SETSCI conference proceedings
SETSCI Conference Proceedings
15 , 45-50, 2023

Cognitive Models and Artificial Intelligence Conference
26-28 October 2023, Ankara, Türkiye

https://doi.org/10.36287/setsci.6.1.019
2687-5527/ 

BMYZ 2023

# ELECTRA ve XLNET Modellerini Kullanarak X Verilerinden İntihara Meyilli İçerikleri Tespit Etme

Tolga Aydın[1], Muhammed Coşkun Irmak[2*]

[1]Bilgisayar Mühendisliği Bölümü, Atatürk Üniversitesi, Erzurum, Türkiye
[2] Bilgisayar Mühendisliği Bölümü, Van Yüzüncü Yıl Üniversitesi, Van, Türkiye
[*](coskunirmak@yyu.edu.tr)

***Özet*** **–** İntihar konusu, birçok farklı disiplin tarafından incelenen ve tarihin her döneminde karşılaşılan bir konu olmuştur. Teknolojinin gelişimi ve akıllı telefonların her yerde bulunması sebebiyle, kullanıcılar, özellikle çevrimiçi sosyal platformlar aracılığıyla duygu ve düşüncelerini rahatlıkla ifade edebilmişlerdir. Paylaşılan bu düşüncelerden biri de intihar düşüncesidir. İntiharın önlenebilmesi ve genç popülasyonda artan intihar oranlarının kontrol altına alınabilmesi için son zamanlarda sosyal medyada paylaşılan intihar düşüncelerinin tespiti üzerine yapılan araştırmaların sayısı her geçen gün artmaktadır. Bu çalışmada, en popüler sosyal medya platformlarından birisi olan X'te paylaşılan gönderilerden hazırlanan bir veri seti üzerinde makine öğrenmesi ve doğal dil işleme teknikleri kullanılarak bu paylaşımı yapan kişilerin intihara meyilli olup olmadığının tespitinin yapılmasına odaklanılmıştır. Önceden eğitilmiş transformatör yöntemlerinin klasik makine öğrenmesi yöntemleri ile karşılaştırmalı bir analizi sunulmuştur. Tüm modeller içerisinde en yüksek doğruluk oranı, ELECTRA modeli ile %96.61 olarak elde edilmiştir. Bu bulgular, intihara meyilli tweetlerin tespitinde transformatör modellerinin potansiyelinin yüksek olduğunu göstermekte ve özellikle ELECTRA'nın bu alanda etkili bir araç olabileceğini ortaya koymaktadır.

*Anahtar Kelimeler – İntihar tespiti, Doğal dil işleme, ELECTRA, XLNET, Transformatör model*

## I. GİRİŞ

Günümüzde, sosyal medya platformları, milyonlarca insanın düşüncelerini, duygularını ve deneyimlerini paylaştığı güçlü bir iletişim aracı haline gelmiştir. Ancak bu dijital platformlar, sadece pozitif deneyimleri değil, aynı zamanda zihinsel sağlık sorunlarının da izlerini taşımaktadır. İnternetin derinliklerinde, insanların kendilerini ifade etmek için kullandıkları bu alanlarda, intihar düşünceleri gibi ciddi zihinsel sağlık sorunlarının da izleri bulunmaktadır.

İntihar, toplumun karşı karşıya olduğu en ciddi sağlık sorunlarından biri haline gelmiştir. Özellikle genç ve orta yaşlılar arasında intihar düşünceleri ve eylemleri endişe verici bir artış göstermektedir. Dünya Sağlık Örgütü (World Health Organization – WHO) istatistiklerine göre, her yıl yaklaşık bir milyon kişi intihar nedeniyle hayatını kaybetmektedir ve ortalama olarak her 40 saniyede bir intihar gerçekleşmektedir [1]. WHO, intiharın gençler arasında ölümün ana nedeni ve yetişkinler arasında altıncı önde gelen neden olduğunu belirtmektedir. Amerikan İntiharı Önleme Vakfı (American Foundation for Suicide Prevention – AFSP) intiharla ilişkilendirilen çeşitli risk faktörlerini tanımlamıştır [1]. Bu faktörler arasında umutsuzluk, zararlı madde kullanımı, anksiyete, şizofreni gibi kişisel sorunlar; toplumdan izolasyon, sevdiklerinin kaybı, işsizlik, zorbalık veya kötü muamele gibi sosyal faktörler; ya da hastalık, duygusal bozukluklar ve önceki intihar girişimleri gibi olumsuz olaylarla ilgili faktörler yer almaktadır. İntihar düşünceleri, sıkça zorlu yaşam koşulları, psikolojik sıkıntılar ve duygusal çalkantılarla ilişkilendirilirken, bu düşüncelerin erken teşhisi ve yardım sağlanması hayati önem taşımaktadır [1, 2]. Bu olumsuzluklara rağmen, teknoloji, bu soruna dikkat çekme ve potansiyel intihar eylemlerini önceden belirleme konusunda umut verici bir rol oynamaktadır.

Yapılan çalışmalarda, Makine Öğrenimi (Machine Learning – ML) ve Doğal Dil İşleme (Natural Language Processing – NLP) tekniklerinin, sosyal medya verilerini analiz ederek intihar düşüncelerini tespit edebileceği gösterilmiştir [1-10]. Özellikle Derin Öğrenme (Deep Learning – DL) yapıları, bu konuda oldukça başarılı sonuçlar elde etmekte ve bu alandaki potansiyelini ortaya koymaktadır [2-4, 9, 10]. Ancak bu tekniklerin başarısını garantilemek için doğru veri setlerini oluşturmak ve etiketlemek büyük önem taşımaktadır. Bu noktada, sosyal medya platformları, intihar düşüncelerini ve risklerini belirlemek için bir veri seti oluşturmada güçlü bir kaynak olmaktadır. Kullanıcılar, bu dijital alanlarda duygularını ve düşüncelerini serbestçe ifade etmektedir, ancak bu ifadeler sıkça gözden kaçmakta veya görmezden gelinmektedir [5, 6].

Bu makale, günümüzde en çok kullanılan sosyal medya platformlarından birisi olan X (Twitter)'te paylaşılan içerikleri analiz ederek intihar düşüncelerini tespit etmeye odaklanan bir araştırmanın sonuçlarını sunmaktadır. Bu çalışmada, kullanılan makine öğrenme ve veri madenciliği teknikleri sayesinde, intihar riski taşıyan içeriklerin belirlenmesi ve erken müdahale imkânı sağlanmaktadır. Ayrıca, bu çalışma, zihinsel sağlık sorunlarını daha iyi anlamak ve önlemek için sosyal medya platformlarının potansiyelini keşfetmeye yönelik bir adım olarak da görülmektedir.

Bu çalışmada, intihara meyilli tweetleri tespit etmede, güçlü transformatör tabanlı makine öğrenme yöntemlerinden ELECTRA ve XLNET'in, klasik makine öğrenmesi yöntemleri ile karşılaştırmalı bir performans analizine yer

verilmiştir. ELECTRA, var olan metinleri daha iyi anlamak ve anlam değişimlerini yakalamak için tasarlanmış yenilikçi bir modeldir. XLNET ise dil modellemesi konusundaki sınırlamaları aşarak daha iyi sonuçlar elde etmek için kendinden önceki modellerden farklı bir yaklaşım benimsemiştir.

Çalışmanın geri kalan kısmı şu şekilde organize edilmiştir: Bölüm 2'de, literatürdeki intihara meyilli tweetlerin tespiti alanında yapılan çalışmalar sunulmuştur. Bölüm 3'te çalışmada kullanılan veri seti ve metodoloji detaylı bir şekilde anlatılmıştır. Bölüm 4'te deneysel sonuçlar verilmiş, ardından bu sonuçlar tartışılmıştır.

## II. LİTERATÜR İNCELEMESİ

Bu bölümde, intihara meyilli paylaşımların tespiti üzerine literatürde yer alan güncel çalışmalara yer verilmiştir.

Haque ve ark. (2022) çalışmalarında, sosyal medya platformlarından biri olan twitter verilerini kullanarak intihar düşüncelerinin erkenden tespiti için makine öğrenmesi (Random Forest – RF, Support Vector Machine – SVM, Stochastic Gradient Descent, Logistic Regression – LR ve Naive Bayes) ve derin öğrenme modellerinin (Long Short-Term Memory – LSTM, Bidirectional LSTM - BiLSTM, Gated Recurrent Unit – GRU, Bidirectional GRU ve Convolutional Neural Network – CNN ve LSTM'in birleştirilmiş modelleri) karşılaştırmalı analizini sunmuşlardır. Çalışmada, intihar düşüncesinin yüksek doğrulukla ve erkenden tespit edilebilmesi için önceki araştırma çalışmalarındaki sonuçlardan daha iyi bir performans değeri elde edilebilmesi amaçlanmıştır. Özellikle metin ön işleme ve öznitelik çıkarma aşamalarına odaklanılmıştır. On sekiz anahtar kelime dikkate alınarak intihar niteliğinde olan ve olmayan toplamda 49178 örneği içeren bir veri kümesi üzerinde deneysel çalışmalar yapılmıştır. Yapılan çalışmalarda en yüksek sınıflandırma performansına %93 doğruluk değeri ile RF modeli kullanılarak ulaşılmıştır. Ayrıca, derin öğrenme modellerinin kelime yerleştirme ile eğitilmesi sonucunda BiLSTM modelinin %93.6 doğruluğa ulaştığı görülmüştür [2].

Baghdadi ve ark. (2022) çalışmalarında, Arapça olarak paylaşılan tweetlerden oluşan bir veri seti hazırlamışlardır. Hazırlanan bu veri setinde 1074 normal ve 956 intihara meyilli olmak üzere toplamda 2030 tweet yer almaktadır. Ayrıca çalışmada, Arapça tweetlerin ön işlemesinin gerçekleştirilmesi için yeni bir algoritma önerilmiştir. Model eğitiminde ise yaygın olarak kullanılan transformatör yöntemlerinden olan Bidirectional Encoder Representations from Transformers – BERT ve Universal Sentence Encoder – USE kullanılmıştır. Eğitilen modeller dengeli doğruluk, özgüllük, F1 puanı, IoU, ROC, Youden İndeksi, NPV ve ağırlıklı toplam ölçüt (weighted sum metric – WSM) performans metriklerine göre karşılaştırılmıştır. USE modelleri için, en iyi WSM değeri %80.2 iken BERT modelleri ile en iyi WSM değeri %95.26 olarak bulunmuştur [3].

Abdulsalam ve ark. (2022) çalışmalarında, Arapça tweetlerden intihar düşüncelerini tespit edebilmek için, Ağustos 2021 – Nisan 2022 tarihleri arasında atılan toplam 5719 tweeti toplayarak yeni bir Arapça intihar tweetleri veri kümesi geliştirmişlerdir. Model eğitimlerinde NB, SVM, K-En Yakın Komşu (k-Nearest Neighbors – KNN), RF ve XGBoost makine öğrenimi modellerini ve AraBERT, AraELECTRA ve AraGPT2 önceden eğitilmiş derin öğrenme modellerini kullanmışlardır. Sonuçlar, karakter n-gram özelliklerinde eğitilen SVM ve RF modellerinin, makine öğrenimi modelleri arasında %86 doğruluk ve %79 F1 puanı ile en iyi performansı sağladığını göstermektedir. Derin öğrenme modellerinin sonuçları, AraBERT modelinin diğer makine ve derin öğrenme modellerini geride bıraktığını, Arapça tweetler veri kümesinde intihar düşüncelerinin tespitini önemli ölçüde geliştirerek %91 doğruluk ve %88 F1 puanı elde ettiğini göstermektedir [4].

Chatterjee ve ark. (2022) çalışmasında, Reddit ve Twitter'da intihar düşünceleri olan iyi etiketlenmiş bir veri kümesi oluşturulması amaçlanmıştır. Sadece klinik intihar belirtileri değil, aynı zamanda sosyal medyada çevrimiçi davranışları da içeren altı özellik grubu tanımlanmıştır. Bu özellik gruplarını kullanarak sosyal medyada intihar düşüncelerini tanımlamak için çok modlu bir model önerilmiştir. Önerilen modeller içerisinde en yüksek doğruluk değeri %87 olarak LR ile elde edilmiştir [7].

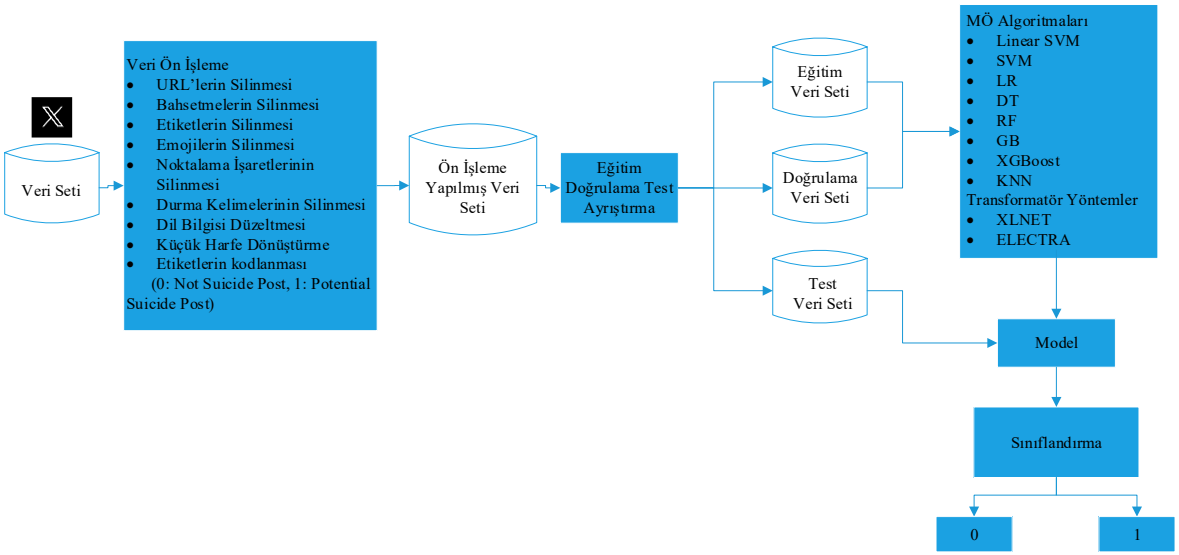
Sabri ve Mohamad (2022) çalışmasında, intihar düşüncesi olan tweetleri tespit etmek için Naive Bayes algoritmasının yeteneği araştırılmıştır. Twitter verileri, Mayıs 2021'deki Malezya'nın pandemi kilitlenmesi sırasında "stres", "anksiyete", "depresyon" ve "intihar" anahtar kelimeleri temel alınarak Tweepy kütüphanesi kullanılarak çekilmiştir. Değerlendirme sonuçları, algoritmanın intihar içerikli tweetleri tespit etmede %80.39 doğrulukla başarılı ve kabul edilebilir bir performans gösterdiğini ortaya koymuştur [8].

Deepa ve ark. (2023) çalışmasında, tweet verisini değerlendirerek twitter kullanıcı profillerindeki intihar derecesini otomatik olarak erken tespit etmeyi amaçlamışlardır. 26 Aralık 2022 – 10 Ocak 2023 tarihleri arasında atılan tweetleri, intiharla ilgili bir dizi ifade ve terimler ile izlemişlerdir. Kriterlere uyan tweetler tespit edildiğinde, bu kullanıcıların zaman çizelgesinden diğer tweetleri de toplamışlardır. Bu süre zarfında 529 kullanıcıdan 16820 tweet toplanmıştır. Bu tweetlerin %40'ı rastgele seçilip ve manuel olarak '1' intihar içeren ve '0' intihar içermeyen tweetler olarak etiketlenmiştir, geri kalanı modeli test etmek için kullanılmıştır. Veri üzerinde bir LSTM modeli uygulanmış ve intihar içeren tweetleri intihar içermeyen tweetlerden ayırmada %90.8 doğruluk elde edilmiştir. Ayrıca bu tweetler üzerine bir makine öğrenimi modeli daha uygulanmış ve 'Derinlemesine Rahatsız Edici', 'Muhtemelen Rahatsız Edici' ve 'Göz Ardı Edilebilir' olmak üzere üç derece verilmiştir [9].

Priyamvada ve ark. (2023) çalışmasında ise intihara meyilli tweetlerin tespiti için makine öğrenmesi ve derin öğrenme mimarilerinden faydalanılmıştır. Stacked CNN – 2 katmanlı LSTM modelini kullanmışlardır. Stacked CNN – 2 katmanlı LSTM mimarisi kelime gömme teknikleriyle önceki CNN – LSTM yaklaşımlarına kıyasla %93.92 sınıflandırma doğruluğu elde etmiştir [10].

## III. MATERYAL AND METOT

Bu başlıkta, çalışmada kullanılan veri setinin içeriği ve yapılan ön işleme çalışmalarına değinilmiştir. Ayrıca model eğitiminde kullanılan makine öğrenmesi ve transformatör yöntemler hakkında detaylı bilgi verilmiştir. Çalışmanın akış şeması Şekil 1'de gösterilmiştir.




Şekil 1 İntihara meyilli tweet tespitinin genel akış diyagramı

### *A. Veri Seti*

Bu çalışmada, intihara meyilli tweetlerin tespiti için, Syeda Aunanya Mahmud tarafından X sosyal medya platformu kullanılarak hazırlanan ve Kaggle üzerinde açık erişimli paylaşılan Suicidal Tweet Detection Dataset (STDD) olarak adlandırılan veri seti kullanılmıştır. Bu veri seti 2 sınıf etiketi (not suicide post ve Potential suicide post) ve 1787 etiketli tweetten oluşmaktadır [11]. Veri setinin güncel tweetleri içermesi ve tweet sayısının literatürdeki diğer çalışmalarda kullanılan veri setlerinden daha az olması çalışmada bu veri setinin kullanılmasının sebeplerindendir. Literatürdeki çalışmalar incelendiğinde çok sayıda veri ile doğruluk değerinin iyileştirilmesine odaklandığı görülmüştür. Bu çalışmada ELECTRA ve XLNET modellerinin az sayıda veri üzerindeki yüksek başarılarını göstermek amaçlanmaktadır.

### *B. Veri Ön İşleme*

Yapılan ön işlemler;

-URL'lerin silinmesi: Paylaşımlarda yer alan yönlendirme linkleri temizlenmiştir.

-Etiketlerin silinmesi: Sosyal medya platformlarında belli konuları sınıflandırmak, bir araya toplamak, öne çıkarmak gibi farklı amaçlarla kullanılan etiketleme (Hashtag) işlemi vardır. Bu amaçla kullanılan # işareti veri setinden kaldırılmıştır.

-Bahsetmelerin silinmesi: Sosyal medya platformlarında yapılan paylaşımları, özellikle görmesi istenilen kişilere ulaştırmak amacıyla bu kişilerin kullanıcı isimlerinin paylaşımlara eklenmesi işlemine bahsetme (Mention) denir. Bu amaçla kullanılan @ işareti yardımıyla veri setinde yer alan bütün bahsetmeler kaldırılmıştır.

-Noktalama işaretlerinin kaldırılması: Python dilinde tanımlı noktalama işaretleri (!"#$%&'()*+,-./:;<=>?@[\]^_`{|}~) model doğruluğunu etkilememesi açısından paylaşımlardan kaldırılmıştır.

-Etkisiz kelimelerin silinmesi: İngilizce dilinde etkisiz olan ve bütün paylaşımlarda yer aldığı için model eğitimini olumsuz yönde etkileyecek bazı kelimeler (stopwords) vardır. Bu kelimeler Python dilinde farklı diller için önceden belirlenerek bir kütüphane içerisine eklenmiştir. Paylaşımlarda yer alan bu etkisiz kelimeler temizlenmiştir.

-Dil bilgisi düzeltme: Yine Python dilinde tanımlı language_tool_python kütüphanesi ile İngilizce dilinde yazılan herhangi bir metnin dil bilgisi (grammer) uygunluğunu kontrol etmek mümkündür. Tüm paylaşımlar bu düzeltme işleminden geçirilmiştir.

-Büyük – küçük harf düzeltmesi: Paylaşımları belli bir forma sokmak amacıyla bütün alfabetik karakterler küçük harfe çevrilmiştir.

-Emojilerin silinmesi: Gönderilerden emoji, ifade (emoticons), bayrak vb. kaldırılmıştır.

Ön işleme adımlarının tamamlanmasının ardından bazı tweetler tamamen yok olmuştur ve tweet sayısı 1787'den 1771'e düşmüştür. Veri setinde en sık kullanılan kelimeler tespit edilerek Şekil 2'de verilen kelime bulutu ile gösterilmiştir.


Şekil 2 Ön işleme yapılmış veri setinin kelime bulutu

### *C. Sınıflandırma*

Veri temizleme işlemlerinin tamamlanmasının ardından sınıflandırma işlemine geçilmiştir. Sınıflandırma işleminin yapılabilmesi için verilerin sayısal değerlere dönüştürülmesi gerekmektedir. Bu amaçla, veriler üzerinde önce Count Vectorizer işlemi uygulanmıştır. Count Vectorizer işlemi bir kelimenin metin içerisinde kaç kere geçtiğini sayar ve bu değeri ağırlık olarak kullanır. Bu işlemin ardından metin içerisinde geçen kelimelerin bulundukları metni ne kadar temsil ettiklerini ölçmek amacıyla Terim Sıklığı – Ters Doküman Sıklığı (Term Frequency – Inverse Document

Frequency – TF-IDF) işlemi yapılmıştır. Buradaki parametreler ise şunlardır;

TF (Term Frequency – Terim Sıklığı): İlgili kelimenin dokümandaki sıklığıdır. Kelimenin dokümanda geçme sayısını, dokümandaki toplam kelime sayısına bölerek elde edilir (Eşitlik 1).

$$tf(t,d) = \frac{f_d(t)}{\max_{w \in d} f_d(w)} \quad (1)$$

DF (Document Frequency – Doküman Sıklığı): TF ile benzemektedir ama bu kez diğer dokümanlara odaklanır. Doküman sayısının ilgili kelimenin geçtiği doküman sayısına bölünmesi ile hesaplanır (Eşitlik 2).

$$df(t,D) \frac{|D|}{|\{d \in D : t \in d\}|} \quad (2)$$

IDF (Inverse Document Frequency – Ters Doküman Sıklığı): DF değerinin logaritması alınarak hesaplanır (Eşitlik 3).

$$idf(t,D) = \ln\left(\frac{|D|}{|\{d \in D : t \in d\}|}\right) \quad (3)$$

TF-IDF (Term Frequency-Inverse Document Frequency): Terim sıklığı ve ters dokuman sıklığının çarpılması ile elde edilir (Eşitlik 4).

$$tfidf(t,d,D) = tf(t,d) * idf(t,D) \quad (4)$$

*Transformatör Modeller*

*XLNET:* Otoregresif ve otokodlama yöntemlerinin avantajlarını birleştirmek için bir permütasyon dili modellemesini kullanan genelleştirilmiş bir otoregresif ön eğitim yöntemidir (Yang et al., 2019; Fodeh et al., 2021). Permütasyon dili modelleme, sağdan sola veya soldan sağa yerine bir cümledeki tüm olası kelime permütasyonları üzerinde çalışılmasını sağlamaktadır (Naveen et al., 2021). XLNET ve BERT modeli arasındaki temel fark, XLNET yaklaşımının Permütasyon dil modeli sayesinde, BERT yaklaşımındaki gibi belirteçlerin %15'ini maskelemeye gerek kalmadan modeli çift yönlü bağlamda eğitmesine izin vermesidir. Ayrıca, XLNET, LSTM gibi dizi modellerinin tekrarlama fonksiyonu ile BERT'in aynı anda terimlerle ilgilenme avantajlarını harmanlamaktadır. Bu sayede, BERT'ten farklı olarak XLNET'in cümle uzunluğu sınırlaması yoktur. Herhangi bir uzunluktaki cümleleri parçalar olarak ele alarak işler ve durumu parçalar arasında taşır. Model, bir diziyi başka bir diziye çevirmek için bir diziyi diğerine eşleyerek bir kodlayıcı-kod çözücü modeli aracılığıyla tüm cümleden seçici olarak bilgi çıkarır. Modelin kod çözücü kısmı, kodlayıcının gizli durumlarının her birini ağırlıklandırarak ilgili gizli durumları bulmak için kodlayıcının tüm gizli durumlarını kullanır ve bu ağırlıklar basit bir ileri beslemeli sinir ağı tarafından belirlenir [12].

*ELECTRA:* BERT gibi dil modelleme ön eğitim yöntemleri, bazı belirteçleri [MASK] ile değiştirerek girdiyi bozar ve ardından orijinal belirteçleri yeniden oluşturmak için bir model eğitir. NLP görevlerine aktarıldıklarında iyi sonuçlar vermelerine rağmen, genellikle etkili olmaları için büyük miktarda bilgi işlem gerektirirler. Ancak ELECTRA, ön eğitim için değiştirilmiş bir belirteç algılama görevi kullanır. Bir üreticiden örnek alarak bazı belirteçleri değiştirerek girdiyi bozar, ardından her bir belirtecin orijinal mi yoksa yedek mi olduğunu tahmin etmek için bir ayrımcı eğitilir. Ayrımcının önemli bir avantajı, modelin yalnızca küçük maskelenmiş alt kümesi yerine tüm girdi belirteçlerinden öğrenmesi ve bu da onu hesaplama açısından daha verimli hale getirmesidir. BERT, küçük bir maskelenmiş alt kümeden (genellikle %15) öğrenirken, ELECTRA hesaplama açısından daha verimli olan tüm girdi belirteçlerinden öğrenebilir [13].

*Makine öğrenmesi modelleri*

*SVM:* Vapnik tarafından 1999 yılında önerilen ve son yıllarda sınıflandırma-regresyon problemlerinde sıklıkla kullanılan bir makine öğrenmesi tekniğidir. SVM'nin temel amacı, tüm destek vektörleri arasında en büyük geometrik aralığa sahip olan ve yeni verileri sınıflandırmak için kullanılan ayırma hiperdüzlemini hesaplama yoluyla bulmaktadır [14]. Bu hiperdüzlemler, ML modellerinin karar sınırları olarakta adlandırılır. Destek vektörleri, hiperdüzlemin konumunu ve yönünü belirlemek için kullanılan SVM hiperdüzlemine en yakın veri noktalarıdır.

*Lineer SVM:* Lineer destek vektör sınıflandırıcısı, büyük veri setlerinde çalışabilme avantajına sahip çok sınıflı sınıflandırma görevi gören bir algoritmadır.

*LR:* Lojistik regresyon, esas olarak tahminler için kullanılan denetimli bir sınıflandırma algoritmasıdır. Cevap değişkenin ikili (binary) olarak gözlendiği durumlarda bağımlı ve bağımsız değişkenler arasındaki ilişkiyi belirlemede kullanılır. Her bir veri noktasının sınıfını tahmin etmek için kullanılan olasılık değerlerini elde ederken aktivasyon fonksiyonu olarak sigmoid fonksiyonunu kullanır [15].

*DT:* Karar ağaçları, düğüm, dal ve yapraktan oluşmaktadır. Öznitelikler düğümler ile temsil edilir. Eğitim verilerine ait öznitelik bilgilerine göre ağaç yapısı oluşturulur ve karar sorularına verilen cevaplara göre karar kuralları oluşturulur. Oluşturulan ağacın yeni bir veri seti için genelleme kabiliyetinin belirlenmesi için test verisi kullanılır. Yeni gelen bir test verisi, ağacın kökünden başlar ve karar kuralına göre bir yaprağa gidene kadar devam eder. Her bir yaprak düğümü sadece bir sınıfa ait gözlem değerleri içerene kadar karar kuralı işletilir. Bu şekilde veriler sınıflandırılır. Karar ağaçlarında en önemli adım ağaçtaki dallanmanın hangi kritere göre yapılacağı veya hangi öznitelik değerlerine göre ağaç yapısının oluşturulacağıdır [16].

*RF:* Rastgele orman, öğrenme aşamasında eğitim veri örneklerinden rastgele bir seçimle oluşturulan birçok karar ağacını kullanır. Rastgele orman aynı zamanda bir veri örneğine ilişkin tüm karar ağaçlarının tahminlerini birbiriyle karşılaştırdığı için bir topluluk yöntemidir.

*GB:* Gradyan Arttırma, Friedman tarafından 2001 yılında tanıtılan güçlü bir makine öğrenme tekniğidir. Sırayla çok sayıda zayıf öğreniciyi inşa etmek ve onları karmaşık bir modele dahil etmeyi amaçlamaktadır [17].

*XGBoost:* Temeli GB ve DT algoritmalarına dayanan bir makine öğrenme tekniğidir. XGBoost, genel performansı iyileştiren ve aşırı uyum ya da aşırı öğrenmeyi azaltan bir dizi düzenlileştirme içermektedir. Bu sayede, ağaçların karmaşıklığını kontrol ederek daha iyi bir performans elde etmeyi başarmaktadır [17].

*KNN:* Etiketsiz bir test verisi verildiğinde eğitim veri setindeki en yakın k noktayı bulur ve bu veriye en uygun olan etiketi atar. Modeldeki tek ayarlanabilir parametre olan k, sınıf üyeliği tahminine dahil edilecek en yakın komşuların sayısını göstermektedir.

## IV. Araştırma Bulguları

Deneysel çalışmalarda, Python dilinde kodları çalıştırmak için Google tarafından geliştirilen ve uç noktada yüksek GPU'lara erişim izni vererek model eğitimlerinin daha hızlı gerçekleşmesine olanak tanıyan Colaboratory ortamı kullanılmıştır. Sınıflandırma aşamasında tüm veri seti %70 eğitim (1239), %15 (266) doğrulama ve %15 (266) test verisi olacak şekilde ayrıştırılmıştır. Çalışmada kullanılan bütün modeller doğruluk, kesinlik, duyarlılık ve F1 skor performans metriklerine göre karşılaştırılmıştır. Bu değerler Eşitlik 5-8 ile hesaplanır.

$$Doğruluk = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

$$Kesinlik = \frac{TP}{TP+FP} \quad (6)$$

$$Duyarlılık = \frac{TP}{TP+FN} \quad (7)$$

$$F1\ Skor = \frac{2*Kesinlik*Duyarlılık}{Kesinlik+Duyarlılık} \quad (8)$$

Çalışmanın ilk aşamasında klasik makine öğrenmesi yöntemleri eğitilmiş ve performans değerleri karşılaştırılmıştır. Bu aşamada elde edilen sonuçlar Tablo 1'de verilmiştir.

Tablo 1. Klasik makine öğrenmesi yöntemlerinin performans metriklerinin karşılaştırılması

| Metrikler | Modeller (%) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | LSVM | SVM | RF | XGB | GB | LR | DT | KNN |
| Doğruluk | 90.6 | 89.1 | 89.1 | 89.1 | 88.7 | 87.2 | 85.0 | 83.4 |
| Kesinlik | 90.9 | 95.3 | 97.5 | 94.2 | 98.7 | 93.9 | 86.7 | 79.8 |
| Duyarlılık | 84.9 | 76.4 | 74.5 | 77.3 | 72.6 | 72.6 | 73.6 | 78.3 |
| F1-Skor | 87.8 | 84.8 | 84.5 | 85.0 | 83.7 | 81.9 | 79.6 | 79.0 |

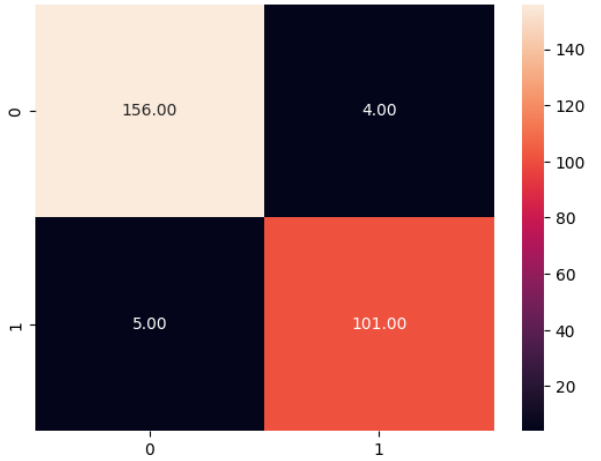
Makine öğrenmesi yöntemlerinde en yüksek doğruluk değeri LSVM ile %90.6 olarak elde edilmiştir. Çalışmanın ikinci aşamasında ise Transformatör yöntemlerinden ELECTRA ve XLNET'in performans değerleri incelenmiştir. Eğitilen modellerin hiperparametreleri şu şekildedir: Devir (epoch) 50, yığın boyutu (batch size) 32 ve öğrenme oranı (learning rate) 0.00005. Elde edilen sonuçlar Tablo 2'de verilmiştir.

Tablo 2. Transformatör yöntemlerinin performans metriklerinin karşılaştırılması

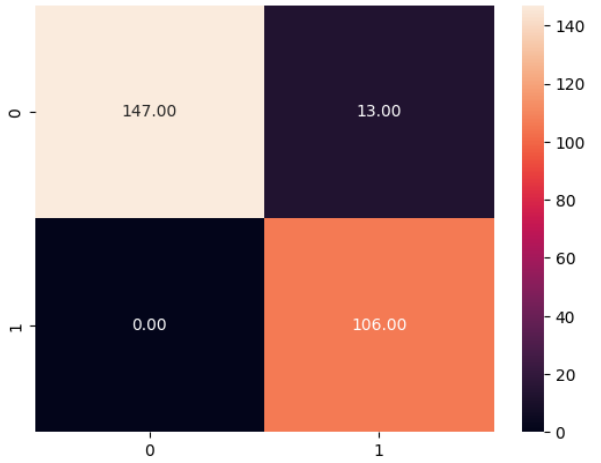
| Metrikler | Modeller (%) | |
|---|---|---|
| | ELECTRA | XLNET |
| Doğruluk | 96.61 | 95.11 |
| Kesinlik | 96.19 | 89.07 |
| Duyarlılık | 95.28 | 100 |
| F1-Skor | 95.73 | 94.22 |

Transformatör yöntemleri ile elde edilen sonuçlar incelendiğinde son yıllarda geliştirilen ELECTRA ve XLNET yöntemlerinin az sayıda veri üzerinde bile yüksek başarılara ulaştığı görülmektedir. Bu modellere ait karmaşıklık matrisleri Şekil 3'te verilmiştir.

Şekil 3 incelendiğinde ELECTRA yöntemi ile 160 intihara meyilli olmayan tweetten 156'sı doğru bulunurken, 106 intihara meyilli tweetten sadece 5'i hatalı olarak bulunmuştur. XLNET modelinde ise intihara meyilli tweetlerin tamamı doğru sınıflandırılırken intihara meyilli olmayan 13 tweet intihara meyilli olarak yanlış etiketlenmiştir. Çalışmada kullanılan tüm modellerin ROC AUC grafikleri ise Şekil 4'te verilmiştir.

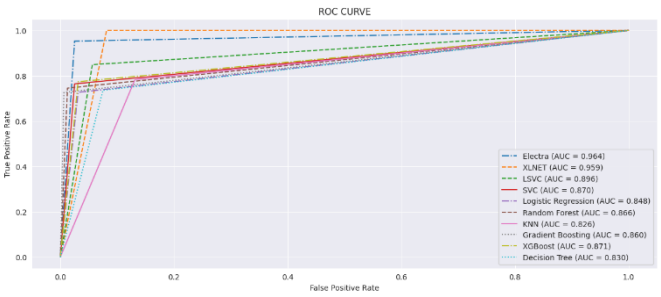



(a)




(b)

Şekil 3 Transformatör yöntemlerine ait karmaşıklık matrisleri (a) ELECTRA (b) XLNET




Şekil 4. Çalışmada kullanılan modellerin ROC-AUC grafiklerinin karşılaştırılması

## V. Sonuç ve Tartışma

Bu çalışmada intihara meyilli olarak atılan tweetlerin otomatik olarak tespiti amaçlanmıştır. Bu doğrultuda, çeşitli makine öğrenmesi yöntemleri ve transformatör modelleri kullanılarak analizler gerçekleştirilmiştir. Makine öğrenmesi yöntemlerinden LSVM, SVM, LR, RF, KNN, GB, XGBoost ve DT modelleri ele alınmıştır. Bununla birlikte transformatör yöntemleri olarak ELECTRA ve XLNET modelleri kullanılmıştır. Tüm bu modeller arasında en yüksek doğruluk değerine ELECTRA yöntemi ile ulaşılmıştır. Bu model ile elde edilen değerler şu şekildedir: doğruluk (%96.61), kesinlik (%96.19), duyarlılık (%95.28) ve F1 skoru (%95.73). Bu sonuçlar, intihara meyilli tweetlerin otomatik olarak tespitinde ELECTRA'nın etkili bir model olduğunu göstermektedir.

Literatürde yer alan çalışmalar incelendiğinde doğruluk değerini artırmak için veri artırma yöntemlerine odaklanıldığı görülmüştür. Buna rağmen elde edilen doğruluk değerleri düşük kalmaktadır. Bu çalışmada az sayıda veri ile literatürdeki sonuçlardan daha yüksek bir doğruluk değeri elde edilmiştir. Bu durum ELECTRA yönteminin doğal dil işleme çalışmalarında başarılı bir şekilde kullanılabileceğini göstermektedir.

## REFERANSLAR


[1] Rabani, S. T., Khan, Q. R., and Khanday, A. M. U. D. "Detection of suicidal ideation on Twitter using machine learning & ensemble approaches". *Baghdad science journal*, 17(4), 1328-1328, 2020.

[2] Haque, R., Islam, N., Islam, M., and Ahsan, M. M. "A comparative analysis on suicidal ideation detection using NLP, machine, and deep learning." *Technologies*, 10(3), 57,2022.

[3] Baghdadi, N. A., Malki, A., Balaha, H. M., AbdulAzeem, Y., Badawy, M., and Elhosseini, M. "An optimized deep learning approach for suicide detection through Arabic tweets." *PeerJ Computer Science*, 8, e1070, 2022.

[4] Abdulsalam, A., Alhothali, A., and Al-Ghamdi, S. "Detecting Suicidality in Arabic Tweets Using Machine Learning and Deep Learning Techniques." *arXiv preprint* arXiv:2309.00246, 2023.

[5] Abdulsalam, A., and Alhothali, A. "Suicidal ideation detection on social media: A review of machine learning methods." *arXiv preprint* arXiv:2201.10515, 2022.

[6] Ramírez-Cifuentes, D., Freire, A., Baeza-Yates, R., Puntí, J., Medina-Bravo, P., Velazquez, D. A., ... and Gonzàlez, J. "Detection of suicidal ideation on social media: multimodal, relational, and behavioral analysis." *Journal of medical internet research*, 22(7), e17758, 2020.

[7] Chatterjee, M., Kumar, P., Samanta, P., and Sarkar, D. "Suicide ideation detection from online social media: A multi-modal feature based technique." *International Journal of Information Management Data Insights*, 2(2), 100103, 2022.

[8] Sabri, N. M., and Mohamad, N. A. "Detection of Suicidal Tweets Based On Naïve Bayes Algorithm." *International Journal of Advanced Research in Technology and Innovation*, 4(3), 47-59, 2022.

[9] Deepa, J., Shriraaman, S., Shruti, V. V., and Vasanth, G. "Detecting and Determining Degree of Suicidal Ideation on Tweets Using LSTM and Machine Learning Models." *Journal of Survey in Fisheries Sciences*, 10(2S), 3217-3224, 2023.

[10] Priyamvada, B., Singhal, S., Nayyar, A., Jain, R., Goel, P., Rani, M., and Srivastava, M. "Stacked CNN-LSTM approach for prediction of suicidal ideation on social media." *Multimedia Tools and Applications*, 1-22, 2023.

[11] Mahmud, S. A., *Suicidal Tweet Detection Dataset*, (Son Erişim Tarihi: 30.09.2023), https://www.kaggle.com/datasets/aunanya875/suicidal-tweet-detection-dataset

[12] Naveen, S., Kiran, M. S. R., Indupriya, M., Manikanta, T. V., and Sudeep, P. V. "Transformer models for enhancing AttnGAN based text to image generation." *Image and Vision Computing*, 115, 104284, 2021.

[13] Clark, K., Luong, M. T., Le, Q. V., and Manning, C. D. "Electra: Pre-training text encoders as discriminators rather than generators." *arXiv preprint* arXiv:2003.10555, 2020.

[14] Li, X., Li, L., Ma, W., & Wang, W. "Two-phase flow patterns identification in porous media using feature extraction and SVM." *International Journal of Multiphase Flow*, 104222, 2022.

[15] Singh, N., Jena, S., and Panigrahi, C. K. "A novel application of Decision Tree classifier in solar irradiance prediction." *Materials Today: Proceedings*, 58, 316-323, 2022.

[16] Kavzoğlu, T., and Çölkesen, İ. "Karar ağaçları ile uydu görüntülerinin sınıflandırılması." *Harita Teknolojileri Elektronik Dergisi*, 2(1), 36-45, 2010.

[17] Yangın, G. "XGBoost ve Karar Ağacı Tabanlı Algoritmaların Diyabet Veri Setleri Üzerine Uygulaması" Mimar Sinan Güzel Sanatlar Üniversitesi, Fen Bilimleri Enstitüsü, İstatistik Anabilim Dalı, İstanbul, 2019.