Konya Mühendislik Bilimleri Dergisi, c. 10, s. 4, 986-1001, 2022
Konya Journal of Engineering Sciences, v. 10, n. 4, 986-1001, 2022
ISSN: 2667-8055 (Elektronik)
DOI: 10.36306/konjes.1061807

TRANSFORMATÖR-TABANLI EVRİŞİMLİ SİNİR AĞI MODELİ KULLANARAK TWITTER VERİSİNDE SALDIRGANLIK TESPİTİ

Erdal ÖZBAY

Fırat Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Elazığ, TÜRKİYE
erdalozbay@firat.edu.tr

(Geliş/Received: 23.01.2022; Kabul/Accepted in Revised Form: 13.10.2022)

ÖZ: Çevrimiçi ortamlar, insanların sosyal etkileşimlerinde anti-sosyal davranışların artmasını kolaylaştırmaktadır. Sosyal medya kullanımının yaygınlaşmasıyla özellikle son yıllarda nefret söylemleri, siber zorbalık ve trolleme gibi davranışlar önemli ölçüde artmıştır. Saldırgan ve nefret içerikli söylemlerin tespiti siber zorbalıkların azaltılması ve engellenmesinde önemli bir adımdır. Siber zorbalık, sosyal medya üzerinden nefret dolu, saldırgan, kaba, aşağılayıcı ve alaycı ifadeler kullanarak diğer bireylere zarar vermek adına yapılan yorumlar olarak adlandırılmaktadır. Hızla büyüyen verilerin varlığı, bunun insan denetimiyle gerçekleştirilmeye çalışılması yavaş ve pahalı olduğundan saldırganlığın otomatik tespitiyle siber zorbalığın durdurulması sağlanabilir. Bu çalışmada Twitter veri seti olan Cyber-Trolls üzerinden saldırganlık tespitini otomatik olarak belirlenmesi ele alınmaktadır. LMTweets adında bir kodlayıcı, veri kümesinin özelliklerinin çıkarılması için 20001 adet tweet üzerinden eğitilmiştir. Çıkarılan öznitelikler, metni saldırgan / saldırgan olmayan olarak sınıflandırmak üzere evrişim sinir ağı modeline girdi olarak verilir. Ayrıca Naïve Bayes, Destek Vektör Makinesi, K-En Yakın Komşu, olmak üzere üç sınıflandırma algoritması uygulanmıştır. Bunun yanında, Evrişimli Sinir Ağı, Uzun Kısa-Süreli Bellek ve Kapılı Tekrarlayan Birim üç öğrenme algoritması ile birlikte BERT, XLNet ve ULMFIT olmak üzere üç transformatör modeli uygulanmıştır. Önerilen modelde Python, Keras API ve Tensorflow birlikte kullanılmıştır. Deneysel sonuçlarda elde edilen performans parametreleri doğruluk, kesinlik, duyarlılık, F_1-ölçütü ve AUC olarak belirlenmiş ve LMTweets + CNN modelinin kullanılan tüm modeller arasında daha iyi performans gösterdiği ortaya konmuştur.

Anahtar Kelimeler: *Saldırganlık, CNN, Derin Öğrenme, Twitter, Transformatör Modeller*

Aggression Detection in Twitter Data Using Transformer-Based Convolutional Neural Network Model

ABSTRACT: Online environments facilitate the increase of anti-social behaviors in people's social interactions. Behaviors such as hate speech, cyberbullying, and trolling have increased significantly, especially in recent years, with the widespread use of social media. Detection of aggression and hateful speech is an important step in reducing and preventing cyberbullying. Cyberbullying is defined as comments made on social media to harm other individuals by using hateful, offensive, rude, humiliating, and sarcastic expressions. It is slow and expensive to try to achieve this with human control with the existence of rapidly growing data, so cyberbullying can be stopped by automatic detection of aggression. In this study, the automatic determination of aggression detection via Cyber-Trolls, which is the Twitter dataset, is discussed. A coder named LMTweets was trained on 20001 tweets to extract the features of the dataset. The extracted features are given as input to the convolutional neural network model to classify the text as aggressive / non-aggressive. In addition, three classification algorithms, namely Naïve Bayes, Support Vector Machine, K-Nearest Neighbors, were applied. In addition, three transformer models, BERT, XLNet, and ULMFIT were applied along with the Convolutional Neural Network, Long Short-Term Memory, and Gated Recurrent Unit three learning algorithms. Python, Keras

API, and Tensorflow are used together in the proposed model. The performance parameters obtained in the experimental results were determined as accuracy, precision, recall, F1-score, and AUC, and it was revealed that the LMTweets + CNN model performed better among all the models used.

Keywords: *Aggression, CNN, Deep Learning, Twitter, Transformer Models*

GİRİŞ (INTRODUCTION)

Web son birkaç yıl öncesinden itibaren salt okunur bir platform olmaktan çıkıp kademeli olarak kullanıcılar tarafından oluşturulan bir web'e evrimleşmiştir. Kullanıcı yorumları bazı çevrimiçi forum, blog ve sosyal medya web sitelerinde yayınlanmakta ve diğer kullanıcılar ile etkileşime girmektedir. Erişim kolaylığı ve çevrimiçi topluluğun büyümesi küresel iletişimin artışına neden olmuştur. Günümüzde, Twitter ve Facebook gibi sosyal medya platformları aracılığıyla çeşitli ürünler, hizmetler ve bilgi birikimleri küresel olarak insanlarla paylaşılabilir hale gelmiştir. Ancak bu özgürlük beraberinde büyük bir risk doğurmaktadır. İnsanlar paylaştığı çok daha az kişisel bilgiyle veya anonim olarak çeşitli yorumlarda bulunup saldırgan bir tutum sergileyebilmektedir (Grigg ve diğ., 2010).

Çeşitli sosyal medya platformlarındaki siber zorbalığın etkisinin artmasındaki en önemli neden güvenlik açıklarıdır (Smit ve diğ., 2015). Bunun yanında, sahte kimliklerin varlığı ve online hesaplardaki anonimlik, saldırganlığın kontrol edilemez bir hızla artmasına sebep olmuştur. Bu olaylar milyonlarca insanın hayatını psikolojik veya zihinsel travmalarla etkilemenin yanında kişilerin intihar etmesine kadar gitmektedir. Siber saldırganlık, kişilere kasıtlı olarak zarar vermeyi hedefleyen, bir birey veya grubun elektronik ortamda saldırgan, istenmeyen veya zararlı olmak üzere tekrarlanan düşmanca davranışlar olarak nitelendirilmektedir (Grigg ve diğ., 2010). Saldırganlık, cinsiyet, ırk, renk, milliyet, etnik köken ve din gibi özellikler temelindeki nefret söylemlerini içermektedir (John ve diğ., 2000).

Günümüzde Twitter üzerindeki aktif bir kullanıcı günde ortalama 500 milyon tweet ile çeşitli hizmetler, ürünler ve devlet politikaları ve şirketler/kuruluşlar, politikacılar için ana inceleme / öneri / geri bildirim bakımından etkileşimde bulunmaktadır. Pazar payını artırmak ya da güncel politikalar tasarlayabilmek adına gerekli adımları atabilmeleri için sosyal medya üzerinden mevcut görüşlerin analizine başvurulur. Twitter kullanıcıları görüşlerini, tweet adı verilen 280 karakter uzunluğunda kısa mesajlar vasıtasıyla iletmektedir. Tweetler, gerçek (literal) veya mecazi / gerçek olmayan (non-literal) tweetler olmak üzere iki ayrı sınıfta değerlendirilmektedir (Abulaish ve diğ., 2020). Mecazi ifadeler insanların düşüncelerini yanıltmak amacıyla yazılmış aslında yazıldığı gibi olmayan ifadelerdir. Gerçek tweetlerin duygu analizi, yorumlarda standart kelimeler bulunduğundan daha kolaydır. Ancak gerçek olmayan tweetlerin duygu analizi yorumlardaki mecazi dilin varlığı nedeniyle nispeten zordur. Figüratif dil, fikir ve düşüncelerin, gramer ve mantıksal ifade tarzlarının dışındaki türevleriyle, canlılık ve güçle aktarılması olarak tanımlanmaktadır (Hepburn, 1875).

Sosyal medya üzerinde günlük tüketilen veri miktarı her geçen gün artmakla birlikte çok büyük bir veri birikimi oluşmaktadır. Saldırganlık içeren davranışların tespitini elle işletilerek (manual) gerçekleştirmek büyük veri dezavantajından dolayı pratik değildir. Bu durum bizleri otonom veya yarı otonom sistemleri geliştirmeye yöneltmektedir. Saldırganlık önleyici tedbirler her ne kadar sosyal medya ağı ve devletler aracılığıyla alınmaya çalışılsa da bu davranışların azaltılması adına etkili çözümlere ihtiyaç duyulmaktadır. Bu sorun Twitter'da çeşitli doğal dil işleme araçları, metin madenciliği ve çeşitli makine öğrenmesi yaklaşımları ile giderilmeye çalışılsa da dilbilgisi ve sözdizimsel kusurlar ile birlikte kısa tweet uzunluklarının varlığı nedeniyle özniteliklerin çıkarılmasında zorluklar yaşanmaktadır (Van der Walt ve diğ., 2018). Sosyal medyada tacizci kullanıcıları filtrelemek zorlu bir görev olabilir çünkü trolleme ya da alaycılık gibi yollarla saldırganlık ve zorbalık göstermenin farklı yolları bulunmaktadır. Çevrimiçi sosyal medyadaki mecazi dil kategorilerinden bazıları alaycılık, ironi, benzetme, metafor, hiciv, komedi ve abartıdır (Joshi ve diğ., 2017).

Doğal dil işleme görevleri arasında saldırganlık tespiti önemlidir ve temel adımı metni işleyerek nihai hedefe göre analiz etmek ve beklenmedik bir şekilde bilgi çıkarmaktır. Günümüzde metin sınıflandırma, Makine öğrenmesi sınıflandırıcılarında en iyi özellik çıkarımı hakkında bilgi talep

etmektedir. Bir cümleyi kelime torbası kullanarak örneklemeye, ardından Destek Vektör Makinesi (Support Vector Machine - SVM) veya Naïve Bayes (NB) aracılığıyla sınıflandırma yapmaya yönelik çok geleneksel yöntemler bulunmaktadır (Joachims ve diğ., 1998). Ancak yüz ifadesi ve ses tonu olmadan bu duyguları anlamak zordur. Metinsel verilerden duyguları anlamanın bu zorluğu aynı zamanda sınıf dengesizliğinden, Twitter'da sınırlı metin uzunluklarından, internet argosunun kullanımından, alaycılık ve doğal dil belirsizliklerinden de kaynaklanmaktadır. Ancak sağlam derin öğrenme algoritmaları ile bu sorunun üstesinden gelinebilir. Özellikle son birkaç yıldır Evrişimli Sinir Ağı (Convolutional Neural Network - CNN) büyük veri kümeleri için karmaşık modeller geliştirmede büyük ilerleme kaydetmiştir. Derin Sinir Ağı (Deep Neural Network - DNN) teknikleri, sınıflandırmada öznitelik çıkarma yöntemlerini hayata geçirmektedir. Bir belgedeki kelime dizileri, DNN yöntemlerinde ağırlıkla çarpıldıktan sonra sıcak vektörler (hot vectors) olarak temsil edilmektedir. Girdi aldıktan sonra, sinir ağları sıcak vektörleri sıralı gizli katmanlarla besleyerek tahminlerde bulunmaktadır (Shen ve diğ., 2014). Sinir ağı modelleri, Çok-katmanlı Algılayıcı (Multi-layer Perceptron - MLP), CNN, Tekrarlayan Sinir Ağı (Recurrent Neural Network - RNN) ve bunların türevleri, doğru modeli seçerek ve hiper parametre ayarları ile metin sınıflandırmasında sağlam sonuçlar elde etmektedirler.

Bu çalışmada Twiter metin verileri üzerinden saldırganlık tespiti ele alınmaktadır. Transformatörlerden çift yönlü kodlayıcı temsilleri (Bidirectional Encoder Representations from Transformers - BERT) tabanlı mimariye dayalı Linguistics-Models (LMTweets) adlı alana özgü bir transformatör modeli önerilmektedir. LMTweets, bir dizi bağlamsal yerleşim oluşturmak ve dizideki her bir kelime için bir öz-dikkat mekanizması aracılığıyla bağlamsal bilgiyi elde etmede kullanılmaktadır. LMTweets tarafından oluşturulan bu kelime yerleştirmeleri, metni sınıflandırmak için CNN'e iletilmektedir. Literatürdeki mevcut çalışmaların dezavantajları arasında; esas olarak büyük veri kümeleri ile çalışmaları, modellerini kıyaslamada veri kümeleri yerine uygulama programlama arayüzü kullanarak oluşturdukları veri kümeleriyle deneysel sonuçları elde etmeleridir (Kumar ve diğ., 2019). Bu nedenlerle onların modellerini karşılaştırmak ve değerlendirmek pek sağlıklı değildir. Birçok yaklaşım, kelimeleri büyük sözlükler içerisinde aramaktadır, bu durum daha fazla zaman harcanmasına neden olduğundan pratik değildir (Potamias ve diğ., 2019). Bu çalışmada metin ön işlemenin gerekli olmadığı, alana özgü transformatör tabanlı bir kodlayıcı geliştirerek bu sınırlamaları kaldırmaktayız.

Bu çalışmamızın başlıca katkıları şöyledir; CNN ile LMTweets adlı bir model önerilerek Twitter veri kümesinde saldırganlık tespiti performansı iyileştirilmektedir. Makine öğrenmesi, derin öğrenme ve transformatör tabanlı olmak üzere farklı modeller ile önerilen yöntemin performansı veri seti üzerinde test edilerek kıyaslanmaktadır. Önerilen model (LMTweets + CNN), literatürde sunulan modellerden daha iyi performans gösterdiği gözlemlenmiştir.

Makalenin geri kalan kısımları şöyle düzenlenmiştir; 2. bölümde konuyla ilgili çalışmalara, 3. bölümde önerilen yöntemin açıklanmasına, 4. bölümde deneysel sonuçlar ve tartışmaya, 5. bölümde sonuçlara yer verilmiştir.

İLGİLİ ÇALIŞMALAR (RELATED WORKS)

Sosyal medyada nefret ifadeleri, saldırganlık ve zorbalık gibi tüm davranışlar trolleme olarak ifade edilmektedir. Bir metin içerisinde saldırganlığı ele almak karmaşık bir olgudur ve birçok alanda bu konuyla ilgili çözüm yolları ortaya konmuştur. Bilimsel topluluklarda saldırganlığı temsil etmek için farklı terminolojiler kullanılmaktadır. Bu çalışmalardan biri olan siber zorbalık Dinakar ve diğ. tarafından modellenmiştir (Dinakar ve diğ., 2012). Bunun yanında, trolleme (Mihaylov ve diğ., 2015), ırkçılık (Greevy ve diğ., 2004), müstehcenlik (Su ve diğ., 2017), aşırılık (Prentice ve diğ., 2011), hakaret (Bansal ve diğ., 2012), küfürlü dil (Mubarak ve diğ., 2017) ve nefret söylemi (Schmidt ve Wiegand, 2017) gibi konularda da mevcut yaklaşımlar ortaya konmuştur.

Semeval-2019 etkinliğindeki katılımcıların İngilizce ve İspanyolca Twitter mesajlarından geleneksel makine öğrenmesi ve çoğunlukla derin öğrenme yaklaşımları kullanılarak kadın ve göçmenlere yönelik nefret söylemlerinin tespiti gerçekleştirilmiştir. Basile ve diğerlerinin çalışmasında girdi olarak

kullanılan tweet'lerde önceden eğitilmiş kelime gömme (word embedding) ve model olarak RNN tercih edilmiştir (Basile ve diğ., 2019).

Sözlü saldırganlık, açık veya gizli saldırganlık olarak ifade edilebilmektedir. Açık saldırganlık bazı sözdizimsel yapılar tarafından doğrudan ifade edilirken, gizli saldırganlık dolaylı bir saldırı şeklidir. Coling-2018'de saldırganlığın belirlemesinde, kullanıcı gönderilerinin saldırganlık düzeylerini sınıflandırmak için trollük, saldırganlık ve siber zorbalık konulu bir çalıştay yapılmıştır. Bu görev için sınıflandırıcı, açık saldırganlık, gizli saldırganlık ve saldırgan olmayan metinler arasında ayrım yapmaktadır. Veri seti olarak saldırganlık açıklamalı 15000 adet Facebook yorumu veya gönderisi kullanılmıştır. Geliştirilen sistemlerin performansları, saldırganlık tespitinin zorlu bir görev olduğunu ortaya koymaktadır (Kumar ve diğ., 2018). Otomatik saldırganlık ve siber zorbalık algılama / sınıflandırma problemlerinde, çevrimiçi kullanıcı yorumları ve metinleri çoğunlukla zorbalık ve zorbalık dışı olarak ayırt edilmektedir (Salawu ve diğ., 2017). Salawu ve diğ. bir sosyal medya veri akışındaki bireysel saldırgan mesajların belirlenmesi, saldırganlığın ciddiyetinin değerlendirilmesi, ilgili bireylerin rollerinin belirlenmesi ve saldırganlık olayının bir sonucu olarak meydana gelen olayların sınıflandırılması gibi siber zorbalık tespitindeki dört farklı görevi içeren bir çalışma ortaya koymuşlardır (Salawu ve diğ., 2017).

Siber zorbalığın otomatik tanımlanmasında birinci adım olarak saldırganlık tespitini içermektedir. Chatzakou ve diğ. tarafından yapılan bir çalışmada Twitter üzerindeki zorbalık ve saldırganlık içeren davranışlar gösterilmektedir (Chatzakou ve diğ., 2017). Önerilen yaklaşımda üç aylık süreçle edinilen 1,6 milyon tweet üzerinden deney sonuçlarını ortaya koymuşlardır. Kullanıcı tabanlı, metin tabanlı ve ağ tabanlı özelliklerle makine öğrenmesi algoritmasını kullanarak %90 Eğri Altında Kalan Alan (Area Under Curve - AUC) elde etmişlerdir. Kullanıcı söylemlerini kabadayı, saldırgan, spam ve normal kullanıcılar olarak sınıflandırmışlardır. Saldırganların diğer gruplar arasında en popüler olduğunu ve nefret dolu yorumlar ve troller yayınlayarak olumsuzlukları yaydıklarını ortaya koymuşlardır.

Bir diğer çalışmada Nobata ve diğ. Yahoo!'nun 2 milyon çevrimiçi yorumundaki nefret söylemini tespit etmeyi amaçlamıştır. Bu çalışmada finans ve haber içeriklerinden dilbilimsel, sözdizimsel, n-gramlar ve gömülü anlamsal özellikler bakımından dört tür özellik dikkate alınmıştır. Buradaki yorumlar ve sayıları normalleştirilerek, bilinmeyen sözcükler aynı simgeyle değiştirilerek, tekrarlanan noktalama işaretleri değiştirilerek ön-işlenmiştir. Deneysel sonuçlar, tüm özelliklerin birleştirilmesinin en iyi F-ölçütü sonuçlarını elde ettiğini göstermiştir (Nobata ve diğ., 2016). Benzer şekilde, Chavan ve diğ. saldırganlık ve saldırganlık içermeyen yorumları ayırt etmek üzere seçilen özellikleri, TF-IDF ağırlıklı n-gramlar, ve zamirler içermekteydi. Chavan ve diğerlerinin yaptığı çalışmada yalnızca en yüksek 3000 adet özellik seçilmiştir. Deneysel sonuçlar, içeriği belirsiz bir siteden gelen yaklaşık 6500 yoruma dayanmaktadır. Yorumlardaki kelime sayılmayacak karakterler, kısa çizgiler ve noktalama işaretleri kaldırılarak ön-işlenmiştir. Beraberinde, olası yazım hatalarını düzeltmek adına bir yazım denetleyicisi uygulanmıştır. Deneyler, zamirler ve atlama gramları dikkate alındığında en iyi performansın elde edildiğini göstermiştir (Chavan ve diğ., 2015).

Karakter düzeyinde metin sınıflandırması Xiao ve diğ. tarafından üst düzey özellikleri öğrenmek için CNN ve RNN kullanılarak gerçekleştirilmiştir (Xiao ve diğ., 2016). Benzer bir görevi Tai ve diğ. metin anlambilimini geliştirmek için Uzun Kısa-Süreli Bellek (Long Short-Term Memory - LSTM) ile cümle düzeyinde gerçekleştirmiştir (Tai ve diğ., 2015). Tai ve diğerlerinin önerdiği CNN, cümlelerdeki ardışık sıradaki ifadelerden veya kelimelerden yerel özellikleri çıkarmaktadır. Tweet sayısı, takipçi sayısı ve kullanıcı sözleri gibi metinsel özellikler, saldırganlığın otomatik tespitini geliştirmektedir (Al-Garadi ve diğ., 2016). Kullanıcıların kişilik özellikleri ayrı ayrı birleştirilmesinde Rastgele Orman (Random Forest - RF) kullanılarak siber saldırganlığın daha iyi tespit edilebildiğini ortaya koymuşlardır (Balakrishnan ve diğ., 2019). Balakrishnan ve diğerleri çalışmalarında Big-Five ve Dark-Triad modellerini kullanmışlardır. Saldırgan, spam gönderen, zorba ve normal olmak üzere dört farklı rolü ele alarak kullanıcı kişiliği ile siber saldırganlık tutumu arasında güçlü bir ilişki olduğunu kanıtlamışlardır.

Doğal dil işleme görevleri de DNN tabanlı modeller tarafından gerçekleştirilmektedir. Saldırganlık tespiti önem kazandıkça bu sorun da derin öğrenme yaklaşımları kullanılarak çözülmeye

çalışılmaktadır. Gambäck ve Sikdar tarafından nefret söylemini sınıflandırmak için derin öğrenme yaklaşımı olan Max-Pooling'li CNN'den faydalanılmıştır (Gambäck ve Sikdar, 2017). Karakter n-gramları ile kelime vektörleri öznitelik olarak kullanılmıştır. Madisetty ve Desarkar tarafından CNN, LSTM ve Çift-Yönlü (Bi-LSTM) birleştirilerek topluluk yöntemi (çoğunluk oyu) kullanılmıştır (Madisetty ve Desarkar, 2018). Deneysel sonuçlarını Facebook ve sosyal medya yorumları üzerinden gerçekleştirmişlerdir. Sosyal medya gönderilerindeki saldırganlığın otomatik tespitinde CNN, LSTM, Bi-LSTM, LSTM-CNN, CNN-LSTM, Bi- LSTM-CNN, CNN-Bi-LSTM olmak üzere yedi derin öğrenme modeli ile farklı deneysel sonuçlar elde edilmiştir. Buna göre LSTM'e dayalı sınıflandırıcı, en iyi ağırlıklı F_1 ölçütü olarak 0,6425 değerini elde etmiştir (Aroyehun ve Gelbukh., 2018).

Literatürde birçok çalışmada Twitter verileri kullanılarak saldırganlık ve türevlerinin tespitinde makine öğrenmesi algoritmaları kullanılmıştır. RF, SVM, NB ve Lojistik Regresyon (LR)'un Twitter'da saldırganlığın tahmin edilmesinde yaygın olarak kullanılan algoritmalar olduğu ortaya konmuştur (Farías ve diğ., 2018, Chia ve diğ., 2019, Sarsam ve diğ., 2020). Bununla birlikte yine birçok çalışmada transformatör tabanlı yaklaşımlar ele alınmıştır. Bunlardan birinde metin gösterimi için RoBERTa transformatör modeli kullanılmış ve metnin sınıflandırılması için Bi-LSTM modeli bunun üzerine uygulanmıştır. Önerilen model dört farklı veri kümesine uygulanmış ve NB, SVM, XLNet, Bert-tabanlı, RoBERTa, ELMo ve Fast-Text olmak üzere çeşitli modellerle karşılaştırılmıştır (Potamias ve diğ., 2020). Avvaru ve diğ. Twitter ve Reddit'te LSTM modellerinin (LSTM, yığılmış (stacked) LSTM, Bi-LSTM ve CNN-LSTM) ve BERT ve XLNet transformatör modellerinin çeşitli versiyonlarını uygulamışlardır (Avvaru ve diğ., 2020). Bir başka çalışmada ise Twitter ve Reddit veri kümelerinde alaycılığın tespit edilmesi için LSTM modeli, BERT, XLNet, ALBERT ve RoBERTa gibi çeşitli transformatör modelleri ile birleştirilmiştir. Transformatör topluluğu modelinin uygulanan tüm modellerden daha iyi performans gösterdiğini ortaya koymuşlardır (Gregory ve diğ., 2020). Mevcut çalışmaların özeti Çizelge 1'de ayrıntılı olarak gösterilmektedir.

Khan ve diğ.'nın 2022 yılında yaptığı bir çalışmada, saldırganlık tespiti için Cyber-Troll veri kümesi kullanılarak word embedding ve sekiz farklı duygusal özelliğin birleşimiyle oluşan beslemeyle yürüttükleri DNN ile 97% F_1 ölçütü sonucunu elde etmişlerdir (Khan ve diğ., 2022). Aynı yıl yapılan bir diğer çalışmada, sosyal medya üzerinden saldırganlık tespiti için CNN ve Bi-LSTM ve olmak üzere iki farklı sınıflandırmanın kullanıldığı bir derin öğrenme modeli önerilmiştir (Pareek ve diğ., 2022).

Başka bir çalışmada, Bengalce'deki saldırgan metinleri tanımlamak ve sınıflandırmak için temel sınıflandırıcılar olarak m-BERT, distil-BERT, Bangla-BERT ve XLM-R'yi içeren ağırlıklı bir topluluk tekniği önerilmiştir. Bu çalışmada %93.43 ile en yüksek F_1 ölçütü sonucuna ulaşılmıştır (Sharif ve Hoque., 2022).

Çizelge 1. Siber Saldırganlık / Zorbalık tespiti çalışmaları
Table 1. *Cyber Aggression / Bullying detection studies*

Yazar	Yöntem	Veri Kümesi	Özellik Gösterimi
Tulkens ve diğ., 2016	SVM	Yahoo!	Sözdizimsel, anlamsal, gömme
Al-Garadi ve diğ., 2016	NB, RF , KNN	Manual tweet	Ağ, aktivite ve kullanıcı
Saravanaraj ve diğ., 2016	NB ve RF	Tweet	Word2vec
Djuric ve diğ., 2015	NL modeli	Yahoo!	Paragraph2vec
Chatzakou ve diğ., 2017	RF	Rastgele tweet	Kullanıcı, metin ve ağ tabanlı
Balakrishnan ve diğ., 2019	RF	Twitter	Big-five and dark-triad
Gambäck ve Sikdar, 2017	CNN	Tweet	Word2vec, character, 4-gram
Davidson ve diğ., 2017	LR-L2 reg	Rastgele tweet	TF-IDF
Risch ve Krestel., 2018	LR ve SA	Facebook post	Gömme, n-gram, sözdizimsel
Aroyehun ve Gelbukh, 2018	CNN,LSTM,Bi-LSTM	Facebook post	Fast-Text model
Khan ve diğ., 2022	DNN	Twitter	TF-IDF, Word2Vec

Özet olarak literatürdeki mevcut çalışmalarda yazarlar, çevrimiçi saldırganlık tespiti için çeşitli makine öğrenmesi, derin öğrenme, topluluk yöntemleri ve önceden eğitilmiş transformatör modellerini kullanmışlardır. Etki alanına özgü veriler için hiçbir çalışmada transformatör modelleri eğitilmemiştir. Alana özel metinle ilgili olarak burada önerilen transformatör modeli, saldırganlık algılama performansını iyileştirebilmektedir. Bu boşluğu gidermek için, verileri sayısal değerlere dönüştürmede yalnızca sosyal medya metnine dayalı bir LMTweets dönüştürücü (kodlayıcı) modeli hazırlanmıştır. Daha sonra bunları saldırgan / saldırgan olmayan olarak sınıflandırmak için CNN modeli uygulanmıştır.

MATERYAL VE METOT (MATERIAL AND METHOD)

Sosyal medya kullanıcı telaffuzlarında, argo, kaba sözler ve gerçek olmayan kelimelerin yaygın olarak bulunması nedeniyle saldırganlık içeriklerinde, metafor gibi mecazi bir dil kullanımı yaygındır. Bu nedenle bu çalışmada en çok kullanılan sosyal medyalardan en önde geleni Twitter veri seti kullanılmıştır. Bu bakımdan önce deneyler için kullanılan veri seti sunulmakta, ardından önerilen mimarinin uygulaması ve tüm parametre ayarları tartışılmaktadır. Daha sonra, deneysel sonuçlar ve tartışmayı ele almaktayız, ardından önerilen yaklaşım saldırganlık tespiti alanındaki en gelişmiş metodolojiler ile karşılaştırılmaktadır.

Cyber-Trolls Veri Seti (Cyber-Trolls Dataset)

Cyber-Trolls veri seti, Data Turks tarafından metin sınıflandırma amacıyla oluşturulmuştur. Saldırgan trollere yardım etmek veya onları önlemek için tweet'leri saldırgan veya saldırgan olmayan olarak sınıflandırmada kullanılan bir veri setidir. Kaggle platformu üzerinden genel paylaşıma açık olan bu veri setinde iki ayrı kategoriden oluşmaktadır:

- Siber Saldırgan (SS) - 1
- Siber Saldırgan Olmayan (SSO) - 0

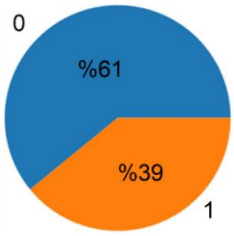
Veri kümesi, Çizelge 2'de gösterildiği gibi 7822'si Siber Saldırgan ve 12179'u Siber Saldırgan Olmayan olmak üzere toplamda 20001 adet tweet'e sahiptir. Şekil 1'de tüm veri setinin pasta olarak gösterildiği üzere Siber Saldırgan tweet oranı %39 ve Siber Saldırgan Olmayan tweet oranı %61'dir. Bu veri kümesi elle işletilmiş insan etiketli bir veri kümesidir. Siber Saldırgan: Tweet'in içeriği (kelimeleri) siber saldırganlık davranışı göstermektedir. Sosyal medya kullanıcılarından birinin, başkalarına zarar verme veya hakaret etme niyetiyle olumsuz anlamlar içeren paylaşımlarda bulunduğu anlamına

gelmektedir. Siber Saldırgan Olmayan: Tweet'in içeriği (kelimeleri) siber saldırganlık davranışı göstermemektedir. Kullanıcının, başkaları için olumsuz bir anlamı veya hakareti olmayan yorumlarını yayınlamaktadır.

Çizelge 2. Cyber-Troll veri kümesi

Table 2. *Cyber-Troll Dataset*

Veri Seti	Toplam Tweet	Siber Saldırgan - 1	Siber Saldırgan Olmayan - 0
Cyber-Trolls	20001	7822	12179

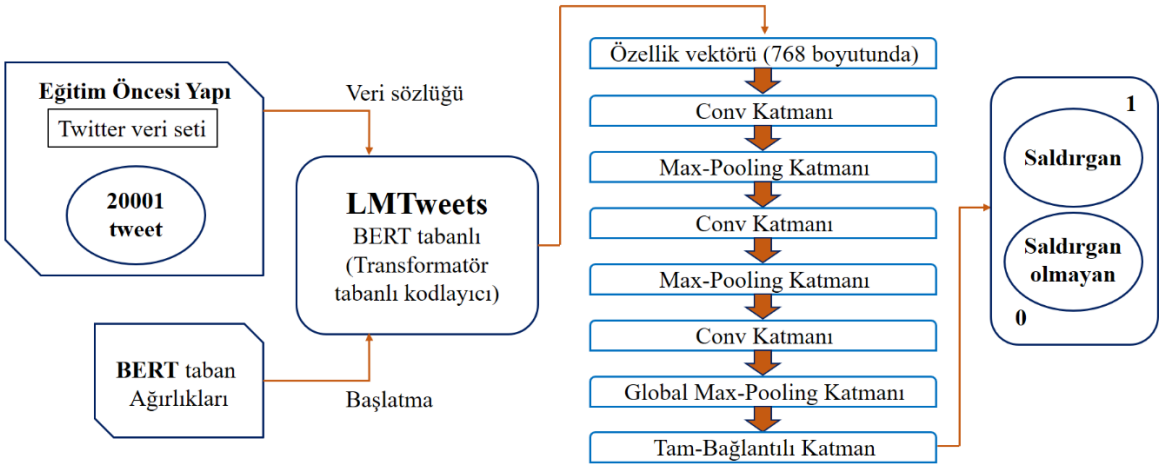


Şekil 1. Cyber-Troll veri kümesi dağılımı

Figure 1. *Cyber-Troll dataset distribution*

Önerilen Yöntem (Proposed Method)

Önerilen yöntem bu bölümde açıklandığı üzere iki aşamadan oluşmaktadır. Önerilen modelin mimarisine Şekil 2'de yer verilmiştir.

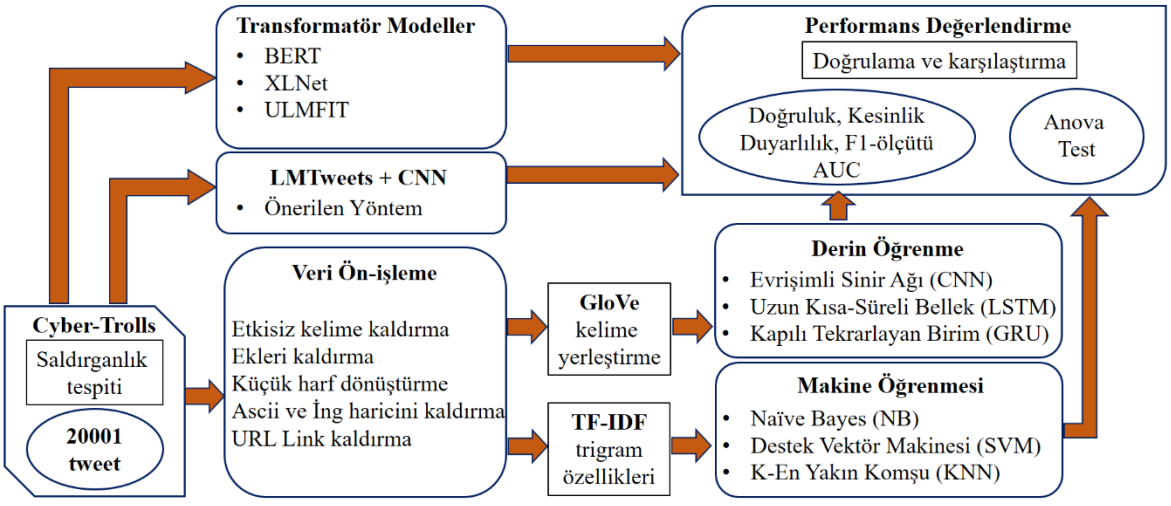


Şekil 2. Önerilen yöntemin LMTweets mimarisi

Figure 2. *LMTweets architecture of the proposed method*

İki aşamadan oluşan önerilen yöntemin mimarisinin ilk aşamasında LMTweets modelinin tasarlanması yer almaktadır. Bu aşamada yer alan LMTweets, yalnızca çok katmanlı transformatör tabanlı çift yönlü dil modeli olan bir kodlayıcıdan oluşmaktadır. Bu yaklaşım bir bakıma BERT mimarisiyle benzeşmektedir. BERT, İngilizce Wikipedia ve Books Corpus'un genel bilgi külliyatı üzerinden önceden eğitildiğinden, sosyal medya, tıbbi ve bilimsel metinler gibi alana özgü görevlerde genellikle düşük performans göstermiştir. Bu durum kısmen, alana özgü ve genel belgeler arasındaki sözcüklerdeki, tümceciklerdeki ve diğer dilsel özelliklerdeki büyük farklılıklardan kaynaklanmaktadır. LMTweets temel modelinin eğitimi için 20001 tweet kullanılmıştır. BERT tarafından ima edilen ön eğitimin hedefi, bir sonraki cümlenin tahmin görevi ile birlikte maskeli bir dil modeli oluşturmaktır. Aynı konfigürasyonu kullanmak yerine, eğitim verilerini daha doğru bir şekilde tamamlayan farklı bir eğitim prosedürü kullanılmıştır. Temel model ilk olarak, BERT tabanlı kontrol noktaları ile başlatılmış

ve bu ön eğitim için başlangıç noktası olarak belirlenmiştir. Modelin konfiürasyonu şöyledir; transformatör block sayısı: 12, hidden layers: 768, self-attention heads sayısı:12. Sözcük boyutunun küçültülmesinde, sözcük dağarcığındaki sözcüklerin verimli bir şekilde işlenebilmesi için hızlı Bayt-Çifti Kodlayıcı (Byte-Pair Encoding - BPE) tokenizer kullanılmıştır. Ön eğitim aşamasında, BERT'deki gibi sözcükleri rastgele maskelemek yerine, maskeleme düzeninin rastgele metin parçaları için değiştirildiği ve ayrıca maskelenmiş belirteçlerin yüzdesinin %30 ile %50 aralığında değiştirildiği alternatif bir strateji kullanılmıştır. Daha düşük maskeleme aralıkları modelin dili anlama kapasitesini olumsuz etkileyebilmektedir. Geniş maskeleme aralıklarında ise ön eğitim ve ince ayar görevleri arasında uyumsuzluklara yol açabilmektedir. Bu bakımdan cümlenin öğrenilmesi için sonraki cümle tahmini görevi cümle benzerliği göreviyle değiştirilmiştir. Bu durumda, bu iki cümlenin bağlamsallaştırılmış yerleştirmelerinin kosinüs benzerliği belirlenen bir eşik değerinin üzerinde seyrederse, çıktının '1' olacağı, aksi takdirde '0' olacağı bir ikili sınıflandırma görevi yerine getirilmiştir. Eşik değeri *t*'nin 0,75 gibi yüksek bir değere ayarlanması, cümle benzerliği görevi için güvenilir tahminlerle sonuçlanacağını garanti etmektedir. Model, 20001 adım için 512 parti boyutu (batch-size) kullanılarak eğitilmiştir. Giriş metni 256 token uzunluğunda doldurulmaktadır. Derin öğrenme modellerini eğitmek için Olasılıksal Dereceli Azalma (Stochastic Gradient Descent - SGD) için bir yedek optimizasyon algoritması olan Adam kullanılmıştır. Adam optimize edicinin eğitim sırasında kullandığı hiper parametreler β_1=0,90, β_2=0,80 ve öğrenme oranı (learning rate) 0,0005'e ayarlanmıştır. Eğitim sırasında modelin otomatik olarak uyarlanabilmesi ve farklı bir ortalama ve varyansa sahip olacak şekilde kaydırılabilmesi için eğitilmiş ağırlık vektörleri uygulanmaktadır. Farklı örnekler arasındaki ortalama ve standart sapma bağımsız olarak hesapladığından, normalizasyonun parti boyutu boyunca ve dolayısıyla partideki diğer örneklere bağlı olduğu şekilde gerçekleştirilir. BERT için önceden eğitilmiş ağırlıklar, transformatör kitaplığında bulunmaktadır. Bu görev için BERT uygulamasında önceden eğitilmiş ağırlıklar kütüphane içerisine dahil edilmiştir ve bu önceden eğitilmiş ağırlıklar, modelin veri kümesi üzerinde ince ayar yapabilmesi adına kullanılmıştır. Bu LMTweets modeli bir öz-dikkat mekanizması aracılığıyla bir dizi bağlamsal yerleştirme oluşturarak dizideki her bir kelime için bağlamsal bilgiyi almakta ve eğitim verilerinin sözdizimsel, sözcüksel ve anlamsal özelliklerini belirgin bir şekilde yakalamaktadır. Böylece önceki eğitim yinelemesi sırasında edinilen bilgiler aracılığıyla yeni görevlerde en iyi performansı gösterebilir.



Şekil 3. Saldırganlık tespiti için kullanılan metodoloji ve veri kümesi üzerinde analizi

Figure 3. *The methodology used for aggression detection and analysis on dataset*

Önerilen yöntemin ikinci aşaması olarak LMTweets'lerinin CNN modeliyle birleştirilmesi yer almaktadır. Eğitim ve değerlendirme setindeki tüm cümle çiftleri için LMTweet'lerden bağlamsallaştırılmış yerleştirmeler çıkarılmaktadır. Bunlar, son transformatör bloğunun son gizli

durumlarıdır ve bu, CNN modelinin girdisi olacaktır. Geliştirilen model, CNN içindeki yerel mekansal yapıyı kullanarak, yakındaki kelimelerin zamansal ilişkisini tanımlayabilmektedir, yani birlikte görünen kelimeler, ister saldırganlık ister duygu ifadeleri olsun, belirli bir türün tespiti için önemli bilgiler içermektedir. Bu şekilde CNN, elle işletilerek veya geleneksel özelliklerden daha üstün olan özellikleri öğrenebilmekte ve bu da mevcut yaklaşımlara kıyasla genelleme kabiliyetini artırmaktadır. Kullanılan CNN modeli, üç katmanlı (her biri 64 gizli birimden oluşan) ve filter-size'ı üç olan 1-B CNN'dir. Her CNN katmanından sonra filter-size=2 olan Max-Pooling katmanı uygulanmıştır. Son CNN katmanı, 128 unit'li tam-bağlantılı bir katmandır ve ikili çapraz entropi (binary cross-entropy) ile eğitilmiştir. Bu çalışmada, 0,003 learning rate oranına sahip Adam optimizer kullanılarak 10 dönem (epochs) için 256 batch-size kullanılmıştır.

Algoritma-1: LMTweets-CNN

Girdi: Veri seti (Metin)
Çıktı: Saldırgan / Saldırgan olmayan
1. her bir text t_i girdisi için;
  Simgeleştirilmiş sözcük (S_{tokens}) ← t_i simgeleştirme
  Sözcük temsili (boyut 768) ← LMTweets (Simgeleştirilmiş sözcükler)
2. CNN ($Ö_1$) ← Her sözcük için sözcük temsili
3. Max-Pooling (MH_1) ← CNN ($Ö_1$)
4. CNN ($Ö_2$) ← Max-Pooling (MH_1)
5. Max-Pooling (MH_2) ← CNN ($Ö_2$)
6. CNN ($Ö_3$) ← Max-Pooling (MH_2)
7. Global Max-Pooling (KMH) ← CNN ($Ö_3$)
8. Dropout Katmanı (SK) ← Global Max-Pooling (KMH)
9. Dense Katmanı (YK) ← Dropout Katmanı (SK)
10. Olasılıklar (O_1 ve O_2) ← Dense Katmanı (YK)
11. eğer ($O_1 \geq 0{,}5$) **ise** Saldırgan,
  değilse Saldırgan olmayan

Önerilen model olan LMTweets + CNN'in çalışması Algoritma-1'de sözde kod olarak verilmiştir. Algoritma başlangıç sürecinde, tweet metinlerini girdi biçiminde almakta ve metnin saldırgan / saldırgan olmayan olup olmadığına bakılmaksızın çıktı üretmektedir. Algoritmada kullanılan sembollerden t_i: yorumları, S_{tokens}: sözcük belirteçlerini, $Ö_1$: CNN'in birinci katmanını, $Ö_2$: CNN'in ikinci katmanını, $Ö_3$: CNN'in üçüncü katmanını, MH_1: birinci Max-Pooling katmanını, MH_2: ikinci Max-Pooling katmanını, KMH: Global Max-Pooling katmanını, SK: Dropout Katmanını, O_1: Metnin saldırgan olma olasılığını, O_2: Metnin saldırgan olmama olasılığını ifade etmektedir.

DENEYSEL SONUÇLAR VE TARTIŞMA (EXPERIMENTAL RESULTS AND DISCUSSION)

Çeşitli makine öğrenmesi, derin öğrenme ve transformatör tabanlı modeller incelenmiş ve Şekil 3'te gösterilen kriterler doğrultusunda önerilen modelin deneysel sonuçları ile karşılaştırılmıştır. Performans değerlendirmesi için, doğruluk, kesinlik, duyarlılık, F_1-ölçütü ve eğri altındaki alan olarak ifade edilen AUC değeri olmak üzere beş ayrı kriter ile analiz edilmiştir.

Bu çalışmada, Cyber-Trolls veri seti kullanılarak aşağıdaki verilen temel modeller önerilen yaklaşım ile karşılaştırılmıştır.

(i) Makine öğrenmesi modelleri: NB, SVM ve K-En Yakın Komşu (K-Nearest Neighbors -KNN)

(ii) Derin öğrenme modelleri: CNN, LSTM ve Kapılı Tekrarlayan Birim (Gated Recurrent Unit - GRU)

(iii) Transformatör modeller: BERT, XLNet ve ULMFIT

Metin analizinde gerçekleştirilen ilk adım, verileri önişleme adımlarından geçirmek olmuştur. Sosyal medya sitelerinde bulunan veriler genellikle yapılandırılmamış ve çok fazla gürültü içermektedir. Gürültünün varlığı duygu analizi performansını olumsuz etkilemektedir (Jianqiang ve diğ., 2017). Bu sebeple öznitelikler çıkarılmadan önce çeşitli ön işleme teknikleri uygulanmaktadır. Makine öğrenmesi ve derin öğrenme algoritmaları için verileri önceden işlemek gerekmektedir. Uygulanan önişleme teknikleri şunlardır; (i) etkisiz kelime kaldırma, (ii) ekleri kaldırma, (iii) küçük harf dönüştürme, (iv) Ascii ve İngilizce haricini kaldırma, (v) URL Link kaldırma. Veri önişleme sonrasında veri seti, makine öğrenmesi modellerinde TF-IDF yaklaşımı kullanılarak sayısal özelliklere dönüştürülmüştür. Derin öğrenme modelleri için, metni sayısal özelliklere dönüştürmek için GloVe word embedding adımı kullanılmıştır. Transformatör modellerinde bir önişleme ve özellik gösterimi tekniklerinin uygulanmasına ihtiyaç yoktur.

Çizelge 3. Önerilen modelin Cyber-Troll veri kümesi ile mevcut modellerle performans karşılaştırması

Table 3. *Performance comparison of the proposed model with existing models with Cyber-Troll Dataset*

Model	Doğruluk	Kesinlik	Duyarlılık	F_1-ölçütü	AUC
Makine Öğrenmesi (*TF-IDF* (*n-gram = 3*))					
NB	0,68	0,42	0,45	0,42	0,74
SVM	0,67	0,51	0,43	0,44	0,80
KNN	0,53	0,34	0,30	0,32	0,69
Derin Öğrenme (*GloVe* (*Boyut = 300*))					
CNN	0,83	0,70	0,69	0,69	0,69
LSTM	0,85	0,70	0,73	0,71	0,73
GRU	0,88	0,80	0,78	0,78	0,71
Transformatör Modeller					
BERT	0,83	0,56	0,70	0,62	0,80
XLNet	0,85	0,53	0,74	0,62	0,80
ULMFIT	0,79	0,57	0,72	0,64	0,69
Mevcut Yaklaşımlar					
Sadiq ve diğ., 2021	0,86	0,93	0,78	0,85	-
Singh ve diğ., 2018	0,73	-	-	-	0,58
Gambäck ve Sikdar, 2017	-	0,86	0,70	0,77	-
Risch ve Krestel, 2018	-	-	-	-	0,63
Önerilen (LMTweets + CNN)	0,96	0,91	0,96	0,93	0,96

Tüm modellerin makine öğrenmesi, derin öğrenme, transformatör modellerinin, mevcut yaklaşımların ve önerilen yaklaşımın nicel analizine Çizelge 3'te yer verilmiştir. Klasik makine öğrenmesi algoritmalarıyla sınıflandırma işlemi için yani NB, SVM ve KNN uygulamak için Python programlama dili kullanılmıştır. Bu algoritmalar, saldırganlık tespitinde en sık uygulanan algoritmalar olduğundan tercih edilmiştir. Cyber-Trolls veri seti üzerinde yürütülen makine öğrenmesi algoritmalarının sonuçları Çizelge 3'te gösterilmiştir. En yüksek performans doğruluk değeri 0,68, kesinlik değeri 0,51, duyarlılık değeri 0,45, F_1-ölçütü değeri 0,44 ve AUC değeri 0,80 ile rapor edilmiştir. SVM algoritması NB ile karşılaştırılabilir şekilde performans göstermiştir. Doğruluk değeri 0,53, kesinlik değeri 0,34, duyarlılık değeri 0,30, F_1-ölçütü değeri 0,32 ve AUC değeri 0,69 olan KNN algoritmasında düşük performans gözlenmiştir.

Günümüzde, duygu analizi ve benzeri çeşitli doğal dil işleme problemlerinde derin öğrenme modelleri en yaygın kullanılan yaklaşımdır. Yakın zamana kadar, RNN modellerinden LSTM ve GRU en popüler yaklaşımlardı, ancak son yıllarda RNN modellerinden daha iyi performans gösterdiği ortaya çıkan bazı dikkat mekanizmalarına rastlanmıştır. Derin öğrenme algoritmalarını uygulamak için arka uç olarak tensör akışına sahip Keras API kullanılmıştır. Burada uygulanan derin öğrenme algoritmalarından CNN parametreleri için (filters=128, filter-size=5, activation='ReLU', dropout=0,5,

batch-size=256, epoch=40, optimizer='Adam') değerlerine ayarlanmıştır. LSTM parametreleri için (dropout= 0,3, activation='Relu', unit=128, epoch =40, optimizer='Adam') değerlerine ayarlanmıştır. GRU parametreleri için (dropout=0,3, activation='ReLU', unit=128, epoch=40, optimizer='Adam') değerlerine ayarlanmıştır. Cyber-Trolls veri seti üzerinde yürütülen derin öğrenme sonuçları sırasıyla Çizelge 3'te gösterilmiştir. Buna göre GRU, 0,88 doğruluk, 0,80 kesinlik, 0,78 duyarlılık ve F_1-ölçütü ile en iyi performansı göstermiştir. Bunun yanında LSTM ve CNN modelleri, karşılaştırmalı olarak neredeyse eşit performans göstermiştir.

Son yıllarda transformatörler, duygu analizi görevi için tüm modellerden daha iyi performans göstermiştir (Maslej ve diğ., 2020). Uygulama için bilinen transformatör modellerinden BERT, XLNet ve ULMFIT olmak üzere üç transformatör modeli kullanılmıştır (Liu ve diğ., 2019). Cyber-Trolls veri seti üzerinde yürütülen transformatör modellerinin sonuçları sırasıyla Çizelge 3'te gösterilmiştir. Buna göre en yüksek performans 0,85 doğruluk, 0,53 kesinlik, 0,74 duyarlılık, 0,62 F_1-ölçütü, 0,80 AUC değerleri ile XLNet ile elde edilmiştir. BERT modeli ile ULMFIT modeli yaklaşık olarak eşit performans göstermiştir. Transformatör modellerinde kullanılan hiper parametreler train-epochs=3, learning rate= 2e-6, max-seq =32, batch-size=32, Adam-epsilon=1e-9'dur. Veri seti için maksimum dizi uzunluğu 1024'e ayarlanmıştır. Önerilen LMTweets + CNN modelinin sonuçları Çizelge 3'te gösterilmiştir. Saldırganlık tespiti için önerilen bu model, literatürde bildirilen temel değerler ve yaklaşımlarla karşılaştırılmıştır. Literatürde bildirilen yaklaşımların sonuçları ilgili araştırma makalelerinden edinilmiştir. Sadiq ve diğ. tarafından önerilen yaklaşım mevcut yöntemler arasında en iyi performansa sahip olan modeldir. Buna göre 0,86 doğruluk, 0,93 kesinlik, 0,78 duyarlılık, 0,85 F_1-ölçütü ile en iyi sonuçları elde etmiştir. Önerilen model ile 0,96 doğruluk, 0,91 kesinlik, 0,96 duyarlılık, 0,93 F1-ölçütü ve 0,96 AUC değeri elde edildiği göz önüne alındığında, bu modelin konuyla ilgili yapılmış son teknolojik çalışmaların deneysel sonuçlarına kıyasla yaklaşık %8 oranında daha yüksek performansa ulaştığı söylenebilir.

Önerilen model, bağlamsal bilgileri çıkarmak için BERT-tabanlı mimariye dayalı bir LMTweets modelini tasarlamaya dayanmaktadır. GloVe ve Word2Vec gibi statik sözcük yerleştirmeleri, bağlamlarından bağımsız olarak sözcükleri temsil etmektedir. Ancak LMTweets ve BERT, bağlama dayalı olarak kelime yerleştirme üreterek kelime bankası için farklı kelime yerleştirmeleri üretecektir. LMTweets, bu modellerin eğitildiği alan ve eğitim prosedürü açısından BERT'den farklıdır. BERT, alana özgü ve genel belgeler arasındaki sözcüklerdeki, tümceciklerdeki ve diğer dilsel özelliklerdeki büyük farklılıklar nedeniyle bu veri kümelerinde düşük performans göstermiştir. Ayrıca, LMTweets tarafından oluşturulan yerleştirmelerde, yakındaki kelimelerin zamansal ilişkisini belirlemek için CNN modeli uygulanmıştır.

Sonuçlar, önerilen modelin performansında her iki bileşenin (LMTweets ve CNN) önemini göstermektedir. Saldırgan içerikli metin, dilbilimsel, örüntü temelli, noktalama temelli, edimbilim, sözdizimsel ve kutupsallık temelli özellikler gibi çeşitli özelliklerle temsil edilebilmektedir. Modelimiz, kalıp tabanlı özelliklere sahip olanların (# işareti içeren) saldırganlığını tespit edememiştir. NB'in tüm makine öğrenmesi algoritmaları arasında daha iyi performans gösterdiği gözlemlenebilir. GRU'nun tüm derin öğrenme algoritmaları arasında daha iyi performans gösterdiği gözlemlenebilir. XLNet, tüm transformatör modelleri arasında daha iyi performans göstermiştir. Sonuçlar, yaklaşımımızın literatürde bildirilen tüm makine öğrenimi, derin öğrenme, transformatör modelleri ve modellerinden daha iyi performans gösterdiğini göstermektedir.

Kombinasyon Çalışması (Combination Study)

Önerilen model üzerinde gerçekleştirilen kombinasyon çalışması ve sonuçları Çizelge 4'te sunulmuştur.

Çizelge 4. Önerilen modelin CNN kombinasyonu çalışma sonuçları
Table 4. *Combination study results of the proposed model*

Model	Doğruluk	Kesinlik	Duyarlılık	F_1-ölçütü	AUC
Kombinasyon Çalışmaları					
LMTweets + CNN	0,96	0,91	0,96	0,93	0,96
LMTweets olmadan	0,83	0,70	0,69	0,69	0,69
CNN olmadan	0,94	0,86	0,94	0,90	0,95

Derinlemesine bir analiz için, her bileşenin önemini belirlemek adına LMTweets + CNN modeli üzerinde bir kombinasyon çalışması yapılmıştır. Bu kombinasyon çalışmasının detayları aşağıdaki gibidir:

(i) LMTweets olmadan: LMTweets kısmı modelden kaldırılmıştır. Metni sayısal forma dönüştürmek için LMTweets yerine GloVe metin gösterimi tekniği kullanılmıştır.

(ii) CNN olmadan: CNN kısmı, LMTweets + CNN modelinden çıkarılmıştır ve sınıflandırma için LMTweets kısmı kullanılmıştır.

Kombinasyon çalışmasının sonuçlarına Çizelge 4'te yer verilmiştir. Modelin sadece CNN kısmı GloVe word embedding birlikte ele alındığında, performansın ortalama %13 oranında düştüğü görülmektedir. Bu durum LMTweets'in etkili performans üzerindeki önemini göstermektedir. Buna karşın CNN kısmı çıkarıldığı taktirde, modelin performansı ortalama %2 oranında azalmıştır.

Özetle bu çalışmada önerilen LMTweets + CNN yaklaşımı daha önce Çizelge 3'ün derin öğrenme modelleri arasında bulunan ve Çizelge 4'te "LMTweets olmadan" olarak adlandırılan CNN (GloVe ile birlikte) yöntemi ve LMTweets + CNN yaklaşımından CNN kısmının çıkarıldığı "CNN olmadan" olarak adlandırılan yöntemler ile karşılaştırılmıştır. Burada amaç LMTweets + CNN yaklaşımının farklı şekillerdeki kombinasyonlarına olan üstünlüğünü vurgulamaktır.

Anova Testi (Anova Test)

Tek yönlü varyans analizi (Anova) normal dağılımlı bir seride iki veya daha fazla bağımsız ortalama arasındaki farkın manidarlığının hesaplanmasıyla deneysel sonuçların anlamlı olup olmadığını bulan istatistiksel bir testtir. Anova tek başına iki veya daha fazla grubun aritmetik ortalamalarını kümülatif olarak karşılaştırmakta, bu karşılaştırmalardan en az birisi anlamlı olduğunda Anova sonucu da anlamlı bulunmaktadır. Olgumuzda sadece bir bağımsız değişken bulunduğundan tek yönlü Anova testi uygulanmıştır. Bu testte, boş hipotez "Grup ortalamasında fark yoktur" ve alternatif hipotez "Grup ortalamasında fark vardır" şeklindedir. Anova testi, gruplar arasındaki ve bir grup içindeki varyasyonları karşılaştırarak gerçekleştirilmektedir. Dördüncü bölümde sunulan sonuçların istatistiksel anlamlılığını değerlendirmek için Python istatistik modelleri kitaplığı kullanılarak tek yönlü Anova testi uygulanmıştır. Üçüncü bölümde önerilen yöntemin Anova testi ile elde edilen olasılık değeri (O) sonuçları, yöntemin etkinliğine istatistiksel olarak karar vermek için kullanılmıştır. Boş hipotezin doğru olduğunda reddedilme olasılığı olan anlamlılık düzeyi 0,05'tir. O değeri, bir Anova testi yapılarak elde edilmektedir. O değeri anlamlılık düzeyinden küçükse, boş (*null*) hipotez reddedilir. Anova testinin doğruluk için veri seti üzerindeki çıktısı Çizelge 5'te verilmiştir. Anova testi sonucunda istatistiksel olarak anlamlı bir olasılık değeri olan (O) değeri elde edilmektedir. Elde edilen O-değeri 1.586×10^{-7}'dir. Bu olasılık değeri gruplar arasında fark olduğunun ortaya konması bakımından $O<0,05$ şeklinde bir karşılaştırma yapılmaktadır. F değeri ile karşılaştırılmak üzere elde edilen olasılık sonucu 0.0000001586 değerini ifade etmektedir. Bu bakımdan anlamlılık seviyesi olan 0,05 değerinden çok daha küçük bir sonuçtur.

Çizelge 5. Anova testinin sonuçları (Tek-yönlü)
Table 5. *Results of Anova Test (One-way)*

Kaynak	Serbestlik Derecesi (df)	Kareler toplamı (sum sq)	Ortalamalar toplamı (mean sq)	F1 ölçütü	Olasılık<F
Doğruluğa Dayalı					
Sınıflandırma	3,00	0,24	0,08	16,33	1.586×10^{-7}
Artık (Resuidal)	50,00	0,24	0,005	-	-

Çizelge 5'teki satır ifadeleri Anova testinin Sınıflandırma (Classification) veya Artık (Resuidal) parametreleri tercih edilerek Serbestlik derecesi (degree of freedom), Kareler toplamı (sum of square), Ortalamalar toplamı (mean square), F1 ölçütü (F-measure) ve Doğruluk (Accuracy) metrikleri üzerinden hesaplandığını ifade etmektedir. Bu test bağımsız grupların ortalamaları arasında istatistiksel olarak anlamlı bir farkın var olup olmadığını ölçmek için kullanılmaktadır. Bu test ile genellikle 0,00 ve 1,00 arasında değişkenlik gösteren sonuçların elde edilmesindeki amaç sadece istatistiksel farkın ortaya konulmasıdır. Elde edilen deneysel sonuçlar ile ilgili bir çıkarım yapılamamaktadır.

Burada, Anova testinin deneysel sonuçları makine öğrenmesi, derin öğrenme, transformatör modellerinin ve önerilen modelin istatistiksel olarak farklı performans gösterebildiğini ortaya koymuştur.

SONUÇLAR (RESULTS)

Özellikle sosyal medya mikro blog platformlarında, dilin sürekli değişen doğası nedeniyle saldırganlık tespiti zordur. Bu çalışmada, LMTweets ve CNN'den oluşan bir mimari önerilmiştir. LMTweets, sosyal medya platformlarından gelen verilerle eğitilmiş BERT-tabanlı transformatöre dayalı bir kodlayıcıdır. Özellikler LMTweets tarafından çıkarılmaktadır ve sınıflandırma için CNN modeline iletilmektedir. Cyber-Trolls veri kümesine üç makine öğrenmesi, üç derin öğrenme ve üç transformatör tabanlı model uygulanmıştır. Tüm modeller arasında en iyi performansı LMTweets + CNN'in gösterdiği gözlemlenmiştir. Önerilen model ile 0,96 doğruluk, 0,91 kesinlik, 0,96 duyarlılık, 0,93 F_1-ölçütü ve 0,96 AUC sonuçları elde edilmiştir. Bu, duruma göre sonuçlar son teknoloji yöntemlere kıyasla %8 daha yüksek performans göstermiştir. Saldırganlık tespiti geniş ve ilgi çekici bir alan olduğu için gelecekteki çalışmalar takip edilebilir. Gelecekte, hinglish gibi makaronik dillerden çok modelli veri kümelerini kullanmayı düşünebiliriz. Uygulanan modellerin yanında grafik sinir ağı veya kapsül ağları da uygulanabilir. Bununla birlikte RNN ve CNN mekanizmalarını dikkate alan hibrit modeller de araştırılabilir.

KAYNAKLAR (REFERENCES)

Abulaish, M., Kamal, A., Zaki, M., 2020, "A survey of figurative language and its computational detection in online social networks." 14(1): 1-52.

Al-Garadi, M. A., Varathan, K. D., Ravana, S. D. J. C. i. H. B., 2016, "Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network." 63: 433-443.

Aroyehun, S. T., & Gelbukh, A., 2018. "Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling. In Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying" (TRAC-2018) (pp. 90-97).

Avvaru, A., Vobilisetty, S., & Mamidi, R., 2020, "Detecting sarcasm in conversation context using transformer-based models. In Proceedings of the second workshop on figurative language processing" (pp. 98-103).

Balakrishnan, V., Khan, S., Fernandez, T., Arabnia, H. R. J. P., 2019, "Cyberbullying detection on twitter using Big Five and Dark Triad features." 141: 252-257.

Bansal, A., Sharma, S. M., Kumar, K., Aggarwal, A., Goyal, S., Choudhary, K., 2012, "Classification of flames in computer mediated communications."

Basile, V., Bosco, C., Fersini, E., Debora, N., Patti, V., Pardo, F. M. R., 2019, "Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter." 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics.

Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., & Vakali, A., 2017, "Mean birds: Detecting aggression and bullying on twitter." Proceedings of the 2017 ACM on web science conference.

Chavan, V. S., & Shylaja, S., 2015, "Machine learning approach for detection of cyber-aggressive comments by peers on social media network." 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), IEEE.

Chia, Z. L., Ptaszynski, M., & Masui, F., 2019, "Exploring machine learning techniques for irony detection." Proceedings of the Annual Conference of JSAI 33rd Annual Conference, 2019, The Japanese Society for Artificial Intelligence.

Davidson, T., Warmsley, D., Macy, M., & Weber, I., 2017, "Automated hate speech detection and the problem of offensive language." Proceedings of the International AAAI Conference on Web and Social Media.

Dinakar, K., Jones, B., Havasi, C., Lieberman, H., & Picard, R. J. A. T., 2012, "Common sense reasoning for detection, prevention, and mitigation of cyberbullying." 2(3): 1-30.

Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., & Bhamidipati, N., 2015, "Hate speech detection with comment embeddings." Proceedings of the 24th international conference on world wide web.

Farías, D. I. H., Montes-y-Gómez, M., Escalante, H. J., Rosso, P., & Patti, V., 2018, "A knowledge-based weighted KNN for detecting Irony in Twitter." Mexican International Conference on Artificial Intelligence, Springer.

Gambäck, B., & Sikdar, U. K., 2017, "Using convolutional neural networks to classify hate-speech." Proceedings of the first workshop on abusive language online.

Greevy, E., & Smeaton, A. F., 2004, "Classifying racist texts using a support vector machine." Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval.

Gregory, H., Li, S., Mohammadi, P., Tarn, N., Draelos, R., & Rudin, C., 2020, "A Transformer approach to contextual Sarcasm detection in Twitter." Proceedings of the Second Workshop on Figurative Language Processing.

Grigg, D. W., 2010, "Cyber-aggression: Definition and concept of cyberbullying." Journal of Psychologists and Counsellors in Schools, 20(2), 143-156.

Hepburn, A. D., 1875, Manual of English Rhetoric, American Book Company.

Jianqiang, Z., & Xiaolin, G. J. I. A., 2017, "Comparison research on text pre-processing methods on twitter sentiment analysis." 5: 2870-2879.

Joachims, T., 1998, "Text categorization with support vector machines: Learning with many relevant features." European conference on machine learning, Springer.

John, T. N., 2000, "Hate Speech." In Encyclopedia of the American Constitution (2nd ed.,edited by Leonard, W. L., Kenneth, L. K. et al., New York: Macmillan), pp. 1277-1279.

Joshi, A., Bhattacharyya, P., & Carman, M. J. J. A. C. S., 2017, "Automatic sarcasm detection: A survey." 50(5): 1-22.

Khan, U., Khan, S., Rizwan, A., Atteia, G., Jamjoom, M. M., & Samee, N. A. 2022. "Aggression Detection in Social Media from Textual Data Using Deep Learning Models." Applied Sciences, 12(10), 5083.

Kumar, A., Sangwan, S. R., Arora, A., Nayyar, A., & Abdel-Basset, M. J. I. a., 2019, "Sarcasm detection using soft attention-based bidirectional long short-term memory model with convolution network." 7: 23319-23328.

Kumar, R., Ojha, A. K., Malmasi, S., & Zampieri, M., 2018, "Benchmarking aggression identification in social media." Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018).

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., 2019, "Roberta: A robustly optimized bert pretraining approach."

Madisetty, S., & Desarkar, M. S., 2018, "Aggression detection in social media using deep neural networks." Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018).

Maslej-Krešňáková, V., Sarnovský, M., Butka, P., & Machová, K. J. A. S., 2020, "Comparison of deep learning models and various text pre-processing techniques for the toxic comments classification." 10(23): 8631.

Mihaylov, T., Georgiev, G., & Nakov, P., 2015, "Finding opinion manipulation trolls in news community forums." Proceedings of the nineteenth conference on computational natural language learning.

Mubarak, H., Darwish, K., & Magdy, W., 2017, "Abusive language detection on Arabic social media." Proceedings of the first workshop on abusive language online.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y., 2016, "Abusive language detection in online user content." Proceedings of the 25th international conference on world wide web.

Pareek, K., Choudhary, A., Tripathi, A., Mishra, K. K., & Mittal, N. 2022. "Hate and Aggression Detection in Social Media Over Hindi English Language." International Journal of Software Science and Computational Intelligence (IJSSCI), 14(1), 1-20.

Potamias, R.-A., Siolas, G., & Stafylopatis, A., 2019, "A robust deep ensemble classifier for figurative language detection." International Conference on Engineering Applications of Neural Networks, Springer.

Potamias, R. A., Siolas, G., Stafylopatis, A.-G. J. N. C., 2020, "A transformer-based approach to irony and sarcasm detection." 32(23): 17309-17320.

Prentice, S., Taylor, P. J., Rayson, P., Hoskins, A., & O'Loughlin, B. J. I. S. F., 2011, "Analyzing the semantic content and persuasive composition of extremist media: A case study of texts produced during the Gaza conflict." 13(1): 61-73.

Risch, J., & Krestel, R., 2018, "Aggression identification using deep learning and data augmentation." Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018).

Sadiq, S., Mehmood, A., Ullah, S., Ahmad, M., Choi, G. S., & On, B.-W. J. F. G. C. S., 2021, "Aggression detection through deep neural model on twitter." 114: 120-129.

Salawu, S., He, Y., & Lumsden, J. J. I. T. o. A. C., 2017, "Approaches to automated detection of cyberbullying: A survey." 11(1): 3-24.

Saravanaraj, A., Sheeba, J., Devaneyan, S. P. J. I. J. o. C. S., 2016, "Automatic detection of cyberbullying from twitter."

Sarsam, S. M., Al-Samarraie, H., Alzahrani, A. I., & Wright, B. J. I. J. o. M. R., 2020, "Sarcasm detection using machine learning algorithms in Twitter: A systematic review." 62(5): 578-598.

Schmidt, A., & Wiegand, M., 2017, "A survey on hate speech detection using natural language processing." Proceedings of the fifth international workshop on natural language processing for social media.

Sharif, O., & Hoque, M. M. 2022. "Tackling cyber-aggression: Identification and fine-grained categorization of aggressive texts on social media using weighted ensemble of transformers." Neurocomputing, 490, 462-481.

Shen, Y., He, X., Gao, J., Deng, L., & Mesnil, G., 2014, "Learning semantic representations using convolutional neural networks for web search." Proceedings of the 23rd international conference on world wide web.

Singh, V., Varshney, A., Akhtar, S. S., Vijay, D., & Shrivastava, M., 2018, "Aggression detection on social media text using deep neural networks." Proceedings of the 2nd Workshop on Abusive Language Online (ALW2).

Smit, D. J. S. A. J. o. E., 2015, "Cyberbullying in South African and American schools: A legal comparative study." 35(2): 1-11.

Su, H.-P., Huang, Z.-J., Chang, H.-T., & Lin, C.-J., 2017, "Rephrasing profanity in chinese text." Proceedings of the First Workshop on Abusive Language Online.

Tai, K. S., Socher, R., & Manning, C. D. J. a. p. a., 2015, "Improved semantic representations from tree-structured long short-term memory networks."

Tulkens, S., Hilte, L., Lodewyckx, E., Verhoeven, B., & Daelemans, W. J. a. p. a., 2016, "A dictionary-based approach to racism detection in dutch social media."

Van der Walt, E., Eloff, J. H., Grobler, J. J. C., 2018, "Cyber-security: Identity deception detection on social media platforms." 78: 76-89.

Xiao, Y. and Cho, K. J. a. p. a., 2016, "Efficient character-level document classification by combining convolution and recurrent layers."