1518 kez doğru sınıflandırılmıştır. Bu yüksek doğruluk oranları, modelin negatif sınıfa ilişkin örüntüleri etkin biçimde öğrendiğini göstermektedir. "1" sınıfında ise diğer modellere göre daha başarılı bir tablo dikkat çekmektedir. Doğru "1" sınıflandırmaları Facebook'ta 511, Reddit'te 292, X Platformu'nda 438 ve Instagram' da 107 olarak gerçekleşmiştir. Yanlış sınıflandırmalar ise görece daha düşük seviyelerde kalmış, örneğin Instagram' da yalnızca 63 hakaret içeren yorum yanlışlıkla "0" olarak etiketlenmiştir. Bu sonuçlar, modelin pozitif sınıfa ait örüntüleri ayırt etme kabiliyetinin gelişmiş olduğunu ve bağlamsal farkları daha iyi yakalayabildiğini göstermektedir. Şekil 4.10'da sunulan hata matrisleri, BERTurk modelinin her iki sınıfı da dengeli bir şekilde ayırt edebildiğini açık biçimde ortaya koymaktadır.

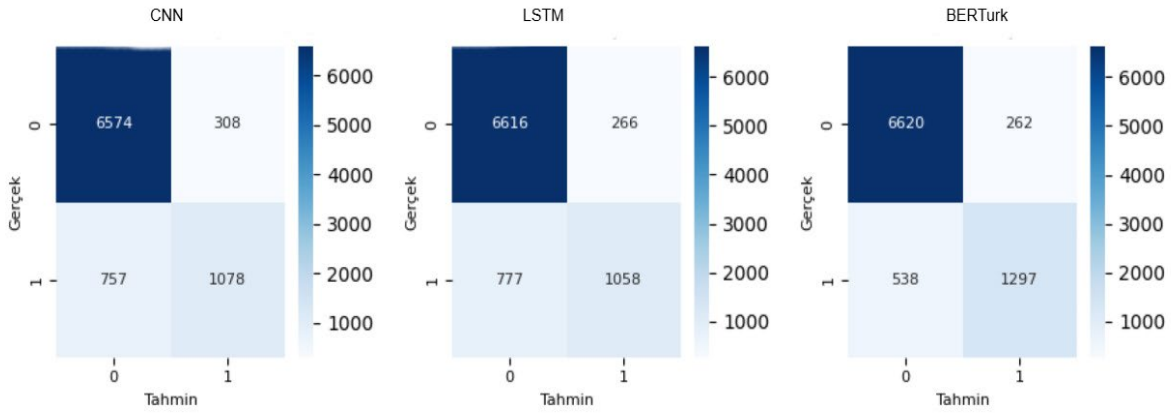
4.4.4. Tüm Platform Verisi ile İkili Sınıflandırma Sonuçları

İkili sınıflandırma görevinde kullanılan CNN, LSTM ve BERTurk modellerinin sonuçları karşılaştırıldığında, BERTurk modelinin tutarlı biçimde en yüksek F1 skorlarına ulaştığı saptanmıştır. Çizelge 4.11'de görüldüğü üzere BERT tabanlı dil modelinin bağlamsal anlama gücü hem hakaret içeren hem de hakaret içermeyen sınıflarda diğer modellere kıyasla daha üstün bir başarı elde etmesini sağlamıştır. CNN ve LSTM modelleri ise özellikle çoğunluk sınıfı olan "hakaret yok" kategorisinde yüksek doğruluk sunsalar da azınlık konumundaki "hakaret var" sınıfında BERTurk kadar başarılı olamamışlardır. Birleşik veri setiyle yapılan değerlendirmelerde, BERTurk modelinin F1 skoru %90'a ulaşırken, CNN ve LSTM modellerinin %87 seviyelerinde kalması da bu farkı ortaya koymaktadır. Bu sonuçlar, çok platformlu geniş bir Türkçe veriyle yürütülen hakaret tespiti çalışmalarında bağlamsal derin öğrenme modellerinin belirgin bir avantaj sağladığını; öte yandan azınlık sınıf performansını iyileştirmek için veri çeşitliliği ve dengeleme tekniklerinin de kritik rol oynadığını göstermektedir. Sonuç itibariyle, BERTurk modeli ikili sınıflandırmada en yüksek başarıyı sağlamış olsa da pozitif sınıfa ait örneklerin daha iyi öğrenilmesine yönelik iyileştirmeler genel sistem performansını daha da artırabilecekti

Çizelge 4.11. Karma veri ikili sınıflandırma sonuçları

Derin Öğrenme Modelleri	Doğruluk	Hassasiyet	Duyarlılık	F1-Skor
CNN	0.88	0.87	0.88	0.87
BERTurk	0.90	0.90	0.90	0.90
LSTM	0.88	0.87	0.88	0.87
Ortalama	0.88	0.88	0.88	0.88

CNN, BERTurk ve LSTM modellerinin tüm sosyal medya platformları birleştirilerek oluşturulan veri seti üzerinde gerçekleştirdiği ikili sınıflandırma sonuçları ele alınmıştır. Yorumlar "hakaret içermeyen" ve "hakaret içeren" olmak üzere iki temel kategoriye ayrılmış, her modelin başarı durumu hata matrisleri üzerinden değerlendirilmiştir. Analizler; modellerin her iki sınıfa yönelik doğru ve yanlış sınıflandırma eğilimlerini ortaya koymakta, özellikle azınlık olan pozitif sınıfa dair ayırt edicilik düzeylerini incelemektedir. Söz konusu sınıflandırma sonuçları, Şekil 4.11'de her model için ayrı ayrı sunulan hata matrisleriyle görsel olarak desteklenmiştir.



Şekil 4.11. Tüm platformlar ikili sınıflandırma karmaşıklık matrisleri

Şekil 4.1'de görülen CNN, LSTM ve BERTurk modellerine ait ikili sınıflandırma hata matrisleri incelendiğinde, BERTurk modelinin her iki sınıfta da daha dengeli ve başarılı sonuçlar ürettiği görülmektedir. CNN ve LSTM modelleri, özellikle hakaret içeren yorumları tespit etmede kısmen zayıf kalırken, BERTurk modeli bu sınıfta daha yüksek doğru sınıflandırma oranına ulaşmıştır. Üç modelde de yanlış pozitif tahminler görece düşük seviyede kalmış, ancak BERTurk modelinin yanlış negatif oranını azaltması ve doğru pozitif tahminlerini artırması, bağlamsal dil anlama kapasitesinin bir göstergesi olarak öne çıkmıştır. Bu genel tablo, Transformer tabanlı modellerin, hakaret tespiti gibi dilin ince anlam farklılıklarını gerektiren sınıflandırma görevlerinde klasik derin öğrenme mimarilerine göre belirgin bir avantaj sunduğunu ortaya koymaktadır.

4.4.5. İkili Sınıflandırma Modellerin Değerlendirmesi

Farklı derin öğrenme mimarileriyle (CNN, LSTM ve BERTurk) gerçekleştirilen ikili sınıflandırma çalışmaları, özellikle BERTurk modelinin hem ayrı platformlarda hem de birleşik veri setinde tutarlı biçimde en yüksek F1 skorlarına ulaştığını ortaya koymuştur. BERTurk, bağlamsal dil modeli altyapısı sayesinde, hakaret içeriklerinin karmaşıklığını ve Türkçedeki dilsel çeşitliliği daha iyi yakalayarak hem pozitif hem de negatif sınıflarda üstün bir başarı sergilemiştir. CNN ve LSTM modelleri ise, özellikle baskın olan hakaret içermeyen sınıfta yüksek doğruluk sunarken, azınlık sınıfta BERTurk kadar başarılı olamamıştır. Bununla birlikte, her üç modelde de veri setindeki dengesizlikten ve azınlık sınıfa ait örneklerin görece az olmasından kaynaklanan yanlış negatifler dikkat çekmektedir. Birleşik veri setiyle yapılan deneylerde, BERTurk %90 F1 skoru ile öne çıkarken, CNN ve LSTM modelleri %87'lik skorlarla iyi düzeyde performans göstermiştir. Sonuçlar, çok platformlu ve geniş ölçekli Türkçe veriyle

yürütülen çalışmalarda, bağlamsal derin öğrenme modellerinin açık bir avantaj sunduğunu; ancak azınlık sınıf performansını iyileştirmek için veri çeşitliliği ve dengeleme tekniklerinin de kritik rol oynadığını göstermektedir.

4.5. Sonuçların Literatürdeki Çalışmalarla Karşılaştırılması

Bu bölümde, tez kapsamında elde edilen sonuçlar ilgili literatürdeki benzer çalışmalarla karşılaştırılmaktadır. Ancak, dört farklı sosyal medya platformunu (Facebook, Instagram, X, Reddit) kapsayan ve yalnızca Türkçe küfür ile hakaret içeriklerine odaklanan herhangi bir çalışmaya rastlanmadığı için, doğrudan karşılaştırmalar yapılamamaktadır. Facebook ve Reddit için Türkçe verilerle yürütülmüş herhangi bir çalışma bulunmadığından bu platformlar için literatür karşılaştırması yapılamamaktadır. Instagram'da ise yalnızca ikili sınıflandırma temelli bir çalışma yer almakta, kategorik sınıflandırmaya dair araştırma bulunmadığı için kategorik karşılaştırma yapılamamaktadır. X platformu için ise hem ikili hem de kategorik sınıflandırma çalışmaları mevcuttur. Literatür karşılaştırması yapılırken platformlara göre mevcut çalışmaların türleri ve F1 skorları baz alınarak değerlendirmeler yapılacaktır.

4.5.1. X Platformu Kategorik Sınıflandırma Sonuçlarının Karşılaştırılması

X platformunda yapılan kategorik sınıflandırma deneyleri, literatürdeki Çöltekin [4] çalışmasıyla karşılaştırılmıştır. Çöltekin'in çalışmasında, 35.282 tweetten oluşan geniş bir veri seti kullanılmış ve SVM (Linear Support Vector Machine) algoritması ile %45 başarı elde edilmiştir. Bu tez çalışmasında ise, Çizelge 4.12'de görülebileceği üzere, CNN ve LSTM modelleri %81 F1 skoru, BERT modeli ise %87 F1 skoru ile daha yüksek performans sergilemiştir.

Elde edilen %87'lik BERT başarımı, hem klasik makine öğrenmesi algoritmalarına hem de diğer derin öğrenme modellerine göre daha üstün bir sonuçtur. CNN ve LSTM modellerinin %81 oranındaki başarıları ise literatürdeki SVM sonucunu neredeyse iki katına çıkarmaktadır. Bu durum, transformer tabanlı modellerin özellikle bağlamsal temsilleri yakalama konusunda avantajlı olduğunu göstermektedir.

Bu tezde ayrıca, Çöltekin'in etiketleme yapısında yer alan ve anlamı belirsiz olan 'X' kategorisi çıkarılarak, yalnızca anlamlı sınıflar kullanılmıştır. Bu değişiklik, model performansını olumlu yönde etkilemiş ve sınıflandırma başarısını yükseltmiştir.

Sonuç olarak, X platformunda gerçekleştirilen bu çalışma, literatürdeki mevcut sonuçlara göre daha yüksek doğruluk sağlamış ve Türkçe sosyal medya verilerinde hem derin öğrenme hem de transformer tabanlı modellerin etkinliğini ortaya koymuştur.

Çizelge 4.12. X platformu kategorik sınıflandırma benzer çalışmalar

Çalışma	Yıl	Kullanılan Yöntemler	Veri Seti	Başarım(%)
Çöltekin [4]	2020	SVM	35.282	45
Bu tez çalışması	2025	CNN	10.674	81
		BERTurk	10.674	87
		LSTM	10.674	81

4.5.2. Instagram İkili Sınıflandırma Sonuçlarının Karşılaştırılması

Bu bölümde, Instagram platformunda gerçekleştirilen ikili sınıflandırma sonuçları Karayiğit ve ark [6] çalışması ile karşılaştırılmaktadır. Çizelge 4.13'te, her iki çalışmada kullanılan yöntemler, veri seti büyüklükleri ve elde edilen başarı oranları özetlenmektedir.

Karayiğit ve ark [6] çalışmasında, 30.084 yorumdan oluşan geniş bir veri seti üzerinde hem geleneksel makine öğrenmesi algoritmaları (SVM, Random Forest, Lojistik Regresyon) hem de derin öğrenme temelli CNN modeli kullanılmıştır. Bu çalışmada CNN ve SVM modelleri %93, Random Forest %92, Lojistik Regresyon ise %90 başarı oranına ulaşmıştır.

Bu tez çalışmasında ise, 11.133 yorum içeren daha küçük fakat dengeli bir veri seti üzerinde üç farklı derin öğrenme modeli test edilmiştir. CNN modeli %92 başarı sağlamış ve literatürdeki CNN sonucuna oldukça yakın bir performans sergilemiştir. LSTM modeli %88 başarı oranı ile diğer modellere göre daha düşük kalmış; ancak RNN tabanlı bir yöntem olarak dildeki sıralı bağımlılıkları yakalama gücünü göstermiştir. BERT modeli ise %95'lik başarımı ile hem bu tezdeki diğer derin öğrenme modellerini hem de Karayiğit ve ark. çalışmasındaki tüm modelleri geride bırakmıştır.

Bu sonuçlar, derin öğrenme modellerinin genel olarak geleneksel makine öğrenmesi yöntemlerine göre daha yüksek performans sergilediğini, özellikle BERT modelinin bağlamsal öğrenme yeteneği sayesinde Instagram verilerinde üstün başarı sağladığını göstermektedir.

Sonuç olarak, Instagram platformu üzerinde gerçekleştirilen bu tez çalışması, CNN, LSTM ve BERT modelleri kullanılarak yapılan ikili sınıflandırma deneyleriyle, hem literatürdeki benzer Türkçe çalışmalarla paralellik göstermiş hem de özellikle BERT modeliyle elde edilen yüksek başarı oranları sayesinde önemli katkılar sunmuştur. BERT modelinin sağladığı üstün performans, derin öğrenme tabanlı yaklaşımların Türkçe hakaret içeriklerinin otomatik olarak tespit edilmesinde güçlü bir alternatif olduğunu ortaya koymaktadır. Bu bağlamda çalışma, Türkçe sosyal medya verileri üzerinde gerçekleştirilen ikili hakaret sınıflandırmasında derin öğrenme yöntemlerinin potansiyelini somut bir şekilde göstermiştir

Çizelge 4.13. Instagram ikili sınıflandırma literatürdeki benzer çalışmalar

Çalışma	Yıl	Kullanılan Yöntemler	Veri Seti	Başarım (%)
Karayiğit ve ark. [6]	2021	CNN	30.084	93
		SVM	30.084	93
		RF	30.084	92
		LR	30.084	90
Bu tez çalışması	2025	CNN	11.133	92
		BERTurk	11.133	95
		LSTM	11.133	88

4.5.3. X Platformu İkili Sınıflandırma Sonuçlarının Karşılaştırılması

X platformu üzerindeki ikili sınıflandırma sonuçları, literatürde yer alan farklı çalışmalarla karşılaştırılmıştır. Çizelge 4.14'te, önceki çalışmaların başarı oranları ile bu tez kapsamında elde edilen sonuçların karşılaştırması sunulmaktadır.

Yılmaz ve ark. [18], 14.752 tweetten oluşan veri seti üzerinde LSTM ile %85, GRU ile %87 başarı oranı elde etmiştir. Mayda ve ark. [17] çalışmasında ise, 1000 tweetlik küçük bir veri seti kullanılarak SMO algoritmasıyla %86 başarı sağlanmıştır. Çöltekin [4] çalışmasında SVM ile %77 başarı oranı rapor edilmiştir. Canbay ve Ekinci [19] çalışmasında ise CNN %73, LSTM %44, BiLSTM %74, GRU %44, BiGRU %73 başarı göstermiştir.

Bu tez çalışmasında ise, 10.674 tweet içeren veri seti üzerinde CNN modeli %85, LSTM modeli %84 ve BERTurk modeli %91 başarı oranına ulaşmıştır. Özellikle BERTurk modelinin %91'lik başarımı, literatürde raporlanan en yüksek başarı oranı olan GRU modeline (%87) göre %4 daha yüksek sonuç vermiştir. CNN ve LSTM modelleri

de literatürdeki CNN ve LSTM başarı oranlarına kıyasla benzer ya da daha iyi sonuçlar elde etmiştir.

Bu farklılıkların temel nedenleri arasında, BERTurk modelinin bağlamsal anlam öğrenme yeteneği, ön işleme adımlarının kapsamlı şekilde uygulanması ve veri setinin daha dengeli bir yapıda hazırlanmış olması gösterilebilir. Ayrıca, bazı çalışmalarda çok büyük veri setleri kullanılmasına rağmen başarı oranlarının düşük kalması, transformer tabanlı modellerin Türkçe veriler üzerindeki avantajını ortaya koymaktadır.

X platformunda gerçekleştirilen bu tez çalışmasında BERTurk modeli literatürdeki tüm çalışmalardan daha yüksek başarı sağlamış; CNN ve LSTM modelleri ise önceki çalışmalardaki aynı modellerle karşılaştırıldığında benzer veya kısmen daha yüksek sonuçlar elde etmiştir. Bu bulgular, özellikle transformer tabanlı modellerin Türkçe ikili hakaret sınıflandırmasında güçlü bir potansiyele sahip olduğunu göstermektedir.

Çizelge 4.14. X platformu ikili sınıflandırma literatürdeki benzer çalışmalar

Çalışma	Yıl	Kullanılan Yöntemler	Veri Seti	Başarım (%)
Çöltekin [4]	2020	SVM	35.282	77
Mayda ve ark. [17]	2021	SMO (Sequential Minimal Optimization)	1.000	86
Yılmaz ve ark. [18]	2022	LSTM	1.4752	85
		GRU	1.4752	87
Canbay ve Ekinci [19]	2023	CNN	35.282	73
		LSTM	35.282	44
		BİLSTM	35.282	74
		GRU	35.282	44
		BİGRU	35.282	73
Bu Tez Çalışması	2025	CNN	10.674	85
		BERTurk	10.674	91
		LSTM	10.674	84

5. SONUÇ VE ÖNERİLER

Bu tez çalışmasında, farklı sosyal medya platformlarından (Facebook, Instagram, X ve Reddit) toplanan Türkçe kullanıcı yorumları üzerinden otomatik hakaret tespiti gerçekleştirmek amacıyla, derin öğrenme temelli sınıflandırma modelleri kullanılmıştır. Çalışmada hem her platforma özgü ayrı modeller hem de tüm platformları kapsayan birleşik modeller eğitilerek, veri yapısının sınıflandırma başarımı üzerindeki etkisi incelenmiştir. Model mimarisi olarak CNN, LSTM ve BERTurk yapıları tercih edilmiş; sınıflandırma görevleri hem çok sınıflı (beş kategori) hem de ikili (hakaret var/yok) düzeyde yürütülmüştür. Buna ek olarak, büyük dil modeli tabanlı GPT-4o kullanılarak sıfır örnekli (zero-shot), tek örnekli (one-shot) ve üç örnekli (three-shot) otomatik etiketleme çalışmaları gerçekleştirilmiştir.

Elde edilen bulgulara göre, ikili sınıflandırma görevinde BERTurk modeli %90 F1 skoru ile en başarılı sonuçları sunmuş, CNN ve LSTM modelleri ise %87'lik skorlarla iyi düzeyde performans göstermiştir. Çok sınıflı sınıflandırma görevinde ise BERTurk %87 F1 skoru ile yine en iyi sonucu verirken, CNN ve LSTM modelleri %82 ve %81 gibi daha düşük performans göstermiştir. Bu durum, BERTurk'un Türkçe dilindeki bağlamsal ilişkileri daha iyi öğrenebilmesiyle açıklanabilir. Ayrıca, platformlar birleştirildiğinde elde edilen modelin, çoğu durumda platforma özel modellerle aynı düzeyde veya daha iyi sonuçlar verdiği gözlemlenmiştir. Bu, modelin platformlar arası örüntüleri öğrenerek farklı veri kaynakları karşısında daha iyi genelleyebildiğini göstermektedir.

GPT-4o modelinin kullanıldığı otomatik etiketleme çalışmasında, zero-shot senaryosunda ortalama %60, one-shot senaryosunda %69 ve three-shot senaryosunda %63 F1 skorları elde edilmiş; özellikle tek örnekli senaryoda daha istikrarlı sonuçlar üretmiştir. Ancak modelin bazı durumlarda eleştirel ifadeleri ya da sert üslupları doğrudan hakaret olarak değerlendirdiği gözlemlenmiştir. Bu durum, modelin bağlamı yeterince derinlemesine analiz edemediği ve yorumlardaki anlam nüanslarını sınıflandırmada zaman zaman hataya düştüğü yönünde önemli bir bulgu sunmaktadır.

Elde edilen başarımlara rağmen, model performansını sınırlayan bazı etkenler gözlemlenmiştir. Özellikle veri setindeki sınıf dengesizliği, azınlık sınıflarda yüksek hata oranlarına ve düşük F1 skorlarına neden olmuş, sınıflar arası benzerlikler ve etiketleme sürecindeki yorum farklılıkları model performansını sınırlamıştır. Ayrıca, platformlara

özgü dilsel çeşitlilik, her modele eşit başarı sağlamamış ve bazı bağlamlarda sınıflandırma doğruluğunu olumsuz etkilemiştir.

Sonuç olarak, bu tez çalışması, dört farklı sosyal medya platformundan (Facebook, Instagram, X ve Reddit) toplanan Türkçe kullanıcı yorumları üzerinden otomatik hakaret tespiti görevinde, derin öğrenme tabanlı modellerin anlamlı düzeyde sınıflandırma başarısı sağlayabildiğini ortaya koymuştur. CNN, LSTM ve BERTurk olmak üzere üç farklı modelle hem çok sınıflı hem de ikili sınıflandırma deneyleri gerçekleştirilmiş; özellikle ikili sınıflandırma görevinde BERTurk modeli %90 F1 skoru ile en başarılı sonuçları vermiştir. Çok sınıflı görevlerde de BERTurk %87 F1 skoru ile en yüksek performansa ulaşmıştır. Ayrıca, platformlar birleştirildiğinde oluşturulan modelin, çoğu durumda platforma özel modellerle benzer hatta daha iyi sonuçlar verdiği gözlemlenmiş; bu da tek ve kapsamlı bir modelin, farklı sosyal medya bağlamlarında genelleyici bir başarı ortaya koyabileceğini göstermiştir. Bu deneysel analizlere ek olarak, çalışmada GPT-4o tabanlı OpenAI API aracılığıyla sıfır, bir ve üç örnekli otomatik etiketleme senaryoları da uygulanmış; büyük dil modeli, özellikle one-shot senaryosunda %69'a ulaşarak daha başarılı sonuçlar vermiş olsa da genel olarak derin öğrenme modelleri kadar yüksek sınıflandırma başarısına ulaşamamıştır. Bu durumun temel nedeni, modelin bazı ifadelerdeki ironik, dolaylı veya eleştirel anlatımları doğru bağlamda yorumlayamaması ve bu tür iletileri yanlış biçimde hakaret olarak etiketleme eğiliminde olmasıdır.

Gelecekte yapılacak çalışmalarda, özellikle azınlık sınıflardaki performans düşüklüğünü gidermek adına veri artırımı ve dengeleme tekniklerine başvurulması faydalı olacaktır. Bu kapsamda eşanlamlı kelime değiştirme, geri çeviri gibi yöntemlerle yapay veri üretimi yapılabilir; ayrıca ağırlıklı kayıp fonksiyonları veya aşırı örnekleme gibi stratejilerle modelin az temsil edilen sınıfları daha iyi öğrenmesi sağlanabilir. Bunun yanı sıra, güncel ve daha büyük dil modellerinin (örneğin XLM-RoBERTa, ConvBERT gibi) denenmesi ya da birden fazla modelin güçlü yönlerini birleştiren ansambl yaklaşımlarının uygulanması, genel sınıflandırma başarımını artırabilir. Ayrıca, mevcut kategorik etiketleme yaklaşımına alternatif olarak çoklu etiketleme ya da iki aşamalı hiyerarşik sınıflandırma sistemlerinin kullanılması, yorumlardaki birden fazla hakaret türünün eşzamanlı olarak tespit edilmesini mümkün kılarak çok boyutlu içeriklerin daha doğru biçimde sınıflandırılmasına imkân tanıyabilir. Etiketleme sürecinde yorumcular arası tutarlılığı artıracak daha sağlam anotasyon protokollerinin uygulanması da veri

kalitesini yükselterek modelin öğrenme sürecine doğrudan katkı sağlayacaktır. Son olarak, farklı platformlara özgü dilsel kalıpları daha iyi temsil edebilmek için platforma özel sözlüklerin veya vektör modellerinin geliştirilmesi, modellerin bağlama duyarlı tepkiler verebilmesini kolaylaştıracaktır.

KAYNAKLAR

[1] M. Çalışkan ve Y. Mencik, "Değişen dünyanın yeni yüzü: Sosyal medya", *Akademik Bakış Dergisi*, sy 50, 2015.

[2] Z. Biricik, "Sosyal Medyada Ünlülere Yönelik Nefret Söylemi: Twitter Üzerine Bir İnceleme (Hate Speech to Famous in Social Media: A Review on Twitter)", *Turk Turizm Arastirmalari Dergisi*, 2022, doi: 10.26677/tr1010.2022.1106.

[3] İ. H. Aydın, "Sosyal Medya Aracılığıyla Hakaret Suçu | Özgün Law Firm". Erişim: 02 Haziran 2025. [Çevrimiçi]. Erişim adresi: https://www.ozgunlaw.com/makaleler/sosyal-medya-araciligiyla-hakaret-sucu-1066?

[4] Ç. Çöltekin, "A corpus of turkish offensive language on social media", içinde *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*, 2020, ss. 6174-6184.

[5] L. Soykan, C. Karsak, I. D. Elkahlout, ve B. Aytan, "A Comparison of Machine Learning Techniques for Turkish Profanity Detection", içinde *International Workshop on Resources and Techniques for User Information in Abusive Language Analysis, ResT-UP 2022 - in conjunction with the International Conference on Language Resources and Evaluation, LREC 2022 - Proceedings*, 2022.

[6] H. Karayiğit, Ç. İnan Acı, ve A. Akdağlı, "Detecting abusive Instagram comments in Turkish using convolutional Neural network and machine learning methods", *Expert Syst Appl*, c. 174, s. 114802, Tem. 2021, doi: 10.1016/J.ESWA.2021.114802.

[7] B. Büyüktanır, Ö. Yakar, ve A. B. A. Girgin, "Sosyal medyada nefret söylemi çerçevesi: Geçmişten günümüze kapsamlı bir derleme", *Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi*, c. 40, sy 1, ss. 685-712, Ağu. 2024, doi: 10.17341/GAZIMMFD.1327840.

[8] "Mevzuat Bilgi Sistemi". Erişim: 02 Haziran 2025. [Çevrimiçi]. Erişim adresi: https://www.mevzuat.gov.tr/mevzuat?MevzuatNo=5237&MevzuatTur=1&MevzuatTertip=5

[9] F. Acay, “Sosyal Medya Aracılığıyla Hakaret Suçu ve Suçun Tespitine İlişkin Uygulamalar”, *İstanbul Aydın Üniversitesi Hukuk Fakültesi Dergisi*, c. 7, sy 1, ss. 71-140, Haz. 2021, doi: 10.17932/IAU.HFD.2015.018/hfd_v07i1003.

[10] B. Y. Çakmakkaya, E. Lokmanoğlu, ve T. Akpinar, “Sosyal Medya Üzerinden Kişilik Haklarının İhlali Ve Hakaret Suçu Davalarında Hukuka Uygunluk Sebebi Olarak Mefruz Rıza”, *Balkan and Near Eastern Journal of Social Sciences Balkan ve Yakın Doğu Sosyal Bilimler Dergisi*, c. 2024, sy 01, s. 10, 2024.

[11] S. Zeybek, B. Alkın, ve Y. Kaya, “Derin öğrenme ve makine öğrenmesi yöntemleri ile sosyal medya verilerinden suç tespiti”, *Niğde Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi*, c. 14, sy 1, ss. 175-182, Oca. 2025, doi: 10.28948/NGUMUH.1551734.

[12] S. Özar, “Türk Hukukunda Nefret Suçlarına -Avrupa Güvenlik ve İşbirliği Teşkilatı Taahhütleri Çerçevesinde- Genel Bir Bakış”, *Türkiye Adalet Akademisi Dergisi*, c. 0, sy 48, ss. 87-108, Eki. 2021, doi: 10.54049/TAAD.1009196.

[13] E. K. Sinanlıoğlu, “Sosyal Medyada Gerçekleşen Kişilik Hakkı İhlalleri Ve Korunma Yolları”, *İstanbul Ticaret Üniversitesi Sosyal Bilimler Enstitüsü, Yüksek Lisans Tezi*, 2021.

[14] Ç. Çelen, “Sosyal medyada değişen ifade kültürünün hakaret suçuna etkisi | AVESİS”. Erişim: 02 Haziran 2025. [Çevrimiçi]. Erişim adresi: https://avesis.deu.edu.tr/yonetilen-tez/3fbaa203-861c-4033-a8e8-0f32df42ca82/sosyal-medyada-degisen-ifade-kulturunun-hakaret-sucuna-etkisi

[15] B. Doğan, “İnternetten veya Sosyal Medya Üzerinden Hakaret Suçu”. Erişim: 02 Haziran 2025. [Çevrimiçi]. Erişim adresi: https://barandogan.av.tr/blog/ceza-hukuku/internetten-ve-sosyal-medya-uzerinden-hakaret-sucu-cezasi.html

[16] E. E. Beyazit, “Hakaret Suçu ve Cezası”. Erişim: 02 Haziran 2025. [Çevrimiçi]. Erişim adresi: https://www.harbiyehukuk.com/hakaret-sucu-ve-cezasi/

[17] İ. Mayda, B. Diri, ve T. Dalyan, “Türkçe Tweetler üzerinde Makine Öğrenmesi ile Nefret Söylemi Tespiti”, *European Journal of Science and Technology Special Issue*, c. 24, ss. 328-334, 2021, doi: 10.31590/ejosat.903854.

[18] Ş. Ş. Yılmaz, İ. Özer, ve H. Gökçen, "Twitter Platformundan Elde Edilen Türkçe Saldırgan Dil Derlemi", *Mühendislik Bilimleri ve Araştırmaları Dergisi*, c. 4, sy 2, ss. 304-316, Eki. 2022, doi: 10.46387/BJESR.1173434.

[19] P. Canbay ve E. Ekinci, "Derin ve Sığ Makine Öğrenmesi Yöntemleri ile Türkçe Tweet'lerden Saldırgan Dil Tespiti", *Bilgisayar Bilimleri ve Mühendisliği Dergisi*, c. 16, ss. 1-10, 2023.

[20] Y. Ashok, F. A. Khan, V. Singh, F. Aslam Khan, ve V. Singh, "A Multi-Architecture Approach for Offensive Language Identification Combining Classical Natural Language Processing and BERT-Variant Models", *Applied Sciences 2024, Vol. 14, Page 11206*, c. 14, sy 23, s. 11206, Ara. 2024, doi: 10.3390/APP142311206.

[21] A. Bihari *vd.*, "Identification of Hate Speech on Social Media using LSTM", *GMSARN International Journal*, c. 17, ss. 468-474, 2023.

[22] R. Hada, S. Sudhir, P. Mishra, H. Yannakoudakis, S. M. Mohammad, ve E. Shutova, "Ruddit: Norms of Offensiveness for English Reddit Comments", *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, ss. 2700-2717, 2021, doi: 10.18653/V1/2021.ACL-LONG.210.

[23] Y. Lee, S. Yoon, ve K. Jung, "Comparative Studies of Detecting Abusive Language on Twitter", *2nd Workshop on Abusive Language Online - Proceedings of the Workshop, co-located with EMNLP 2018*, ss. 101-106, 2018, doi: 10.18653/V1/W18-5113.

[24] J. H. Park ve P. Fung, "One-step and Two-step Classification for Abusive Language Detection on Twitter", *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, ss. 41-45, 2017, doi: 10.18653/V1/W17-3006.

[25] P. Badjatiya, S. Gupta, M. Gupta, ve V. Varma, "Deep Learning for Hate Speech Detection in Tweets", *26th International World Wide Web Conference 2017, WWW 2017 Companion*, ss. 759-760, Haz. 2017, doi: 10.1145/3041021.3054223.

[26] K. B. Zümberoğlu, S. Z. Dik, B. S. Karadeniz, ve S. Sahmoud, "Towards Better Sentiment Analysis in the Turkish Language: Dataset Improvements and Model Innovations", *Applied Sciences 2025, Vol. 15, Page 2062*, c. 15, sy 4, s. 2062, Şub. 2025, doi: 10.3390/APP15042062.

[27] A. B. Altinel, G. K. Baydogmus, S. Sahin, ve M. Z. Gurbuz, "So-haTRed: A Novel Hybrid System for Turkish Hate Speech Detection in Social Media With Ensemble Deep Learning Improved by BERT and Clustered-Graph Networks", *IEEE Access*, c. 12, ss. 86252-86270, 2024, doi: 10.1109/ACCESS.2024.3415350.

[28] C. Balli, M. S. Guzel, E. Bostanci, ve A. Mishra, "Sentimental Analysis of Twitter Users from Turkish Content with Natural Language Processing", *Comput Intell Neurosci*, c. 2022, sy 1, s. 2455160, Oca. 2022, doi: 10.1155/2022/2455160.

[29] Y. Lecun, Y. Bengio, ve G. Hinton, "Deep learning", *Nature*, c. 521, sy 7553, ss. 436-444, May. 2015, doi: 10.1038/NATURE14539.

[30] Y. Kim, "Convolutional Neural Networks for Sentence Classification", *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, ss. 1746-1751, 2014, doi: 10.3115/V1/D14-1181.

[31] Y. Zhang ve B. C. Wallace, "A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification", Eki. 2015, doi: https://doi.org/10.48550/arXiv.1510.03820.

[32] S. Albelwi ve A. Mahmood, "A Framework for Designing the Architectures of Deep Convolutional Neural Networks", 2017, doi: 10.3390/e19060242.

[33] V. Tümen, "Akıllı Ulaşım Sistemleri İçin Karayolu Tipi, Kavşak ve Virajların Derin Öğrenme Yöntemleri ile Belirlenmesi", *Fen Bilimleri Enstitüsü, Fırat Üniversitesi, Elazığ.*, 2019.

[34] E. Başaran, "Timpanik Membran Görüntü Analizi ve Yapay Zeka Kullanılarak Sanal Otitis Media Tanı Sisteminin Geliştirilmesi Erdal Başaran 2020 Doktora Tezi Bilgisayar Mühendisliği", 2020.

[35] A. Vaswani *vd.*, "Attention Is All You Need", *Adv Neural Inf Process Syst*, c. 2017-December, ss. 5999-6009, Haz. 2017, Erişim: 03 Haziran 2025. [Çevrimiçi]. Erişim adresi: https://arxiv.org/pdf/1706.03762

[36] J. Devlin, M. W. Chang, K. Lee, ve K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, c. 1, ss. 4171-4186, Eki. 2018.

[37] "dbmdz/bert-base-turkish-uncased · Hugging Face". Erişim: 03 Haziran 2025. [Çevrimiçi]. Erişim adresi: https://huggingface.co/dbmdz/bert-base-turkish-uncased

[38] M. Arzu ve M. Aydoğan, "Türkçe Duygu Sınıflandırma İçin Transformers Tabanlı Mimarilerin Karşılaştırılmalı Analizi", *Journal of Computer Science*, ss. 1-6, 2023, doi: 10.53070/bbd.1350405.

[39] S. Hochreiter ve J. Schmidhuber, "Long Short-Term Memory", *Neural Comput*, c. 9, sy 8, ss. 1735-1780, Kas. 1997, doi: 10.1162/NECO.1997.9.8.1735.

[40] F. A. Gers, J. Schmidhuber, ve F. Cummins, "Learning to Forget: Continual Prediction with LSTM", *Neural Comput*, c. 12, sy 10, ss. 2451-2471, Eki. 2000, doi: 10.1162/089976600300015015.

[41] İ. Ayaz, "Forecasting CO_2 Emissions with Machine Learning Methods: Türkiye Example and Future Trends", *NATURENGS*, c. 5, sy 2, ss. 82-87, Ara. 2024, doi: 10.46572/NATURENGS.1595329.

[42] M. Ma, R. Chen, X. Wang, Q. Xie, Y. Wang, ve T.-M. Ma, "Study on Ammonia Concentration Prediction Model of Pigsty Based on LTSM Neural Network", *Sch J Agric Vet Sci*, sy 7, ss. 80-84, 2022, doi: 10.36347/sjavs.2022.v09i07.001.

[43] M. Sundermeyer, R. Schlüter, ve H. Ney, "LSTM neural networks for language modeling", *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012*, c. 1, ss. 194-197, 2012, doi: 10.21437/INTERSPEECH.2012-65.

[44] D. M. Fernández, E. Cernadas, S. Barro, D. Amorim, ve A. Fernández-Delgado, "Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?", *Journal of Machine Learning Research*, c. 15, ss. 3133-3181, 2014.

[45] İ. Ayaz, F. Kutlu, ve Z. Cömert, "DeepMaizeNet: A novel hybrid approach based on CBAM for implementing the doubled haploid technique", *Agron J*, c. 116, sy 3, ss. 861-870, 2024.

[46] A. Tharwat, "Classification assessment methods", *Applied Computing and Informatics*, c. 17, sy 1, ss. 168-192, 2018, doi: 10.1016/J.ACI.2018.08.003/FULL/PDF.

[47] D. Arı ve M. Burukanlı, "Real-Tıme Human Bone Fracture Detectıon Usıng Yolo Models", *ASES IX. International Scientific Research Congress*, ss. 172-183, 2025.