International Journal of Multidisciplinary Studies and Innovative Technologies

e-ISSN: 2602-4888
dergipark.org.tr/en/pub/ijmsit
Research Article

2025, 9 (1), 118-124
DOI: 10.36287/ijmsit.9.1.16
Received: June 23, 2025; Accepted: July 22, 2025

Yapay Zeka ve Doğal Dil İşleme ile Türkçe Metinlerden Siber Zorbalık Tespiti

Halid Buğra Gökdeniz[1*], Pelin Canbay[2]

[1]Kahramanmaraş Sütçü İmam Üniversitesi, Bilgisayar Mühendisliği Bölümü, Kahramanmaraş, Türkiye (bugragkdeniz@gmail.com)*
[2]Kahramanmaraş Sütçü İmam Üniversitesi, Bilgisayar Mühendisliği Bölümü, Kahramanmaraş, Türkiye (pelincanbay@ksu.edu.tr) (ORCID: 0000-0002-8067-3365)

Özet – Siber zorbalık, dijital platformların yaygınlaşmasıyla birlikte ciddi bir toplumsal tehdit haline gelmiştir. Anonimliğin sağladığı sahte güven ortamı, bireylerin gerçek dünyada sergilemekten kaçındıkları saldırgan davranışları sanal ortamlarda rahatça uygulayabilmelerine neden olmaktadır. Siber zorbalık sadece mağdurlar üzerinde psikolojik ve sosyal tahribat yaratmakla kalmayıp, zamanla mağdurların da benzer davranış biçimlerini benimsemelerine neden olarak, adeta bulaşıcı bir hastalık gibi yayılabilmektedir. Bu tür içeriklerin manuel olarak tespiti, hem zaman hem de insan kaynağı açısından son derece maliyetli ve verimsizdir. Bu nedenle, doğal dil işleme (NLP) ve yapay zeka tabanlı otomatik tespit sistemlerinin geliştirilmesi büyük önem taşımaktadır. Literatürde İngilizce gibi yaygın diller üzerine yapılan çalışmalar yoğunlukta olsa da, Türkçe gibi morfolojik açıdan zengin ve kaynak bakımından kısıtlı diller için siber zorbalık tespiti hâlâ gelişmekte olan bir araştırma alanıdır. Bu çalışmada, Türkçe metinler üzerinde geleneksel makine öğrenmesi yöntemleri (TF-IDF + SVM gibi) ile derin öğrenme tabanlı modeller (GRU, BiGRU, CNN, BiLSTM) karşılaştırmalı olarak incelenmiştir. Derin öğrenme modellerinde kelime temsili için FastText embedding yöntemi kullanılmış ve BiLSTM ile BiGRU modelleri %96 doğruluk oranına ulaşarak en yüksek performansı göstermiştir. Bulgular, yapay zeka destekli NLP çözümlerinin Türkçe dijital metinlerde siber zorbalığın etkin bir şekilde tespiti için güçlü bir araç sunduğunu ortaya koymaktadır.

Anahtar Kelimeler *– Yapay Zeka, Siber Zorbalık tespiti, NLP, Türkçe, Derin Öğrenme*

Atıf: Gökdeniz, H. B., Canbay, P. (2025). Yapay Zeka ve Doğal Dil İşleme ile Türkçe Metinlerden Siber Zorbalık Tespiti. International Journal of Multidisciplinary Studies and Innovative Technologies, 9(1): 118-124.

Detection of Cyberbullying from Turkish Texts with Artificial Intelligence and Natural Language Processing

Research Problem/Questions – *Is it possible to automatically detect cyberbullying content in Turkish online conversations using artificial intelligence and natural language processing techniques? Can classical machine learning and deep learning models provide significant accuracy and performance outcomes for this purpose?*
Short Literature Review – *Prior research has mostly concentrated on English data, demonstrating the efficacy of NLP and deep learning techniques to detect cyberbullying. However, there is a lack of studies on morphologically complicated and low-resourse languages like Turkish. This circumstance highlights the necessity for linguistic adaptation and modeling.*
Methodology – *This study comparatively employed traditional machine learning and deep learning techniques for the detection of cyberbullying in Turkish texts. Traditional models (LR, SVM, KNN, RF) utilized TF-IDF for text representation, whereas deep learning models employed the FastText word embedding approach. Although deep learning models like GRU, BiGRU, CNN, and BiLSTM show high accuracy, the optimal performance was achieved by the BiLSTM and BiGRU models, with 96% accuracy.*
Results – *Deep learning models have achieved notable success in detecting cyberbullying, achieving accuracy rates reaching up to 96%. The findings reveal that AI and NLP -based solutions can be effective in more complicated and low-resources languages, such as Turkish, and provide considerable potential for better security in digital contexts.*

Keywords *– Artificial Intelligence, Cyberbullying detection, NLP, Turkish, Deep Learning*

Citation: Gökdeniz, H. B., Canbay, P. (2025). Detection of Cyberbullying from Turkish Texts with Artificial Intelligence and Natural Language Processing. International Journal of Multidisciplinary Studies and Innovative Technologies, 9(1): 118-124.

I. GİRİŞ

İnternetin ve sosyal medya platformlarının hayatın her alanına entegre olmasıyla birlikte, dijital iletişim araçları hem bireyler arası etkileşimi artırmış hem de yeni türden sosyal problemleri beraberinde getirmiştir. Bu sorunlardan biri olan siber zorbalık, bireylerin çevrim içi ortamlarda, özellikle çevrim içi oyun platformlarında bilinçli ve tekrar eden bir şekilde zarar verici davranışlar sergilemesiyle karakterize

edilen, giderek yaygınlaşan ve derinleşen bir olgudur. Siber zorbalık, mağdurlar üzerinde yalnızca kısa vadeli psikolojik etkiler değil, uzun vadede sosyal izolasyon, travma, hatta intihar gibi ciddi sonuçlara neden olabilmektedir [1]. Anonimliğin sağladığı psikolojik rahatlık, bireylerin gerçek hayatta sergilemekten çekinecekleri saldırgan eylemleri dijital ortamda gerçekleştirmelerini kolaylaştırmakta, bu da zorbalığın hızla yayılmasına ve içselleştirilmesine yol açmaktadır. Ayrıca, mağdur bireylerin zamanla siber zorba haline gelmesi gibi döngüsel etkiler, bu olgunun toplumsal boyutunu daha da karmaşık hale getirmektedir.

Siber zorbalığın dijital metinler üzerinden otomatik olarak tespit edilmesi, hem erken müdahale hem de çevrim içi ortamların daha güvenli hale getirilmesi açısından oldukça önemlidir [2]. Ancak bu tür içeriklerin manuel olarak tespit edilmesi; büyük veri hacmi, dilsel çeşitlilik ve içeriklerin bağlama bağımlılığı gibi etkenler nedeniyle sürdürülebilir değildir. Bu noktada Doğal Dil İşleme (Natural Language Processing - NLP) ve Yapay Zeka (Artificial Intelligence - AI) tabanlı otomatik sınıflandırma sistemleri devreye girmektedir. Son yıllarda, NLP alanındaki gelişmeler sayesinde metin sınıflandırma, duygu analizi, nefret söylemi tespiti gibi görevlerde yüksek doğruluk oranlarına ulaşan birçok model geliştirilmiştir [3]. Bu alanda yapılan çalışmalarda, dil bağımsız çalışmaların kısıtlı olması [4] sebebiyle ele alınan dilin yaygınlığı önemli bir parametredir. Örneğin, İngilizce veriler üzerinde yapılan çalışmalarda birçok yapay zeka destekli modeller ve NLP teknikleri kullanılarak siber zorbalık tespiti üzerine başarılı sonuçlar elde edilmiştir [5-7]. Ancak, Türkçe gibi morfolojik olarak zengin ve kaynak açısından kısıtlı diller için yapılan, dijital metinlerden siber zorbalık tespiti çalışmaları oldukça sınırlıdır [8-10]. Türkçe'nin eklemeli yapısı, kelime türetme esnekliği ve bağlamdan bağımsız yapılar içermesi, özellikle veri temsili ve model öğrenmesi açısından önemli zorluklar doğurmaktadır [11]. Bu nedenle Türkçe dijital metinlerde siber zorbalık tespiti üzerine yapılacak çalışmalar, hem akademik literatüre katkı sağlamakta hem de yerel dildeki dijital güvenlik politikalarının oluşturulmasına temel teşkil etmektedir.

Bu çalışmada, Türkçe metinlerdeki siber zorbalık içeriklerinin tespiti amacıyla hem geleneksel makine öğrenmesi yöntemleri hem de derin öğrenme tabanlı modeller karşılaştırmalı olarak değerlendirilmiştir. TF-IDF ile temsil edilen veriler üzerinde geleneksel yöntemler (örneğin SVM ve Lojistik Regresyon) test edilmiş, derin öğrenme modellerinde ise FastText embedding (kelime gömme) kullanılarak GRU, BiGRU, CNN ve BiLSTM mimarileri uygulanmıştır. Elde edilen sonuçlar, özellikle BiLSTM ve BiGRU modellerinin %96 doğruluk oranına ulaşarak Türkçe metinlerde siber zorbalık tespiti için etkili birer araç olduğunu göstermektedir.

II. MATERYAL VE METOTLAR

A. Veri Seti

Çalışmamızda, Mikayil Sadigzade tarafından 2022 yılında derlenmiş, "Sosyal Medyada Sanal Zorbalığın Tespiti" isimli Yüksek Lisans tez çalışmasında kullanılan Türkçe veri seti kullanılmıştır [9, 12]. Veri seti "message" ve "cyberbullying" olmak üzere iki adet sütundan oluşmaktadır. Message sütunu incelenecek metni içermekte iken cyberbullying sütunu metnin zorbalık içerip içermediğini 0 ve 1 etiketleri ile belirtmektedir (1 zorbalık içeren, 0 içermeyen metinler için).

Doğal dil yapısının sağlıklı bir şekilde işlenebilmesi bakımından veri kümesinde bulunan metinler ilk olarak veri temizleme işlemine tabi tutulmuştur. NLP çalışmalarında, metinlerin semantik analizinin hatalardan arındırılması adına; noktalama işaretlerinden, emojilerden ve alfa-numerik olmayan karakterlerden arındırılması, büyük/küçük harf dönüşümleri gibi ön işlemler çalışmamızda kullanılan metinlere de uygulanmıştır. Ayrıca yine semantik analize katkısı olmayan Türkçe durak kelimeleri (stop words), NLTK (Natural Language Toolkit) kütüphanesi kullanılarak veri setindeki metinlerden çıkarılmıştır.

Temizlenmiş metinler yapay zeka modellerine girdi olarak verilmeden önce yine NLP çalışmalarında metinlerin anlamsal bağlamının daha kuvvetlendirilmesi amacıyla yapılan lemmatizasyon işlemi veri kümesindeki metinlere uygulanmıştır. Bu işlem için Türkçe metinlere özgü geliştirilen Stanza_TR [13] NLP aracı kullanılmıştır. Kullanılan araç kütüphanesi aracılığıyla, tokenize, pos, lemma işlem hattı tanımlanmış ve aşamalar sırayla gerçekleştirilmiştir. Veri kümesinde kullanılan metinlerin kelime dağılımı Şekil 1'de kelime bulutu olarak sunulmuştur.

Şekil 1. Kullanılan veri kümesinin kelime bulutu

Şekil 1'de bulunan kelimeler, veri setinde bulunan metinlerde geçme sıklıklarına göre farklı büyüklüklerde gösterilmektedir. Kullanılan veri seti 2200 adet Türkçe metin içermektedir. Metin etiketleri dengeli olarak dağılmıştır. Yapay zeka modellerine girdi olarak verilmeden önce son işlem olarak veri seti %80 eğitim, %20 test olmak üzere, python kütüphanesinden train_test_split fonksiyonu kullanılarak rastgele ayrılmıştır. Tablo 1'de veri setinin ayrım sonucu örnek sayıları sunulmuştur.

Tablo 1. Veri Dağılımı

Sınıf	Eğitim Verisi Sayısı	Test Verisi Sayısı
0	897	204
1	863	237

Veri seti ayrımı sonucu elde edilen eğitim seti, çalışmamızdaki yapay zeka modellerinin eğitimi için kullanılmakta olup, test seti ise eğitim sonrası üretilen modelin başarısını değerlendirmek amacı ile kullanılmaktadır. Kullanılan modellerin başarısını değerlendirmek için; doğruluk (accuracy), kesinlik (precision), duyarlılık (recall) ve F-skor ölçekleri kullanılmıştır.

B. *Geleneksel Makine Öğrenmesi Yöntemleri (Baseline Modeller)*

Makine öğrenmesi, sistemlerin açıkça kural tabanlı çözümlerinin yanı sıra verilerden örüntüler öğrenmesini ve bu öğrenme sonucunda kararlar almasını sağlayan bir yapay zeka yaklaşımıdır. Bu çalışmada, modern derin öğrenme yaklaşımlarının performansını değerlendirmek amacıyla, geleneksel makine öğrenmesi algoritmaları karşılaştırma (baseline) modelleri olarak kullanılmıştır. Bu tür bir karşılaştırma, ileri düzey modellerin ne derece anlamlı iyileştirmeler sağladığını görmek açısından oldukça değerlidir. Bu kapsamda, ilk olarak metin verileri Term Frequency–Inverse Document Frequency (TF-IDF) ile sayısal forma dönüştürülmüş ve Destek Vektör Makineleri (Support Vector Machines - SVM), Lojistik Regresyon (Logistic Regression - LR), K-En Yakın Komşuluk (K-Nearest Neighbor - KNN) ve Rastgele Orman (Random Forest - RF) sınıflandırıcı algoritmaları ile eğitim gerçekleştirilmiştir. Bu algoritmalar, sınıflandırma problemlerinde yaygın olarak kullanılan ve farklı karar mekanizmalarına dayanan yöntemlerdir. SVM, karar sınırlarını maksimum marj (sınır) ile belirlerken; LR, doğrusal ayrılabilirliği olasılıksal temelde modellemektedir. KNN, sınıflandırmayı komşu örneklerin etiketlerine göre yaparken; RF, birden fazla karar ağacının çoğunluk oyunu ile karar vermesini sağlar. Tablo 2, her bir geleneksel yöntem için kullanılan parametre yapılandırmalarını ve bu yapılandırmaların doğruluk üzerindeki etkisini detaylı şekilde sunmaktadır. Bu klasik yöntemler, hem temel performans eşiği sağlaması hem de derin öğrenme modellerine kıyasla daha açıklanabilir olmaları açısından araştırmada önemli bir yer tutmaktadır.

Tablo 2. Geleneksel Modellerde kullanılan parametreler

Model	Model Tipi	Parametreler
LR	Doğrusal	solver='lbfgs' penalty='l2' C=1.0 max_iter=100 class_weight=None
SVM	Doğrusal Olmayan	kernel='rbf' C=1.0 gamma='scale' class_weight=None
KNN	Doğrusal Olmayan	n_neighbors=5 metric='minkowski' (p=2) weights='uniform' algorithm='auto'
RF	Topluluk (Ensemble)	n_estimators=100 criterion='gini' max_depth=None bootstrap=True class_weight=None

Tablo 2'de verilen parametreler, kullanılan modelin varsayılan değerleri olup bazı parametrelerde ele alınan probleme özgü ayarlamalar yapılmıştır. Parametre ayarlamaları aşağıda detaylandırılmıştır.

- Lojistik Regresyon (LR) – Doğrusal Model
  - solver='lbfgs': Optimizasyon problemi için L-BFGS algoritması kullanılmıştır (küçük ve orta ölçekli veri kümeleri için uygundur).
  - penalty='l2': Aşırı öğrenmeyi engellemek amacıyla L2 (ridge) düzenleme uygulanmıştır.
  - C=1.0: Düzenleme katsayısı; daha düşük değerler daha güçlü düzenleme (ceza) anlamına gelir.
  - max_iter=100: Modelin öğrenme sürecinde en fazla 100 yineleme (iterasyon) yapmasına izin verilmiştir.
  - class_weight=None: Sınıflar arasında ağırlıklandırma yapılmamıştır; tüm sınıflar eşit önemdedir.

- Destek Vektör Makineleri (SVM) – Doğrusal Olmayan Model
  - kernel='rbf': Karar sınırlarını doğrusal olmayan biçimde ayırmak için RBF (Radial Basis Function) çekirdek fonksiyonu kullanılmıştır.
  - C=1.0: Hata toleransı ile margin genişliği arasındaki dengeyi belirleyen ceza parametresidir.
  - gamma='scale': RBF çekirdeği için gama değeri otomatik olarak hesaplanmıştır (1 / (n_features * X.var())).
  - class_weight=None: Sınıflar arası dengesizlik dikkate alınmamıştır.

- K-En Yakın Komşu (KNN) – Doğrusal Olmayan Model
  - n_neighbors=5: Sınıflandırma sırasında en yakın 5 komşu örnek dikkate alınmıştır.
  - metric='minkowski' (p=2): Minkowski mesafesi kullanılmıştır; p=2 değeri ile bu, Öklidyen mesafeye eşdeğerdir.
  - weights='uniform': Tüm komşular eşit ağırlıkta değerlendirilmiştir.
  - algorithm='auto': Uygun algoritma veri setine göre otomatik olarak seçilmiştir (ball_tree, kd_tree veya brute-force olabilir).

- Rastgele Orman (RF) – Topluluk Model
  - n_estimators=100: Model 100 adet karar ağacından oluşan bir orman şeklinde yapılandırılmıştır.
  - criterion='gini': Ağaçların bölünme kararlarında Gini saflık ölçütü kullanılmıştır.
  - max_depth=None: Ağaçların maksimum derinliği sınırlanmamıştır; gerektiği kadar derinleşebilirler.
  - bootstrap=True: Her ağaç için örnekler rastgele seçilerek (yeniden örneklemeyle) eğitilmiştir (bootstrap örnekleme).
  - class_weight=None: Tüm sınıflar eşit ağırlıklı değerlendirilmiştir.

C. *Derin Öğrenme Tabanlı Modeller*

Derin öğrenme modelleri, özellikle bağlam ilişkilerini daha iyi kavrayabilme yetenekleri nedeniyle çalışmanın merkezini oluşturmaktadır. NLP problemlerinde, kelimeler arasındaki sıralı bağıntıları, anlamsal geçişleri ve uzun vadeli ilişkileri yakalayabilen mimariler, geleneksel makine öğrenmesi tekniklerine kıyasla çok daha başarılı sonuçlar verebilmektedir. Bu kapsamda; Gated Recurrent Unit (GRU), Bidirectional GRU (BiGRU), Convolutional Neural Network (CNN) ve Bidirectional Long Short-Term Memory (BiLSTM) mimarileri uygulanmıştır. GRU ve LSTM gibi tekrarlayan

sinir ağı (RNN) temelli modeller, dizisel verilerdeki zaman bağımlılığını modelleyebilmek için geliştirilmiştir. GRU, LSTM'nin daha sadeleştirilmiş bir versiyonu olup benzer performansı daha düşük hesaplama maliyetiyle sağlayabilmektedir. Bu modellerin çift yönlü versiyonları olan BiGRU ve BiLSTM, metin dizisini hem ileri hem geri yönde işleyerek bağlam bilgisini çift yönlü olarak yakalamayı mümkün kılar. Böylece özellikle cümlenin sonundaki bir kelimenin, baştaki bir kelimeye etkisinin öğrenilmesi sağlanır. Diğer yandan, CNN mimarisi genellikle görüntü işleme ile ilişkilendirilse de, yerel özellikleri öğrenme yeteneği sayesinde metin sınıflandırmada da etkili bir yöntemdir. Özellikle kısa metinlerdeki belirgin örüntüleri tespit etme kabiliyeti ile metinlerdeki anlam kümelerini başarıyla modelleyebilir.

Bu modellerde girdi olarak, önceden eğitilmiş Türkçe FastText embedding vektörleri kullanılmıştır. Embedding katmanı ile dilin semantik yapısı modele aktarılmış, bu sayede bağlam bilgisi korunarak daha doğru sınıflandırma yapılması hedeflenmiştir. Tablo 3, 4, 5 ve 6 sırasıyla kullanılan BiLSTM, GRU, BiGRU ve CNN modellerinin mimari yapısı ve özelliklerini içermektedir.

Tablo 3. BiLSTM modelinin özellikleri

Özellik	**Değer**
Embedding Dimension	100
Maksimum Sekans Uzunluğu	100
BILSTM gizli birim sayısı	128 (Çift Yönlü)
Çıkış Katmanı	Dense 1, Sigmoid Act.
Kayıp Fonksiyonu	Binary Crossentropy
Optimizasyon Algoritması	Adam
Epoch	10
Batch size	32

Tablo 4. GRU Modelinin Özellikleri

Özellik	**Değer**
Embedding Dimension	100
Maksimum Sekans Uzunluğu	100
GRU gizli birim sayısı	128
Çıkış Katmanı	Dense 1, Sigmoid Act.
Kayıp Fonksiyonu	Binary Crossentropy
Optimizasyon Algoritması	Adam
Epoch	10
Batch size	32

Tablo 5. BiGRU Modelinin özellikleri

Özellik	**Değer**
Embedding Dimension	100
Maksimum Sekans Uzunluğu	100
BIGRU gizli birim sayısı	128 (Çift Yönlü)
Çıkış Katmanı	Dense 1, Sigmoid Act.
Kayıp Fonksiyonu	Binary Crossentropy
Optimizasyon Algoritması	Adam
Epoch	10
Batch size	32

Tablo 6. CNN Modelinin Özellikleri

Özellik	**Değer**
Embedding Dimension	100
Maksimum Sekans Uzunluğu	100
CNN Filtre Sayısı (Conv1d)	128
Çıkış Katmanı	Dense 1, Sigmoid Act.
Kayıp Fonksiyonu	Binary Crossentropy
Optimizasyon Algoritması	Adam
Epoch	10
Batch size	32

D. Eğitim ve Değerlendirme

Tüm modeller, eğitim ve test setlerine ayrılan veri kümesi üzerinde 5 katlı çapraz doğrulama (5-fold cross validation) yöntemiyle değerlendirilmiştir. Model performansları doğruluk, kesinlik, duyarlılık ve F-skor ölçütleri ile karşılaştırılmıştır.

III. DENEYSEL SONUÇLAR

Çalışmada kullanılan modeller belirtilen metrikler yardımı ile karşılaştırmalı olarak değerlendirilmiştir. Geleneksel makine öğrenmesi yöntemleri kullanılarak, 5 katlı çapraz doğrulama sonucunda elde edilen ortalama sonuçlar Tablo 7'de sunulmuştur.

Tablo 7. Geleneksel Makine Öğrenmesi Model Performansları

Model	**Doğruluk**	**Kesinlik**	**Duyarlılık**	**F-skor**
LR	0.789	0.79	0.79	0.79
SVM	0.780	0.79	0.80	0.78
KNN	0.780	0.78	0.78	0.78
RF	0.757	0.76	0.77	0.75

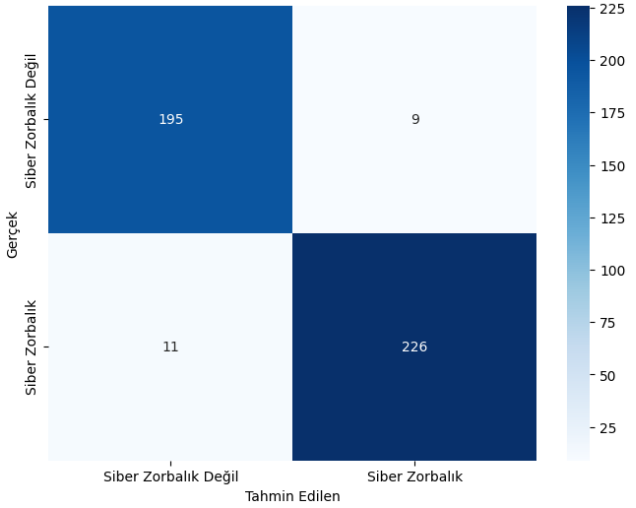
Geleneksel modeller 5 katlı çapraz doğrulama ile gerçekleştirilmiş ve tüm modeller için varyans < 0,001, standart_sapma < 0,1 olarak elde edilmiştir. Modeller arasında en iyi performansı TF-IDF + LR kombinasyonu göstermiş olmasına rağmen, kullanılan diğer yöntemlerden de benzer performans oranları elde edilmiştir.

Derin öğrenme modelleri, güncel çalışmalarda geleneksel yöntemlere kıyasla daha çok kaynak kullanımının yanı sıra daha iyi sonuçlar da sunmaktadır. Çalışmamız kapsamında kullanılan derin öğrenme modellerinden elde edilen sonuçlar Tablo 8'de sunulmuştur.

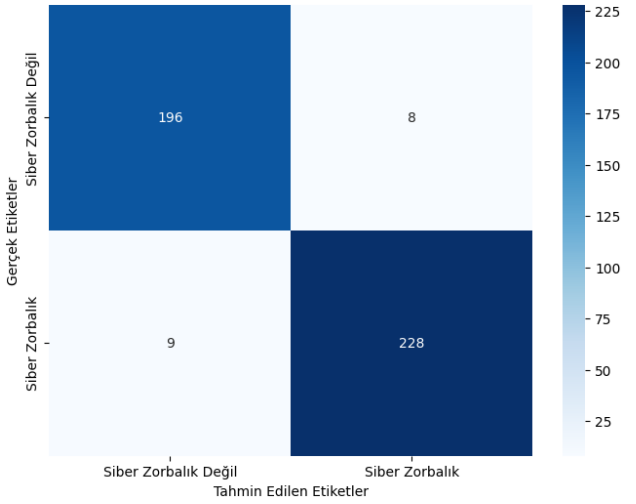
Tablo 8. Derin Öğrenme Modelleri Performansları

Model	Doğruluk	Kesinlik	Duyarlılık	F-skor
GRU	0.94	0.94	0.94	0.94
BiGRU	0.96	0.96	0.96	0.96
CNN	0.95	0.95	0.95	0.95
BiLSTM	0.96	0.96	0.97	0.96

Derin öğrenme tabanlı modeller genel olarak geleneksel modellere kıyasla belirgin şekilde daha yüksek performans sergilemiştir. Derin öğrenme modelleri arasında en yüksek performans BiLSTM modeli ile 0.96 olarak elde edilmiştir. BiLSTM ve BiGRU modelleri ile elde edilen karmaşıklık matrisleri sırası ile Şekil 2 ve Şekil 3'te sunulmuştur.

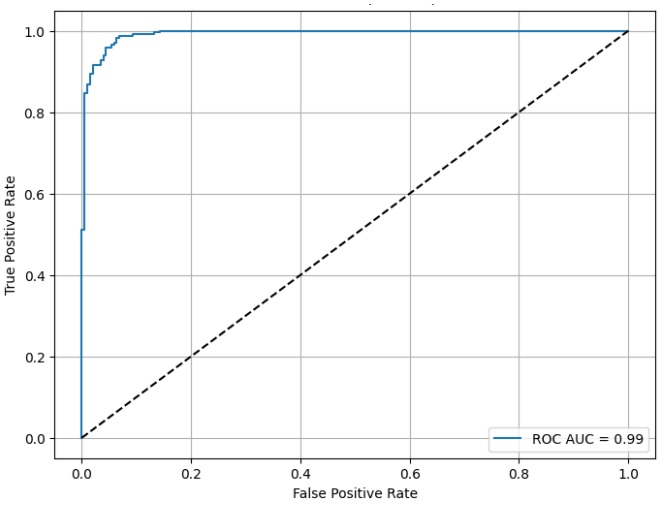


Şekil 2. BiLSTM modeli ile elde edilen karmaşıklık matrisi

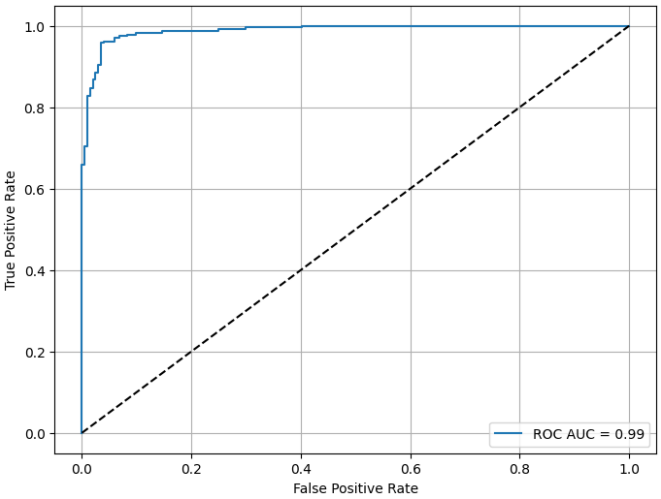


Şekil 3. BiGRU modeli ile elde edilen karmaşıklık matrisi

Birbirine yakın yüksek sonuçlar üreten BiLSTM ve BiGRU modellerinin farkının daha belirgin anlaşılabilmesi için, BiLSTM modelinin ROC-AUC (Receiver Operating Characteristic - Area Under Curve) eğrileri ve BiGRU modelinin ROC-AUC eğrileri sırayla Şekil 4 ve Şekil 5'te verilmiştir.



Şekil 4. BiLSTM modeli ile elde edilen ROC-AUC eğrisi



Şekil 5. BiGRU modeli ile elde edilen ROC-AUC eğrisi

Şekil 4 ve Şekil 5'te verilen ROC-AUC eğrilerine bakıldığında BiLSTM modelinin BiGRU modeline göre çok küçük bir farkla, metinlerde siber zorbalık tespiti üzerine daha iyi performans gösterdiği görülebilmektedir.

Modeller arası F-skor ayrıntı analizi için modellerin Makro/Mikro skorlama değerleri karşılaştırılmıştır. BiLSTm modeli için Makro F1 değeri 0,9612 iken Mikro F1 değeri 0,9615 olarak elde edilmiştir. Aynı metrikler ile BiGRU modelinin sonuçları, Makro F1 için 0,9564, Mikro F1 için 0,9566 olmaktadır.

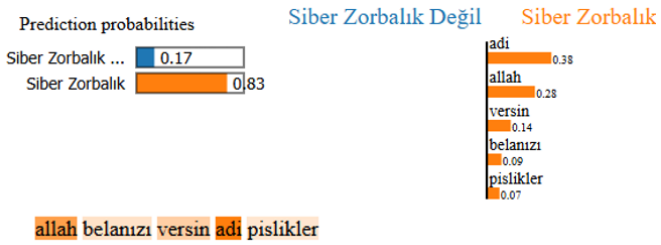
Çalışma kapsamında yapılan tüm uygulamalar ve geliştirilen tüm modeler Google Colab platformu kullanılarak gerçekleştirilmiştir. Derin öğrenme modelleri için platform işletim sistemi Linux 6.1.123+, işlemcisi x86_64 ve grafik işlemcisi T4 GPU olarak ayarlanmıştır. Bu ayarlamalar içerisinde BiGRU modelinin eğitim süresi 10,19 saniye, test süresi ile 0,67 saniye sürmektedir. Aynı ayarlamalar içerisinde BiLSTM modelinin eğitim süresi 11.75 saniye, test süresi ise 0.67 saniye sürmektedir.

Yapılan çalışmalar sonucunda geleneksel yöntemlerde kabul edilebilir başarı oranları alınırken, derin öğrenme modellerinde oldukça iyi bir performans elde edilmiştir. Çalışmanın sonunda makine öğrenmesi yolu ile zorbalığın

tespitinin mümkün olduğu, daha farklı veri kümleleri kullanılarak başarının daha da arttırılabileceği görülmüştür.

IV. MODEL AÇIKLANABİLİRLİĞİ VE YORUMLANABİLİRLİK ANALİZİ

Elde edilen sonuçlar, Türkçe gibi eklemeli ve morfolojik olarak zengin dillerde bile derin öğrenme tabanlı doğal dil işleme modellerinin oldukça başarılı sonuçlar verebileceğini göstermektedir. Özellikle bağlamı iki yönlü değerlendirebilen BiGRU ve BiLSTM gibi modeller, metin içindeki anlamsal ilişkileri daha iyi öğrenerek siber zorbalık ifadelerini daha doğru sınıflandırabilmektedir. Derin öğrenme modelleri yüksek doğruluk oranlarına ulaşsalar da, karar mekanizmalarının şeffaf olmaması bu modellerin "kara kutu" olarak değerlendirilmesine yol açmaktadır. Bu durum özellikle etik, adil ve güvenilir yapay zeka uygulamaları açısından önemli bir sınırlılık oluşturmaktadır. Bu bağlamda, çalışmada en yüksek performansı gösteren BiLSTM ve BiGRU modellerine yönelik olarak Local Interpretable Model-agnostic Explanations (LIME) [14] yöntemiyle açıklanabilirlik analizi gerçekleştirilmiştir. LIME, modelden bağımsız bir şekilde çalışarak, sınıflandırma kararlarını etkileyen kelimeleri yerel düzeyde analiz eder ve bu sayede modelin belirli bir tahmini neden yaptığına dair sezgisel yorumlar sunar. Kullanılan veri kümesinde bulunan siber zorbalık içeren ve siber zorbalık olarak etiketlenmeyen örnek metinlerin açıklanabilirliği Şekil 6 ve Şekil 7'de sunulmuştur.



Şekil 6. Siber zorbalık içeren metnin açıklanabilirliği

Uygulanan analizlerde, LIME aracı yardımıyla metinlerde hangi kelime veya kelime gruplarının siber zorbalık sınıfına atanmasında belirleyici rol oynadığı görselleştirilmiş ve modelin davranışları daha anlaşılır hale getirilmiştir. Bu sayede, modellerin yalnızca yüksek doğruluk elde etmeleri değil, aynı zamanda insan yorumuna açık karar mantıklarıyla güvenilir bir biçimde kullanılabilir hale gelmeleri amaçlanmıştır. Elde edilen bulgular, LIME'ın hem model doğruluğunu hem de kullanıcı güvenini destekleyecek şekilde etkili bir açıklama aracı olduğunu göstermiştir.

Şekil 7. Siber zorbalık değil etiketli metnin açıklanabilirliği

V. BULGULAR VE ÇIKARIMLAR

Bu çalışmada, Türkçe metinlerde siber zorbalık içeriklerinin otomatik olarak tespiti amacıyla geleneksel makine öğrenmesi yöntemleri ile modern derin öğrenme mimarileri karşılaştırmalı olarak incelenmiştir. Geleneksel yöntemlerde TF-IDF vektörleştirme tekniği ile Logistic Regression modeli en iyi sonucu %78.9 doğruluk oranı ile kabul edilebilir bir zemin sunarken, derin öğrenme tabanlı modeller bu performansı önemli ölçüde aşmıştır. Özellikle BiLSTM ve BiGRU mimarileri, önceden eğitilmiş Türkçe FastText kelime gömme vektörleri ile birlikte kullanıldığında %96 doğruluk oranına ulaşarak en başarılı modeller olarak öne çıkmıştır.

Elde edilen bulgular, bağlam bilgisine duyarlı mimarilerin, Türkçe gibi morfolojik olarak karmaşık dillerde anlamlı içerik sınıflandırmasında daha etkili olduğunu ortaya koymaktadır. Ayrıca, LIME aracı kullanılarak yapılan açıklanabilirlik analizleri sayesinde, bu yüksek doğruluğa sahip modellerin karar alma süreçleri görselleştirilmiş ve yorumlanabilir hale getirilmiştir. Bu durum, derin öğrenme tabanlı yaklaşımların yalnızca performans açısından değil, etik ve güvenilirlik boyutları açısından da uygulamaya uygun olduğunu göstermektedir.

Sonuç olarak, çalışmada geliştirilen yöntemlerin, Türkçe sosyal medya verilerinde siber zorbalık içeriklerinin erken ve doğru bir biçimde tespitinde etkili bir çözüm sunduğu ortaya konmuştur. Gelecek çalışmalarda, daha büyük ve çeşitli veri kümeleri ile transformer tabanlı modellerin entegrasyonu, Türkçe'de siber zorbalıkla mücadelede daha da güçlü sistemlerin geliştirilmesini mümkün kılacaktır. Özellikle dijital oyun platformlarında otomatik siber zorbalık filtreleme sistemlerinin geliştirilmesi için bu tür modellerin etkili bileşenler olarak kullanılabileceği değerlendirilmektedir.

REFERANSLAR

[1] Aktepe, E. (2013, February). Ergenlerde Siber Zorbalık ve Siber Mağduriyet. In Yeni Symposium (Vol. 51, No. 1).

[2] Akca, E. B., & Sayımer, İ. (2017). Siber zorbalık kavramı, türleri ve ilişkili olduğu faktörler: Mevcut araştırmalar üzerinden bir değerlendirme. AJIT-e: Academic Journal of Information Technology, 8(30), 7-19.

[3] Canbay, P., & Ekinci, E. (2023). Derin ve Sığ Makine Öğrenmesi Yöntemleri ile Türkçe Tweetlerden Saldırgan Dil Tespiti. Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, 16(1), 1-10.

[4] Canbay, P., & Bölücü, N. (2022, May). A Language-Free hate speech identification on code-mixed conversational tweets. In The International Conference on Artificial Intelligence and Applied Mathematics in Engineering (pp. 102-108). Cham: Springer International Publishing.

[5] Nadali, S., Murad, M. A. A., Sharef, N. M., Mustapha, A., & Shojaee, S. (2013, December). A review of cyberbullying detection: An overview. In 2013 13th international conference on intellient systems design and applications (pp. 325-330). IEEE.

[6] Iwendi, C., Srivastava, G., Khan, S., & Maddikunta, P. K. R. (2023). Cyberbullying detection solutions based on deep learning architectures. Multimedia Systems, 29(3), 1839-1852.

[7] Rosa, H., Pereira, N., Ribeiro, R., Ferreira, P. C., Carvalho, J. P., Oliveira, S., ... & Trancoso, I. (2019). Automatic cyberbullying detection: A systematic review. Computers in Human Behavior, 93, 333-345.

[8] Bozyiğit, A., Utku, S., & Nasibov, E. (2021). Cyberbullying detection: Utilizing social media features. Expert Systems with Applications, 179, 115001.

[9] S.M. Sadigzade, "Sosyal medyada sanal zorbalığın tespiti," Master's thesis, Dokuz Eylul Universitesi (Turkey), 2022.

[10] Özel, S. A., Saraç, E., Akdemir, S. & Aksu, H. (2017). Detection of cyberbullying on social media messages in turkish. In 2017

International conference on computer science and engineering (UBMK) (pp. 366–370). IEEE.

[11] Oflazer, K. (2016). Türkçe ve doğal dil işleme. Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, 5(2).

[12] Sadigzade, M., & Nasibov, E. (2022). Comparatıve Analysıs Of Count Vectorızatıon Vs Tf-Idf Vectorızatıon For Detectıng Cyberbullyıng In Turkısh Twıtter Messages. Journal of Modern Technology & Engineering, 7(1).

[13] Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. arXiv preprint arXiv:2003.07082.

[14] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).