Sosyal Medya Gönderilerinde İntihar İçeriğinin Derin Öğrenme Modelleriyle Analizi ile Modellerin Karşılaştırılması

Analysis of Suicide Content in Social Media Posts with Deep Learning Models and Comparison of Models

Mahmutcan SAKINCI [1], Enes ÇAKMAK[1], Nail KOCABAY [1], Mustafa TETİK [1], Serhat ÖZEKES [1], Merve PINAR [1], Abdulsamet AKTAŞ [1]

[1]Marmara Üniversitesi (MÜ), Teknoloji Fakültesi, Bilgisayar Mühendisliği Bölümü, İstanbul, Türkiye

Öz

Bu çalışma, sosyal medya gönderilerinde intihar içeriklerinin tespit edilmesi ve derin öğrenme modellerinin bu alandaki performanslarının karşılaştırılması üzerine odaklanmaktadır. Çalışmada, Reddit platformundan alınan iki farklı veri seti, kapsamlı bir ön işleme sürecinden geçirilerek birleştirilmiş ve literatürdeki mevcut çalışmalara kıyasla daha geniş çaplı bir veri seti oluşturulmuştur. Bu veri seti, sosyal medya gönderilerinin doğal dil işleme (NLP) teknikleriyle analiz edilmesine yönelik yenilikçi bir yaklaşım sunmaktadır. Çalışmada, T5-small ve ALBERT modelleri gibi ileri düzey derin öğrenme yöntemleri kullanılmış ve bu modellerin doğruluk oranları sırasıyla %97 ve %97,27 olarak tespit edilmiştir. Bulgular, özellikle geniş çaplı ve dengeli veri setlerinin kullanımının model performansını artırmadaki önemini vurgulamaktadır. Bunun yanı sıra, doğal dil işleme yöntemlerinin sosyal medya verilerinde intihar içeriklerinin tespit edilmesi ve erken müdahaleler için uygulanabilirliği değerlendirilmiştir. Sonuç olarak, bu çalışma hem yöntem hem de veri seti açısından literatüre önemli katkılar sunmakta ve intihar eğilimlerinin tespitinde derin öğrenme modellerinin gücünü ortaya koymaktadır.
Anahtar Kelimeler: Sosyal medya, intihar içerik analizi, derin öğrenme, doğal dil işleme, T5-small, ALBERT

Abstract

This study focuses on detecting suicide-related content in social media posts and comparing the performance of deep learning models in this domain. Two distinct datasets from Reddit were preprocessed and integrated to create a broader and more comprehensive dataset, distinguishing this work from previous studies. This dataset provides an innovative approach to analyzing social media posts using natural language processing (NLP) techniques. Advanced deep learning methods, such as T5-small and ALBERT, were employed, achieving accuracy rates of 97% and 97,27%, respectively. The findings emphasize the significance of utilizing large-scale and balanced datasets to enhance model performance. Additionally, the applicability of NLP methods in detecting suicidal content within social media data and their potential for enabling early interventions were evaluated. In conclusion, this study makes substantial contributions to the literature in terms of both methodology and dataset, demonstrating the power of deep learning models in identifying suicidal tendencies.

Keywords: Social media, suicide content analysis, deep learning, natural language processing, T5-small, ALBERT

I. GİRİŞ

Sosyal medya, modern yaşamın ayrılmaz bir parçası haline gelmiş ve bireylerin düşüncelerini, duygularını ve günlük yaşantılarını ifade ettikleri bir mecra olarak toplumsal etkileşimde büyük bir yer edinmiştir. Facebook, Twitter, Instagram ve Reddit gibi platformlar, bireylerin sosyal çevreleriyle bağlantıda kalmasını sağlamanın yanı sıra, toplumun genel ruh halini yansıtan geniş bir veri havuzu sunmaktadır. Bu platformlardan elde edilen büyük veri, sadece pazarlama ve trend analizi gibi ticari amaçlarla değil, aynı zamanda sağlık ve sosyal iyilik hali için kritik bilgiler sağlayabilecek analizler için de kullanılmaktadır.

İntihar, dünya genelinde ciddi bir halk sağlığı sorunu olarak varlığını sürdürmekte ve her yıl milyonlarca bireyi doğrudan ya da dolaylı olarak etkilemektedir. Dünya Sağlık Örgütü'ne göre, her yıl yaklaşık 726.000 kişi intihar sonucu yaşamını yitirmektedir [1]. Bu sayı, intihar eğilimlerinin erken tespiti ve önleyici müdahalelerin ne kadar hayati olduğunu göstermektedir. Erken müdahaleler, yalnızca bireysel hayatların kurtarılmasını değil, aynı zamanda toplumun genel refahını artırmayı da mümkün kılabilir.

Sosyal medya platformları, bireylerin ruhsal sağlık durumlarına dair önemli ipuçları barındırmakta olup, bu ipuçları intihar eğilimlerinin belirlenmesinde kullanılabilecek önemli veriler sunmaktadır. Özellikle metin tabanlı analizlerde, bireylerin duygusal durumlarına işaret eden dil kalıpları, kelime seçimleri ve içerik bağlamları, intihar riski taşıyan bireylerin tespiti için kritik öneme sahiptir. Bununla birlikte, sosyal medya verilerinin çeşitliliği, doğasındaki belirsizlikler ve veri setlerinin dengesiz yapısı, bu tür analizlerde çeşitli zorluklar yaratmaktadır.

Bu çalışmada, sosyal medya gönderilerinde intihar içeriklerini tespit etmek ve mevcut yöntemlerin doğruluğunu artırmak amacıyla yeni bir yaklaşım önerilmiştir. Reddit platformundan alınan iki farklı veri seti, detaylı bir ön işleme sürecinden geçirilmiş ve bu veri setleri birleştirilerek literatürdeki diğer çalışmalardan farklı olarak daha geniş kapsamlı ve dengeli bir veri kümesi oluşturulmuştur. Bu veri birleştirme işlemi, hem veri setlerinin sınıf dağılımındaki

dengesizlik sorunlarını azaltmış hem de analiz için daha çeşitli örnekler sunarak model performansını artırmıştır. Birleştirilen veri setleri, sosyal medya platformlarındaki intihar içeriklerinin daha doğru bir şekilde tespit edilmesine olanak sağlayacak şekilde optimize edilmiştir.

Çalışmada, T5-small ve ALBERT gibi ileri düzey doğal dil işleme (NLP) modelleri kullanılmış ve bu modellerin doğruluk oranları sırasıyla %97 ve %97.27 olarak tespit edilmiştir. Bu başarı, birleştirilen veri setlerinin analitik gücünü ve önerilen modellerin etkinliğini açıkça ortaya koymaktadır. Bu araştırmanın temel amacı, mevcut literatürdeki boşlukları doldurarak sosyal medya verilerinden elde edilen bilgilerle intihar riskine yönelik erken uyarı sistemlerinin geliştirilmesine katkı sağlamaktır. Bu doğrultuda, yalnızca teknik bir analizden ziyade, sosyal bir sorumluluk duygusuyla hareket edilerek bireylerin hayatlarını iyileştirmeyi hedefleyen bir yaklaşım benimsenmiştir.

Sonuç olarak, bu çalışma, sosyal medya verilerinin analizinde derin öğrenme modellerinin gücünü ve veri birleştirme süreçlerinin önemini ortaya koymaktadır. Ayrıca, intihar içeriklerinin tespitinde kullanılan teknolojilerin sadece akademik alanda değil, aynı zamanda toplum sağlığı ve sosyal farkındalık oluşturma bağlamında da önemli etkiler yaratabileceğini göstermektedir.

II. LİTERATÜR TARAMASI

2.1. Literatürde Bulunan Çalışmalar

Literatürde bulunan çalışmalarda, sosyal medya metin analizi ve intihar eğilimi tespiti üzerine yapılan son yıllardaki literatür incelenmiştir. İntihar, dünya genelinde önemli bir halk sağlığı sorunudur ve erken tespit hayati öneme sahiptir. Bu bağlamda, doğal dil işleme (NLP) ve makine öğrenimi yöntemleriyle sosyal medya gönderilerini analiz eden yaklaşımlar değerlendirilmiş ve veri setlerinin kullanımı incelenmiştir. Araştırmalar, mevcut modellerin doğruluk oranlarını, veri setlerinin özelliklerini ve tespit edilen boşlukları ele almıştır.

Ghelmar Astoveza ve ekibinin yaptığı çalışmada Twitter'da intihar düşüncelerinin tespiti için makine öğrenimi ve sinir ağlarının kullanımını incelemiştir. Araştırmacılar, yapay sinir ağları (ANN) ve Çok Katmanlı Algılayıcı (MLP) sınıflandırıcılarını kullanarak potansiyel olarak riskli tweetleri belirleyen modeller geliştirmiş ve farklı kategorilerde %65-91 oranında doğruluk elde etmişlerdir [2]. İntihar düşüncesi içeren ve içermeyen tweetleri ayırt etmek için çeşitli makine öğrenimi ve ensemble yöntemleri, Twitter'dan manuel olarak toplanan veriler üzerinde kullanılmıştır. Bu yaklaşımlar, kendine zarar verme ve intihar düşüncelerini ifade eden bireylerin kullandığı dil ve kalıpları analiz etmek için doğal dil işleme tekniklerinden yararlanmaktadır. Bu çalışmalar, intihar düşüncelerinin otomatik tespiti için potansiyeli gösterse de araştırmacılar gerçek intihar riski tahmininde bu modellerin kesinliğinin henüz kanıtlanmadığını ve yalnızca bu modellere dayanarak doğrudan müdahalede bulunmanın önerilmediğini vurgulamaktadır. Bu çalışma intihara meyilli tweetleri riskli tweetlerde %65, riskli olmayan tweetlerde ise %91 doğrulukla sınıflandırmak için bir sinir ağı modeli geliştirilmiştir [2].

Rezaul Haque ve ekibi bu çalışmada, Twitter sosyal medya platformunda intihar düşüncelerini belirlemek amacıyla çeşitli makine öğrenimi ve derin öğrenme modellerinin karşılaştırmalı bir analizi yapılmıştır. Araştırmanın temel amacı, önceki çalışmalara kıyasla daha yüksek bir model performansı elde ederek intihar belirtilerini doğru bir şekilde tanımlamak ve intihar girişimlerinin önüne geçmektir. Bu doğrultuda, metin ön işleme ve özellik çıkarma yöntemleri (örneğin, CountVectorizer ve kelime gömme teknikleri) uygulanmış ve çeşitli makine öğrenimi ile derin öğrenme modelleri eğitilmiştir.

Araştırmada kullanılan veri seti, Python Tweepy API'ı aracılığıyla canlı tweetlerden toplanan ve 18 intihar ve intihar dışı anahtar kelimeyi içeren 49.178 örnekten oluşmaktadır. Elde edilen deneysel bulgular, makine öğrenimi algoritmaları arasında RF (Random Forest) modelinin %93 doğruluk ve 0.92 F1 skoru ile en iyi sınıflandırma performansını sağladığını ortaya koymaktadır. Bunun yanı sıra, kelime gömme teknikleri ile eğitilen derin öğrenme modellerinin, makine öğrenimi modellerine göre daha yüksek performans sergilediği ve BiLSTM modelinin %93,6 doğruluk ve 0.93 F1 skoru ile en iyi sonuçları verdiği gözlemlenmiştir [3].

Norlina Mohd Sabri ve Noor Alisa Mohamad'ın yaptığı çalışmada pandeminin insan psikolojisi üstündeki etkilerini gözlemledi. Pandemi, insanların düşüncelerini, görüşlerini ifade etmek, inceleme paylaşmak ve günlük yaşamlarını çevrimiçi paylaşmak için sosyal medyaya bağımlı hale geldiği bir trend oluşturdu. Bu durum, sosyal medya platformlarındaki veri hacminin artmasına katkıda bulundu. Bu bağlamda, sosyal medya verileri işletmeler ve organizasyonlar için analiz edilmiştir. Ürünler ve günlük durumlar gibi konularda kamuoyu görüşlerini elde etmek için çeşitli duygu analizi araştırmaları yapılmıştır. Twitter, mesaj yazmanın kolaylığı ve kullanım rahatlığı nedeniyle en çok ziyaret edilen sosyal medya platformlarından biri haline gelmiştir. Bu araştırma, Twitter verilerinin intihar niyeti tespiti için analiz edilmesini önermektedir. Bu, Covid-19 pandemisinden etkilenen depresif kişilere, çevrimiçi görüş ve düşüncelerini analiz ederek yardımcı olmayı amaçlamaktadır. Tweetlerdeki intihar davranışını tespit etmek, intihar önleme için ilk adım olabilir. Araştırmanın amacı, Naïve Bayes algoritmasının intihar düşüncesi içeren tweetleri tespit etme yeteneğini keşfetmektir. Twitter verileri, Mayıs 2021'de Malezya'nın pandemi karantinası sırasında Tweepy kütüphanesi kullanılarak stres, kaygı, depresyon ve intihar anahtar kelimeleriyle toplanmıştır. Toplamda 5439 tweet toplanmıştır. Değerlendirme sonuçları, algoritmanın %80,39 doğrulukla intihar düşüncesi içeren tweetleri tespit etmede iyi ve kabul edilebilir bir performans sergilediğini göstermiştir. Sonuçlar ayrıca, pandemide, özellikle karantina döneminde, daha fazla insanın depresif olduğunu ortaya koymuştur. Gelecek çalışmalarda, Naïve

Bayes'in performansının diğer tanınmış sınıflandırıcılarla karşılaştırılması ve Facebook ve Instagram gibi diğer sosyal medya platformlarından farklı dillerdeki kelimelerin toplanıp işlenmesi planlanmaktadır [4].

Ning Wang ve ekibi [5], sosyal medya gönderilerinden intihar girişimlerini tahmin etmek amacıyla bir derin öğrenme modeli olan C-Attention (C-Att) ağını geliştirmiştir. Çalışmada, CLPsych 2021 yarışması tarafından sağlanan veri seti kullanılmıştır [6]. Bu veri seti, Twitter gönderilerinden elde edilen, intihar girişiminde bulunan veya bulunma riski taşıyan bireylerin paylaşımlarını içermektedir. Araştırmada kullanılan C-Att modeli, Doc2Vec ile oluşturulan 100 boyutlu metin gömülerini kullanarak multi head self attention (MHA) modülü ile ilişkileri yakalamış ve çıktıları sınıflandırmıştır. Model, 6 ay öncesine ait intihar tahmininde F1 skoru 0.737 ve F2 skoru 0.843 ile literatürdeki diğer yöntemlerden üstün bir performans sergilemiştir. Bunun yanında, Support Vector Machine (SVM) ve K-Nearest Neighbor (KNN) gibi geleneksel makine öğrenimi yöntemleri de Doc2Vec gömüleri ile test edilmiştir. 30 gün öncesine ait intihar girişimlerini tahmin etmede bu yöntemler, F1 skoru 0.741 ve F2 skoru 0.833 ile en iyi sonuçları elde etmiştir. Araştırmacılar, modelin özellikle uzun vadeli intihar tahminlerinde etkili olduğunu, ancak veri seti dengesi ve küçük veri alt gruplarında performans zorlukları yaşandığını vurgulamıştır.

Yaakov Ophir ve ekibi [7], Facebook gönderilerinden intihar riskini tahmin etmek amacıyla iki derin sinir ağı modeli geliştirmiştir: Single Task Model (STM) ve Multi-Task Model (MTM). Çalışmada, 1002 Facebook kullanıcısına ait 83.292 gönderi ve bu kişilerin klinik olarak doğrulanmış psikososyal bilgilerinden oluşan bir veri seti kullanılmıştır. STM, gönderilerden doğrudan intihar riskini tahmin ederken (Facebook metinleri → intihar), MTM, kişilik özellikleri, psikososyal riskler ve psikiyatrik bozukluklar gibi katmanlı risk faktörlerini dikkate alarak daha yüksek doğruluk sağlamıştır (Facebook metinleri → kişilik özellikleri → psikososyal riskler → psikiyatrik bozukluklar → intihar).

Araştırmada, Embeddings from language model (ELMo) algoritmasıyla metin etiketleri çıkarılmış ve modeller bu etiketlerle eğitilmiştir. MTM, genel intihar riski için AUC skorunu 0.746, yüksek intihar riski için ise 0.697 elde ederek STM'ye göre önemli ölçüde daha yüksek performans göstermiştir. Ayrıca, MTM'nin tahminlerinin açıkça intiharla ilgili ifadelerden çok, duygusal olarak yüklü dil ve olumsuz temalar gibi ince metinsel işaretlere dayandığı tespit edilmiştir.

Bu çalışma, sosyal medya gönderilerinden intihar riski tahmini için çok katmanlı bir modelin, basit yöntemlere göre üstün olduğunu göstermektedir. Araştırmacılar, bu tür modellerin psikolojik ve hesaplamalı bilimlerden türetilen teorik bilgileri birleştirerek daha yüksek doğruluk sağladığını vurgulamıştır.

Dheeraj Kodati ve Ramakrishnudu Tene [8], sosyal medya gönderilerinde intihara ilişkin duyguları tespit etmek amacıyla iki derin öğrenme modeli geliştirmiştir: Context-based bidirectional gated recurrent unit with multi-head attention and a convolutional neural network (C-BiGRU-MHA-CNN) ve Lexicon-based bidirectional long short-term memory with multi-head attention and convolutional neural network (L-BiLSTM-MHA-CNN). Çalışmada, Reddit'teki *SuicideWatch* topluluğundan elde edilen 6820 gönderi ve CEASE [9] veri seti kullanılmıştır. C-BiGRU-MHA-CNN modeli, bağlamsal bilgiyi korumak ve uzun vadeli bağımlılıkları ele almak amacıyla BERT gömülerini kullanmıştır. Bu model, dikkat mekanizması ve evrişimli sinir ağı katmanları aracılığıyla duygu sınıflandırmasını gerçekleştirmiştir. Model, Reddit veri setinde %98,12 doğruluk ve CEASE veri setinde %97,68 doğruluk oranına ulaşmıştır. L-BiLSTM-MHA-CNN modeli ise, sözlük tabanlı özelliklerle birlikte bağlamsal bilgiyi kullanarak doğruluk oranını artırmıştır.

Araştırmacılar, bu modellerin duygu kategorilerini başarıyla tespit ettiğini ve bağlamsal özelliklerin yanı sıra sözcük türü etiketlerinin de (POS) model performansına önemli katkılar sağladığını belirtmiştir. Ayrıca, intihar içeriği taşıyan metinlerde en belirgin duyguların korku, üzüntü ve depresyon olduğu tespit edilmiştir.

Emily Lin, Jian Sun, Hsingyu Chen ve Mohammad H. Mahoor'un gerçekleştirdiği bu çalışma [10], sosyal medya gönderilerinden intihar niyetlerini tespit etmeyi amaçlamıştır. Araştırmada, Reddit platformundaki SuicideWatch ve Depression alt başlıklarından toplanan gönderilerle oluşturulan Suicide and Depression Detection (SDD) veri seti kullanılmıştır [11]. 110,040 gönderiden oluşan bu veri seti, intihar ve intihar olmayan olmak üzere iki sınıfa ayrılmıştır. Araştırmacılar, RoBERTa modelini bir Convolutional Neural Network (CNN) katmanıyla birleştirerek RoBERTa-CNN adlı yeni bir model geliştirmiştir. Bu model, dilbilimsel örüntüleri daha iyi kavrayarak intihar niyetlerini etkili bir şekilde tespit etmeyi hedeflemektedir. Model, doğruluk, kesinlik (precision), hatırlama (recall), F1 skoru ve AUC gibi metriklerle değerlendirilmiştir ve %98 doğruluk, %96,81 F1 skoru gibi yüksek sonuçlar elde etmiştir. Ayrıca, model, diğer yöntemlere kıyasla daha üstün bir performans sergileyerek, intihar niyetlerinin tespitinde güvenilir bir araç olarak öne çıkmıştır.

Moumita Chatterjee, Poulomi Samanta, Dhrubasish Sarkar ve Piyush Kumar tarafından gerçekleştirilen bu çalışma [12], Twitter verileri üzerinde çoklu özellik analizi yaparak intihar niyetlerinin tespit edilmesini hedeflemiştir. Çalışmada, Twitter API'ları kullanılarak 188,704 İngilizce tweet toplanmış ve bu veri seti intihar niyetli ve niyetsiz gönderiler olarak iki sınıfa ayrılmıştır. Veri setinin temizlenmesi ve hazırlanmasından sonra, dilbilimsel, duygusal, istatistiksel, konu temelli ve zamansal özellikler çıkarılmıştır. Logistic Regression, Random Forest, Support Vector Machine ve XGBoost gibi algoritmalar kullanılarak sınıflandırma yapılmıştır. Çalışmada en iyi performans, Logistic Regression ile %87 doğruluk ve 0.81 F1 skoru elde edilerek sağlanmıştır. Sonuçlar, öz niteliklerin uygun kombinasyonlarının sınıflandırma başarısını artırmada önemli bir rol oynadığını göstermiştir.

Michael Mesfin Tadesse ve ekibi (2020) tarafından gerçekleştirilen çalışmada [13], Reddit sosyal medya

platformundan toplanan veri seti kullanılmıştır [14]. Veri seti, toplamda 7201 yazıdan oluşmakta olup, bunların 3549'u intihar göstergesi olan ve 3652'si intihar eğilimi taşımayan gönderilerden meydana gelmektedir. Yazılar, kullanıcıların kimlik bilgilerini korumak amacıyla anonimleştirilmiştir. Araştırmada, intihar eğilimi taşıyan yazıların belirlenmesi için birleştirilmiş bir Uzun Kısa Süreli Bellek (LSTM) ve Konvolüsyonel Sinir Ağı (CNN) modeli önerilmiştir. Bu model, LSTM'nin uzun mesafeli bağımlılıkları yakalama yeteneği ile CNN'in yerel kalıpları belirleme gücünü birleştirerek daha yüksek doğruluk oranları elde etmeyi hedeflemiştir. Çalışmada, Word2Vec tabanlı kelime gömme teknikleri kullanılarak metinler temsil edilmiştir. Modelin hiper parametreleri 10 katlı çapraz doğrulama ile optimize edilmiş ve performansı farklı makine öğrenmesi ve derin öğrenme yöntemleriyle karşılaştırılmıştır. Değerlendirme metrikleri olarak doğruluk, F1 skoru, kesinlik ve duyarlılık kullanılmıştır. Önerilen LSTM-CNN modeli, %93,8 doğruluk ve %93,4 F1 skoru ile diğer yöntemleri geride bırakmıştır. Çalışmanın sonuçları, önerilen modelin intihar riski taşıyan yazıları başarılı bir şekilde tespit ettiğini ve bu tür eğilimlerin dil kullanımındaki değişimlerle bağlantılı olduğunu göstermiştir.

Tolga Aydın ve Muhammed Coşkun Irmak [15], günümüzde en çok kullanılan sosyal medya platformlarından birisi olan X (Twitter)'te paylaşılan içerikleri analiz ederek intihar düşüncelerini tespit etmeye odaklanan bir araştırma yapmışlardır. ELECTRA ve XLNET modellerini kullanarak bunları klasik makine öğrenmesi yöntemleri ile kıyaslamışlardır. Çalışmanın sonunda en yüksek doğruluğa (%96,61) ELECTRA modelini kullanarak ulaşmışlardır. Xlnet %95,11, klasik makine öğrenmesi yöntemlerinin en başarılısı (LSVM) ise %90,6 doğruluğa ulaşabilmiştir.

Özay Ezerceli ve Rahim Dehkharghani [16], çalışmalarında Reddit gönderilerinden ve intihar notlarından oluşan üç farklı veri setine (SuicideDetection, CEASEv2.0 ve SWMH) üç ayrı derin öğrenme yaklaşımı uygulamış ve elde ettikleri sonuçları klasik makine öğrenimi yöntemleriyle karşılaştırmışlardır. Araştırmaları sonucunda, veri setlerinde sırasıyla 0.97, 0.75 ve 0.68 F1 skorlarına ulaşmışlardır. Özellikle SuicideDetection ve CEASEv2.0 veri setlerinde, önceki state-of-the-art modellerden daha iyi performans göstermeyi başarmışlardır.

Theyazn H. H. Aldhyani, Saleh Nagi Alsubari, Ali Saleh Alshebami, Hasan Alkahtani Zeyad A. T. Ahmed [17] intihara meyilli sosyal medya gönderilerini tespit etmeyi hedefleyen bir başka çalışmada metin temsili için TF-IDF ve Word2Vec yöntemlerinden yararlanıp, CNN–BiLSTM ve XGBoost modellerini kullanarak Reddit gönderilerinden oluşan bir veri tabanı üzerine çalışmışlardır. Sırasıyla %95 ve %91,5 doğruluk oranına ulaşarak bu konuda CNN–BiLSTM modelinin XGBoost modelinden daha iyi sonuç verdiğini göstermişlerdir.

Tablo 1. Literatürde bulunan çalışmaların sonuçları

Veri setleri	Araştırmacılar	Model/Metot	Performans (Doğruluk)
CLPsych 2021	Ning Wang ve ekibi [5]	C-Att, SVM, KNN	%73,7 F1, %84,3 F2
Facebook	Yaakov Ophir ve ekibi [7]	STM, MTM	%74,6 AUC
Sosyal Medya Platformlarının Verileri	Dheeraj Kodati ve Ramakrishnudu Tene [8]	C-BiGRU-MHA-CNN, L-BiLSTM-MHA-CNN	%98,12, %97,68
Reddit SuicideWatch	Emily Lin ve ekibi [10]	RoBERTa-CNN	%98 doğruluk, %96,81 F1
Tweepy API	Moumita Chatterjee ve ekibi [12]	Logistic Regression, Random Forest, SVM, XGBoost	%87 doğruluk, %81 F1
Reddit SuicideWatch	Michael Mesfin Tadesse ve ekibi [13]	LSTM-CNN	%93,8 doğruluk, %93,4 F1
Reddit SuicideWatch	Tolga Aydın ve Muhammed C. Irmak [15]	ELECTRA, XLNET, LSVM	%96,61 doğruluk (ELECTRA), %95,73 F1
SuicideDetection CEASEv2.0 SWMH	Özay Ezerceli ve Rahim Dehkharghani [16]	Derin Öğrenme Yöntemleri, Klasik ML	%97,42 doğruluk, %97 F1
Reddit SuicideWatch	Theyazn H. H. Aldhyani ve ekibi[17]	CNN-BiLSTM, XGBoost	%95 doğruluk, %91,5 F1
TwitterAdvanced Search (TAS)	Ghelmar Astoveza ve ekibi [2]	ANN, MLP	%65-%91 doğruluk, %88 F1
Tweepy API	Rezaul Haque ve ekibi [3]	Random Forest, BiLSTM	%93,6 doğruluk, %93 F1
Tweepy API	Norlina Mohd Sabri ve ekibi [4]	Naïve Bayes	%80,39 doğruluk, %77 F1

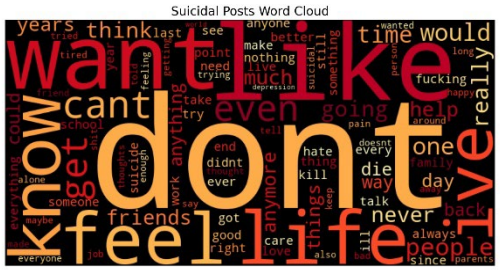
III. MATERYAL VE YÖNTEM

Burada, çalışmamızda kullanılan veri seti, ön işleme ve yöntem hakkında bilgiler verilmiştir.

3.1. Veri Seti

Çalışmamızda, biri Mendeley Data platformundan temin edilen Suicidal Ideation Detection Dataset [18], diğeri ise Kaggle platformunda yer alan Suicide and Depression Detection Dataset olmak üzere iki farklı veri seti kullanılmıştır. Suicidal Ideation Detection Dataset, Reddit platformundaki SuicideWatch topluluğundan intihara meyilli gönderiler ve CasualConversation, BenignExistence ve CongratsLikeImFive topluluklarından intihara meyilli olmayan gönderileri içermektedir. Bu veri seti, 15.477 kayda sahiptir; veriler 1 Haziran 2023 ile 13 Kasım 2023 tarihleri arasında toplanmıştır.

Şekil 1. İntihar meyilli olmayan gönderilen kelime popülasyonu

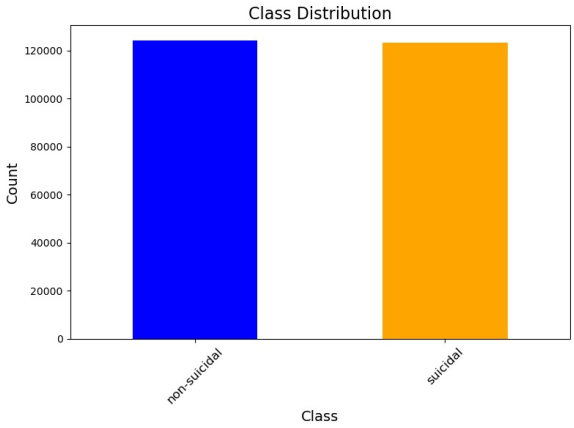


Şekil 2. İntihar meyilli gönderilerin kelime popülasyonu

Öte yandan, Suicide and Depression Detection Dataset [11] ise Reddit platformundaki SuicideWatch topluluğundan intihara meyilli gönderiler ve Depression topluluğundan depresyon içerikli gönderilerin yanı sıra, r/teenagers topluluğundan normal içerikli gönderileri içermektedir. Bu veri seti, 247.551 kayıt ve iki etiketle (intihara meyilli ve intihara meyilli değil) oluşturulmuştur.

Şekil 3. Tüm gönderilerin kelime popülasyonu

Toplamda 247551 kayıttan oluşan bu veri setleri intihara meyilli, intihara meyilli değil şeklinde iki etikete sahipler.

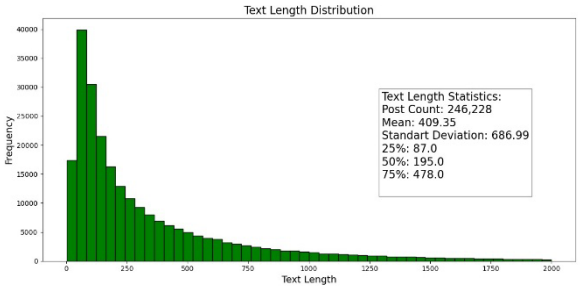


Şekil 4. Veri setinin etiket dağılımını gösteren grafik

3.2. Ön İşleme

Eğitimin daha geniş bir kapsama alanına sahip olması için birleştirme işlemi yapılmıştır. Bu işlem modelin genelleme kapasitesini yükseltmektedir. Birleştirilen iki farklı veri setindeki iki farklı etiket isimlendirmeleri aynı olacak şekilde değiştirilmiştir. Veri setlerindeki nitelikler Post ve Label adında iki farklı değer olarak ayarlanmıştır. Post kısmı kullanıcıların sosyal medyadaki metinlerini belirtmek amacıyla kullanılmıştır. Label kısmı ise kullanıcının metninin suicidal veya non-suicidal olduğunu belirtmek amacıyla kullanılmıştır. Post verilerindeki noktalama işaretleri, anlamsız kelimeler, emojiler kaldırılmıştır.

Post kısmındaki bütün büyük harfler küçük harfe çevrilmiştir. Verilerdeki noktalama işaretleri, anlamsız kelimeler ve emojilerin kaldırılmasında NLTK kütüphanesi kullanılmıştır.

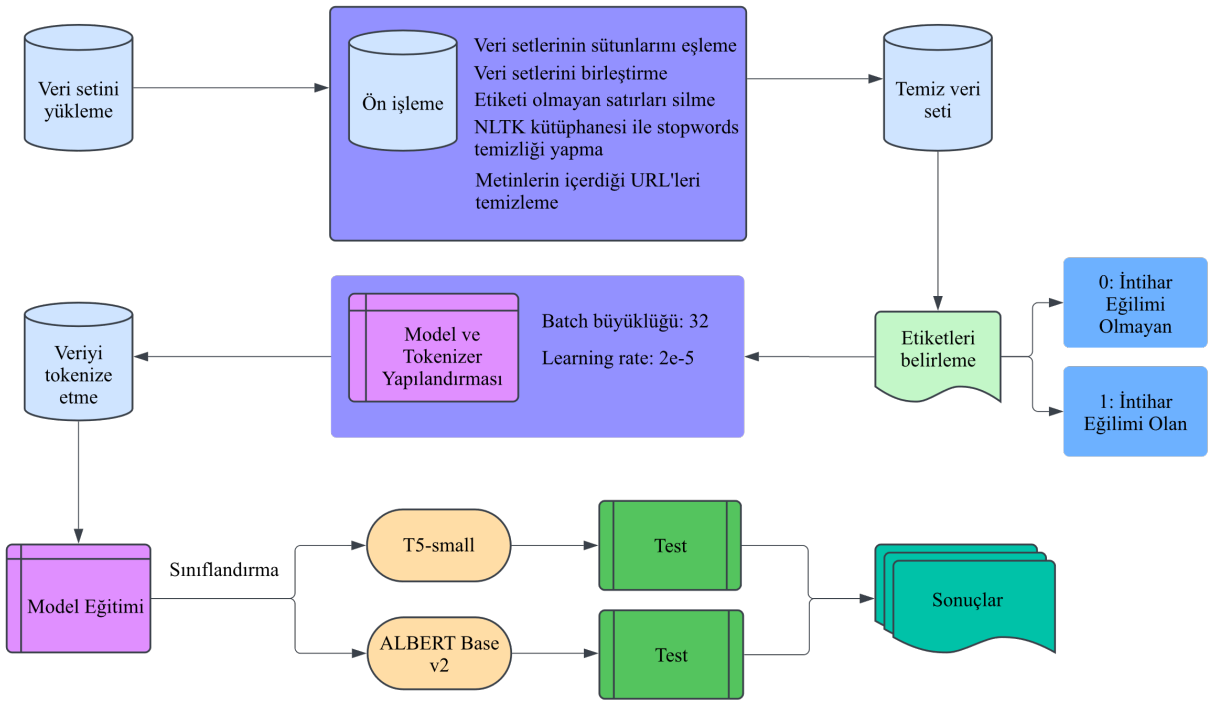


Şekil 5. Veri setindeki gönderilerin uzunluğunu gösteren grafik

3.3. Önerilen Yöntem

T5[19] (Text-to-Text Transfer Transformer), Google tarafından geliştirilmiş bir dil modeli olup tüm doğal dil işleme (NLP) görevlerini metinden-metine formatında ele alan yenilikçi bir yaklaşıma sahiptir. Bu model, görevlerin giriş ve çıkışlarını metin formatında temsil ederek (örneğin, sınıflandırma görevinde bir cümlenin sınıf etiketini metin olarak üretmek) farklı görev türlerini birleştiren bir framework sunar. Transformer tabanlı encoder-decoder mimarisi kullanan T5, "doldurma" (span corruption) adı verilen bir bozma tekniğiyle eğitilir; bu teknikte, metindeki rastgele bölümler kaldırılır ve modelden bu boşlukları doldurması beklenir. Modelin farklı boyutları (ör., T5 Small, Base, Large) arasında, çalışmada düşük hesaplama maliyeti nedeniyle T5 Small tercih edilmiştir. Bu model, dil anlama ve dil üretimi görevlerinde oldukça başarılıdır ve tüm NLP görevlerini birleştirerek daha verimli bir öğrenme süreci sağlar.

Öte yandan ALBERT[20] (A Lite BERT), Google tarafından BERT'ün daha hafif ve verimli bir versiyonu olarak geliştirilmiştir. Model, parametre paylaşımı ve faktörize embedding parametrizasyonu gibi yeniliklerle hesaplama maliyetini azaltırken performansını korur. Parametre paylaşımı, modelin tüm katmanlarında aynı ağırlıkları kullanarak toplam parametre sayısını düşürürken, faktörize embedding parametrizasyonu, embedding boyutunu küçülterek bellekte daha az yer kaplamasını sağlar. ALBERT, maskelenmiş dil modelleme (MLM) ve sentence order prediction (SOP) hedefleriyle eğitilir; SOP, cümle ilişkilerini anlamada BERT'in kullandığı "Next Sentence Prediction" (NSP) görevinden daha etkilidir. Bu optimizasyonlar, ALBERT'i hızlı, hafif ve büyük veri kümelerinde etkili bir model haline getirir.



Şekil 6. Model eğitiminde kullanılan akış şeması

Bu çalışmada, T5-Small ve ALBERT modelleri, intihara meyilli metinlerin tespiti için ayrı ayrı ince ayar sürecine tabi tutulmuş, her iki modelin sonuçları karşılaştırılarak en uygun model belirlenmeye çalışılmıştır.

3.4. Uygulanan Yöntem

Birleştirilen iki farklı veri seti üzerinde farklı sütun isimlerinin tutarlılığını sağlamak için Post ve Label sütunları altında birleştirme işlemi gerçekleştirildi. Bu işlem, veri setlerinin uyumlu bir şekilde birleştirilmesine olanak tanıyarak model eğitiminde kullanılacak düzgün bir veri seti oluşturdu. Daha sonra Post verileri üzerinde detaylı bir ön işleme süreci uygulandı. Bu süreçte metinlerden noktalama işaretleri, cümleye anlam katmayan kelimeler ve emojiler gibi potansiyel olarak dikkat dağıtıcı unsurlar kaldırıldı. Bu temizleme işlemleri, verinin daha anlamlı ve model eğitimi için uygun hale getirilmesini sağladı.

Birleştirilen iki farklı veri seti üzerinde, farklı sütun isimlerinin tutarlılığını sağlamak için Post ve Label sütunları altında birleştirme işlemi gerçekleştirilmiştir. Bu işlem, veri setlerinin uyumlu bir şekilde birleştirilmesine olanak tanıyarak, model eğitiminde kullanılacak düzgün bir veri seti oluşturmuştur. Veri birleştirme işlemi sonrasında, veri setindeki sınıf dağılımı Şekil 4'teki gibi gözlemlenmiştir. Gözlemde elde edilen sonuçlar, non-suicidal ve suicidal sınıflarının yaklaşık olarak eşit sayıda gözleme sahip olduğunu göstermektedir; her iki sınıfta yaklaşık 120.000 örnek bulunmaktadır. Bu durum, veri setinin dengeli bir dağılıma sahip olduğunu ve sınıflar arasında dengesizlikten kaynaklanabilecek olası model yanlılıklarını en aza indirdiğini göstermektedir.

Birleştirme işlemini takiben Post sütununa ait veriler üzerinde detaylı bir ön işleme süreci uygulanmıştır. Bu süreçte, metinlerden noktalama işaretleri, bağlam açısından cümleye anlam katmayan kelimeler ve emojiler gibi potansiyel olarak dikkat dağıtıcı unsurlar temizlenmiştir. Ayrıca, veri setindeki metin uzunluklarının dağılımı Şekil 5'teki gibi analiz edilmiştir. Metin uzunluklarına ilişkin yapılan analizler, veri setindeki toplam 246.228 gönderinin ortalama uzunluğunun 409.35 karakter olduğunu göstermiştir. Standart sapma ise 686,99 olarak hesaplanmıştır. Bu istatistiklere göre metin uzunluklarının çoğunluğu 87 ile 478 karakter arasında yoğunlaşmıştır. Bu bilgi, modelin çok uzun veya çok kısa metinlerden etkilenme durumunu değerlendirmek için önemli bir içgörü sağlamaktadır.

Yapılan bu temizleme işlemleri ve analizler, veri setinin daha anlamlı hale getirilmesini ve model eğitimi için uygun bir yapı oluşturulmasını sağlamıştır. Şekil 4'te gözlemlenen dengeli sınıf dağılımı ve metin uzunluklarının özenle analiz edilmesi hem modelin performansını artırmakta hem de sonuçların güvenilirliğini desteklemektedir.

Temizlenmiş veri seti, doğal dil işleme görevlerinde etkili bir performans gösteren T5-small modeli kullanılarak eğitildi. Eğitim sürecinde veri seti %70 eğitim ve %30 test oranında bölünerek kullanıldı. Bu süreçte modelin metinleri anlamlandırma ve sınıflandırma yeteneği optimize edildi. Eğitim tamamlandıktan sonra, model test aşamasına tabi tutuldu. Test sonuçlarında T5-small modeli 0.97 doğruluk oranı (accuracy) elde ederek oldukça yüksek bir performans sergiledi. Bu sonuç hem veri ön işleme sürecinin hem de model eğitiminin etkin bir şekilde

gerçekleştirildiğini gösterdi. Modelin bu performansı, benzer görevler için güvenilir bir araç olabileceğini ortaya koydu.

Temizlenmiş veri seti, doğal dil işleme görevlerinde güçlü bir performans sunan ALBERT (A Lite BERT) modeli kullanılarak eğitilmiştir. Eğitim sürecinde veri seti %70 eğitim ve %30 test oranında bölünerek işlenmiştir. Bu süreçte ALBERT modelinin hafif ve optimize edilmiş yapısından yararlanılarak metinlerin anlamlandırılması ve sınıflandırılması süreçleri daha verimli hale getirilmiştir. Model, düşük bellek kullanımı ve hızlı işlem kapasitesi sayesinde eğitim sürecinde yüksek bir performans göstermiştir.

Eğitim süreci tamamlandıktan sonra model test aşamasına tabi tutulmuştur. Test sonuçlarına göre ALBERT modeli, 0.9727 doğruluk oranı (accuracy) elde ederek oldukça başarılı bir performans sergilemiştir. Bu sonuç hem veri ön işleme adımlarının hem de modelin mimari avantajlarının eğitim sürecine olumlu bir katkı sağladığını göstermektedir. ALBERT modelinin bu başarısı, benzer metin sınıflandırma ve anlamlandırma görevleri için etkili bir çözüm sunabileceğini ortaya koymaktadır.

IV. SONUÇLAR VE TARTIŞMA

Bu çalışma, sosyal medya gönderilerinde intihar içeriklerinin tespit edilmesi ve derin öğrenme modellerinin performanslarının karşılaştırılması konusunda önemli bir araştırma sunmaktadır. Reddit platformundan elde edilen iki veri seti, kapsamlı bir ön işleme sürecinden geçirilerek birleştirilmiş ve bu birleşim, dengeli ve geniş bir veri kümesi oluşturulmasını sağlamıştır. Çalışmada T5-small ve ALBERT modelleri gibi ileri düzey derin öğrenme yöntemleri kullanılmış, sırasıyla %97 ve %97,27 doğruluk oranları elde edilmiştir. Elde edilen sonuçlar, çalışmanın etkilerini ve derin öğrenme tekniklerinin bu alandaki gücünü gözler önüne sermektedir.

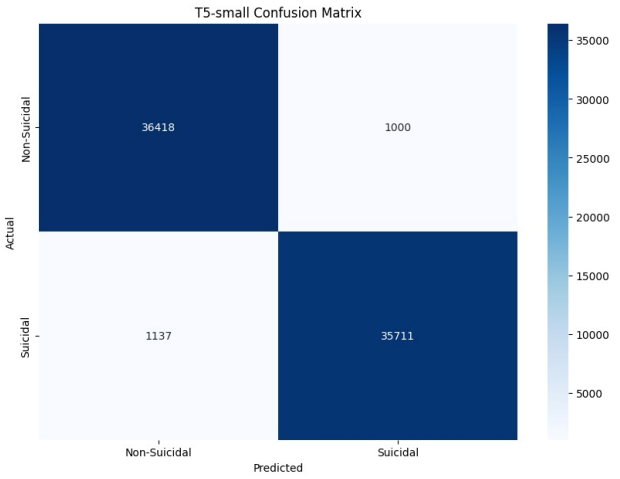
4.1. Veri Setleri ve Ön İşleme Sürecinin Etkileri

Birleştirilen veri setlerinin 247.551 gönderiden oluşması ve intihara meyilli ile meyilli olmayan içeriklerin eşit dağılım göstermesi, model performansını artıran en önemli faktörlerden biridir. Veri setlerinin doğal dil işleme teknikleriyle işlenmesi, özellikle gereksiz sembollerin, emojilerin ve anlamsız kelimelerin temizlenmesi, modelin daha doğru tahminler yapmasını sağlamıştır. Metin uzunluklarının analiz edilmesi ve temizleme sürecinin ardından dengeli bir veri kümesinin oluşturulması, sınıflar arası önyargıların azalmasına yardımcı olmuştur.

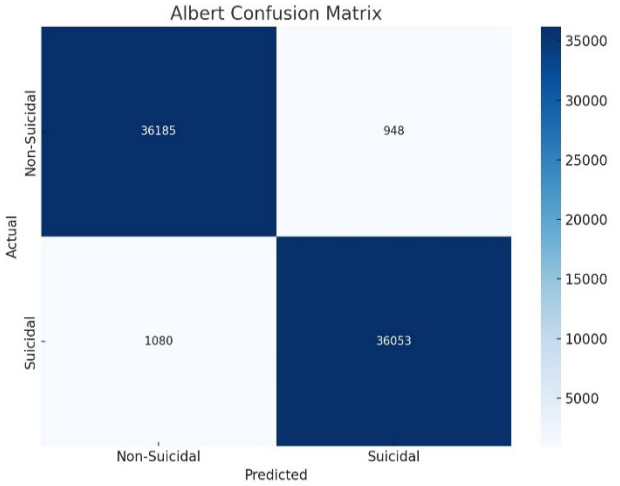
4.2. Kullanılan Modeller ve Performans Karşılaştırmaları

T5-small Modeli: Geniş bir görev yelpazesi için metinden-metine formatında çalışan bu model, %96 doğruluk oranı ile etkili bir performans göstermiştir. Özellikle düşük hesaplama maliyeti, bu modelin geniş çaplı uygulamalarda tercih edilebilirliğini artırmaktadır. T5-small modelinin karmaşıklık matrisi Şekil 7'de verilmiştir.

ALBERT Modeli: Daha hafif bir yapı ve parametre paylaşımı özellikleriyle optimize edilmiş olan bu model, %97,27 doğruluk oranı ile öne çıkmıştır. ALBERT'in hızlı işlem kapasitesi ve düşük bellek gereksinimi, modelin eğitim sürecinde verimli bir şekilde çalışmasını sağlamıştır. Albert modelinin karmaşıklık matrisi Şekil 8'de verilmiştir.



Şekil 7. T5-small modelinin karmaşıklık matrisi



Şekil 8. Albert modelinin karmaşıklık matrisi

4.3. Literatür ile Karşılaştırmalar

Bu çalışmada kullanılan modellerin doğruluk oranları, literatürdeki diğer yöntemlere kıyasla daha yüksek sonuçlar vermiştir. Örneğin, Twitter veri setlerinde kullanılan ANN ve MLP gibi eski yöntemlerle %65-91 doğruluk elde edilirken, çalışmamızdaki modeller bu oranları önemli ölçüde aşmıştır. Literatürde yer alan başka bir çalışmada BiLSTM modeli %93,6 doğruluk oranına ulaşırken, ALBERT modeli bu oranı daha da ileriye taşımıştır.

4.4. Sosyal ve Toplumsal Katkılar

Çalışma, sosyal medya verilerindeki intihar eğilimlerini erken tespit etmek için kullanılabilecek güçlü araçlar sunmaktadır. Bu araçların gerçek zamanlı bir uyarı sistemi olarak kullanılması, bireylerin erken teşhis edilmesini ve uygun müdahalelerin yapılmasını mümkün kılabilir. Çalışma, yalnızca teknik bir analiz sunmanın ötesinde, toplumsal faydayı ön plana çıkararak bireylerin hayatlarını iyileştirme hedefi taşımaktadır.

4.5. Gelecek Araştırmalar için Öneriler

Daha fazla sosyal medya platformunun (örneğin Facebook, Instagram) verilerinin dahil edilmesi, modellerin genelleştirme yeteneğini artırabilir.

Farklı dillerdeki içeriklerin analiz edilmesi, kültürel farklılıkların intihar içeriklerine etkisinin araştırılmasına olanak tanıyabilir.

Modellerin gerçek zamanlı bir sistemde uygulanarak performanslarının test edilmesi, bu teknolojilerin pratikteki etkinliğini değerlendirebilir.

V. SONUÇ

Bu çalışmada, sosyal medya gönderilerinde intihar içeriklerinin tespit edilmesi üzerine iki dil modelinin performansları karşılaştırılmıştır. Reddit platformundan elde edilen iki farklı veri seti, kapsamlı bir ön işleme sürecinden geçirilerek birleştirilmiş ve toplamda yaklaşık 240.000 gönderiden oluşan dengeli bir veri kümesi oluşturulmuştur. Bu veri kümesi, 60M parametreye sahip T5-small ve 12M parametreye sahip ALBERT modelleri ile analiz edilmiştir. Modellerin performansları doğruluk oranları ile ölçülmüş ve her ikisi de %97 doğruluk oranına ulaşmıştır. Bu sonuçlar, her iki modelinde bu görevde yüksek doğruluk oranına ulaşabildiğini ancak aralarında boyut farkına rağmen sonuçlarda kayda değer bir fark olmadığını göstermiştir.

KAYNAKLAR

[1] Suicide. (t.y.). Geliş tarihi 03 Ocak 2025, gönderen https://www.who.int/news-room/fact-sheets/detail/suicide

[2] Astoveza, G., Obias, R. J. P., Palcon, R. J. L., Rodriguez, R. L., Fabito, B. S., & Octaviano, M. V. (2018). Suicidal Behavior Detection on Twitter Using Neural Network. TENCON 2018 - 2018 IEEE Region 10 Conference, 0657-0662. https://doi.org/10.1109/TENCON.2018.8650162

[3] Haque, R., Islam, N., Islam, M., & Ahsan, M. M. (2022). A Comparative Analysis on Suicidal Ideation Detection Using NLP, Machine, and Deep Learning. Technologies, 10(3), 57. https://doi.org/10.3390/technologies10030057

[4] Sabri, N. M., & Mohamad, N. A. (2022). Detection of Suicidal Tweets Based On Naïve Bayes Algorithm. International Journal of Advanced Research in Technology and Innovation, 4(3), Article 3. https://myjms.mohe.gov.my/index.php/ijarti/article/view/20077

[5] Wang, N., Luo, F., Shivtare, Y., Badal, V. D., Subbalakshmi, K. P., Chandramouli, R., & Lee, E. (2021). Learning Models for Suicide Prediction from Social Media Posts (No. arXiv:2105.03315). arXiv. https://doi.org/10.48550/arXiv.2105.03315

[6] MacAvaney, S., Mittu, A., Coppersmith, G., Leintz, J., & Resnik, P. (2021). Community-level Research on Suicidality Prediction in a Secure Environment: Overview of the CLPsych 2021 Shared Task. İçinde N. Goharian, P. Resnik, A. Yates, M. Ireland, K. Niederhoffer, & R. Resnik (Ed.), Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access (ss. 70-80). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.clpsych-1.7

[7] Ophir, Y., Tikochinski, R., Asterhan, C. S. C., Sisso, I., & Reichart, R. (2020). Deep neural networks detect suicide risk from textual facebook posts. Scientific Reports, 10(1), 16685. https://doi.org/10.1038/s41598-020-73917-0

[8] Kodati, D., & Tene, R. (2023). Identifying suicidal emotions on social media through transformer-based deep learning. Applied Intelligence, 53(10), 11885-11917. https://doi.org/10.1007/s10489-022-04060-8

[9] Ghosh, S., Ekbal, A., & Bhattacharyya, P. (2022). A Multitask Framework to Detect Depression, Sentiment and Multi-label Emotion from Suicide Notes. Cognitive Computation, 14(1), 110-129. https://doi.org/10.1007/s12559-021-09828-7

[10] Lin, E., Sun, J., Chen, H., & Mahoor, M. H. (2024). Data Quality Matters: Suicide Intention Detection on Social Media Posts Using RoBERTa-CNN. 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 1-5. https://doi.org/10.1109/EMBC53108.2024.10782647

[11] Suicide and Depression Detection. (t.y.). Geliş tarihi 03 Ocak 2025, gönderen https://www.kaggle.com/datasets/nikhileswarkomati/suicide-watch

[12] Chatterjee, M., Samanta, P., Kumar, P., & Sarkar, D. (2022). Suicide Ideation Detection using Multiple Feature Analysis from Twitter Data. 2022 IEEE Delhi Section Conference (DELCON), 1-6. https://doi.org/10.1109/DELCON54057.2022.9753295

[13] Tadesse, M. M., Lin, H., Xu, B., & Yang, L. (2019). Detection of Suicide Ideation in Social Media Forums Using Deep Learning. Algorithms, 13(1), 7. https://doi.org/10.3390/a13010007

[14] Ji, S., Yu, C. P., Fung, S., Pan, S., & Long, G. (2018). Supervised Learning for Suicidal Ideation Detection in Online User Content. Complexity, 2018(1), 6157249. https://doi.org/10.1155/2018/6157249

[15] Aydın, T., & Irmak, M. C. (2023). ELECTRA ve XLNET Modellerini Kullanarak X Verilerinden İntihara Meyilli İçerikleri Tespit Etme. Cognitive Models and Artificial Intelligence Conference Proceedings, 45-50. https://doi.org/10.36287/setsci.6.1.019

[16] Ezerceli, Ö., & Dehkharghani, R. (2024). Mental disorder and suicidal ideation detection from social media using deep neural networks. Journal of Computational Social Science, 7(3), 2277-2307. https://doi.org/10.1007/s42001-024-00307-1

[17] Aldhyani, T. H. H., Alsubari, S. N., Alshebami, A. S., Alkahtani, H., & Ahmed, Z. A. T. (2022). Detecting

and Analyzing Suicidal Ideation on Social Media Using Deep Learning and Machine Learning Models. International Journal of Environmental Research and Public Health, 19(19), 12635. https://doi.org/10.3390/ijerph191912635

[18] Mafi, M. M. H. M. (2023). Suicidal Ideation Detection Reddit Dataset [Dataset]. Mendeley Data. https://doi.org/10.17632/Z8S6W86TR3.1

[19] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2023). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer (No. arXiv:1910.10683). arXiv. https://doi.org/10.48550/arXiv.1910.10683

[20] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations (Versiyon 6). arXiv. https://doi.org/10.48550/ARXIV.1909.11942