Research Article/Araştırma Makalesi

Gazi AKADEMİK YAYINCILIK

GAZİ
JOURNAL OF ENGINEERING SCIENCES

Determining Cyberbullying Status of Turkish Tweets with Gpt-3 Classification Model

Çilem Koçak[a], Tuncay YİĞİT[b]

Submitted: 21.11.23 Revised: 19.12.2023 Accepted: 20.12.2023 doi:10.30855/gmbd.0705S26

ABSTRACT

Regardless of whether they are young or old, people have quickly stepped into the internet world with phones, tablets, computers and smart devices, which are among today's communication technologies. With the increase in the number of social media users, some negativities are encountered. The most important problem encountered on social media is cyber bullying. Although cyberbullying may seem like daily dialogues between social media users or groups, the incidence of cyberbullying increases day by day with the diversification of shared information, content, and agenda in social media environments. This paper evaluates the performance of a GPT-3 classification model for the task of classifying tweets from tweeters as those that contain cyberbullying or those that do not. The model was first trained and tested on Turkish tweets. It resulted in an accuracy of 55% in total. After the data was translated to English and the model was retrained and tested, accuracy was increased to 66%. Precision, recall, and F1 score for both classes of tweets were 0.65, 0.68, and 0.67, respectively, for tweets without cyberbullying and 0.67, 0.64, and 0, respectively, for tweets with cyberbullying. It was found to be 65. Cyberbullying The confusion matrix of the model showed that 17 tweets correctly contained cyberbullying, while 9 tweets incorrectly contained cyberbullying. The results of this paper show that GPT-3 can be used for the task of classifying tweets into those that contain cyberbullying and those that do not with a reasonable degree of accuracy.

Keywords: Cybebrullying, Twitter, Artificial intelligence, Text Classification, GPT-3

[a,*] Isparta Unıversıty Of Applıed Scıences, Yalvaç Technical Sciences Vocational School - Isparta, Türkiye
Orcid: 0000-0002-4516-2076
e mail: cilemkocak@isparta.edu.tr

[b] Süleyman Demirel University, Faculty Of Engineering,, Dept. of Computer Engineering Isparta, Türkiye
Orcid: 0000-0001-7397-7224

*Corresponding author: cilemkocak@isparta.edu.tr

Gpt-3 Sınıflandırma Modeli İle Türkçe Tweetlerin Siber Zorbalık Durumlarının Belirlenmesi

ÖZ

İnsanlar genç yaşlı fark etmeksizin günümüz iletişim teknolojilerinden olan telefon, tablet, bilgisayar ve akıllı cihazlar ile internet dünyasına hızlı bir şekilde adım atmışlardır. Sosyal medya kullanıcı sayısının artışı ile de bazı olumsuzluklarla karşılaşılmaktadır. Sosyal medya da karşılaşılan en önemli sorun da siber zorbalık durumlardır. Siber zorbalık sosyal medya kullanıcıları ya da gruplar arasında gerçekleşen günlük diyaloglar gibi görünse de paylaşılan bilgi, içerik, gündem sosyal medya ortamlarının çeşitlenmesi ile günden güne karşılaşma durumu artmaktadır. Bu makale, tweetterdan alınan tweetlerin siber zorbalık içerenler ve içermeyenler olarak sınıflandırma görevi için bir GPT-3 sınıflandırma modelinin performansını değerlendirmektedir. Model ilk olarak Türkçe tweet'ler üzerinde eğitimiş ve test edilmiştir. Toplamda da %55'lik bir doğrulukla sonuçlanmıştır. Veriler İngilizce'ye çevrildikten ve model yeniden eğitilip test edildikten sonra doğruluk %66'ya yükseltilmiştir. Her iki tweet sınıfı için kesinlik, hatırlama ve F1 puanı, siber zorbalık içermeyen tweet'ler için sırasıyla 0,65, 0,68 ve 0,67 ve siber zorbalık içeren tweet'ler için sırasıyla 0,67, 0,64 ve 0,65 olarak bulunmuştur. Siber zorbalık Modelin karışıklık matrisi, 17 tweet'in doğru bir şekilde siber zorbalık içerdiğini, 9'unun ise yanlış bir şekilde siber zorbalık içerdiğini görülmüştür. Bu makalenin sonuçları, GPT-3'ün tweet'leri siber zorbalık içerenler ve makul bir doğruluk derecesi ile içermeyenler olarak sınıflandırma görevi için kullanılabileceğini göstermektedir.

Anahtar Kelimeler: *Siber zorbalık, Twitter, Yapay zeka, Metin Sınıflandırma, GPT-3*

To cite this article: Ç. Koçak and T. Yiğit, "Determining Cyberbullying Status of Turkish Tweets with Gpt-3 Classification Model," *Gazi Journal of Engineering Sciences,* vol.9, no.4, pp. 278-285, 2023. doi:10.30855/gmbd.0705S26

1. Giriş (Introduction)

Telefon ve bilgisayar teknolojilerinin gelişmesi kullanım durumunun ve sosyal medya platformlarındaki sayının artışı ile siber zorbalık davranışlarının görülme olasılığı da artmaktadır. Siber zorbalık mağdurları, kullanıcılar tarafından elektronik iletişim araçları ile tehdit edilmekte, sıklıkla yazılı hakaret içeren mesajlar almakta, sahte kimlik ile birini kötü gösterme gibi eylemlerle karşı karşıya kalmaktadırlar. Bu durumda zorba ve kurban arasındaki karşılıklı ilişkiler sorunları ortaya çıkmaktadır [1]. Bu sorunlar insan arasındaki arkadaşlık ve duygusal ilişkilerinin bozulması, farklı görüş ve düşüncelere sahip olan kişiler arasında olan yazılı atışmalar nedenleriyle oluşan intikam duygusundan ortaya çıktığı düşünülmektedir [2]. Siber zorbalık davranışının sergilenmesi için gerçekleştirildiği araç ve ortam fark etmeksizin mağdur üzerinde yıkıcı bir sonuç ortaya çıkarma, kırma, küçük düşürme, aşağılama ve kurbanda kalıcı izler bırakılması istenmekle birlikte mağdurun toplum ilişkileri olumsuz yönde etkilenmekte ayrıca duygusal, sosyal ve psikolojik olarak zarar vermektedir.

Bu makale, metin sınıflandırması kullanarak siber zorbalık tespiti için makine öğrenimi algoritmalarının kullanımına ilişkin bir çalışma sunmaktadır. Siber zorbalık günümüz toplumunda önemli bir sorundur ve onu tespit etmek için etkili ve verimli yöntemlere ihtiyaç duyulmaktadır. Bu nedenle, bu makale siber zorbalık tespiti için makine öğrenimi algoritmalarının potansiyelini keşfetmeyi amaçlamaktadır. Spesifik olarak, çevrimiçi metinlerden siber zorbalığı tespit etmek için denetimli bir makine öğrenimi algoritmasına dayalı bir metin sınıflandırma yaklaşımı önerilmektedir. Bu yaklaşımın performansını mevcut yöntemlerle karşılaştırılmış ve bir siber zorbalık belgeleri külliyatındaki etkinliğini değerlendirilmektedir. Sonuçlar, önerilen yöntemin doğruluk açısından mevcut yaklaşımlardan daha iyi performans gösterdiğini göstermektedir. Ayrıca bulgularımızın çıkarımlarını tartışıyor ve bu konuda gelecekte yapılabilecek olası çalışmaları önerilerde bulunmaktayız.

2. Siber Zorbalık (Cyber Bullying)

Siber zorbalık, e-posta, metin mesajları, sosyal medya veya diğer çevrimiçi platformlar gibi dijital dünyada gerçekleşen bir zorbalık biçimidir [3]. Gençler arasında giderek artan bir endişe kaynağıdır ve duygusal sıkıntı, depresyon ve hatta intihar gibi ciddi sonuçları olabilir. Siber zorbalık, birini taciz etmek, tehdit etmek veya utandırmak için dijital teknolojiyi kullanmayı içerir. Kırıcı mesajlar veya resimler göndermeyi, acımasız yorumlar göndermek için sahte hesaplar oluşturmayı veya birinin kişisel bilgilerini rızası olmadan paylaşmayı içerebilir. Ayrıca, kurbanın kötü görünmesini sağlamak için çevrimiçi içeriğin manipüle edilmesini veya kurbanın istenmeyen resimlerini çevrimiçi olarak yayınlamayı da içerebilir. Siber zorbalığın etkileri yıkıcı olabilir. Siber zorbalık kurbanları utanç, izolasyon ve depresyon duyguları yaşarlar. Ayrıca kaygı, yorgunluk ve düşük benlik saygısı yaşayabilirler. Ek olarak, siber zorbalık kurbanları baş ağrısı, uyku güçlüğü ve mide ağrısı gibi fiziksel semptomlar yaşayabilir [4]. Son yıllarda teknolojinin yaygınlaşması ve gençlerin çevrimiçi geçirdikleri sürenin artması nedeniyle siber zorbalığın yaygınlığı arttı. Bu nedenle, siber zorbalık belirtilerinin farkında olmak ve bunu önlemek için adımlar atmak önemlidir. Siber zorbalık örnekleri arasında birine kaba mesajlar veya resimler göndermek, kırıcı yorumlar göndermek için sahte sosyal medya hesapları oluşturmak ve birinin kişisel bilgilerini rızası olmadan kullanmak yer alır. Ayrıca, birinin utanç verici resimlerini yayınlamayı veya kurbanı kötü göstermek için çevrimiçi içeriği manipüle etmeyi içerebilir. Siber zorbalık birçok şekilde olabilir ve belirtilerin farkında olmak ve bunu önlemek için adımlar atmak önemlidir [5].

Siber zorbalık ile fiziksel ortamda gerçekleştirilen zorbalık türleri her ne kadar zorbalığın gerçekleştirildiği ortam farklı olsa da birbirlerine benzemektedirler. Siber zorbalık davranışlarının sergilenmesi için ortam farkı olarak internet ortamı, sosyal medya gereçleri kullanılmaktadır. Zorbalar birçok yöntemle siber zorbalık yapmaktadırlar. Sık karşılaşılan ve sınıflandırılan siber zorbalık çeşitleri;

- ✓ "Siber takip; Bir kişiyi sanal ortamlarda sürekli takip halinde tutmak,
- ✓ Karalamak; Bir kişi ile ilgili asılsız, zararlı ve kaba beyanlarda bulunmak,
- ✓ Kendini Başkası Gibi Göstermek; İnternet ortamlarında kendi kimliğini gizleyerek hayali biri ya da başkasının kimliğine bürünmek,
- ✓ Taciz Etme; Bir kişiye kırıcı veya cinsel içerikli mesajlar göndermek,
- ✓ Kışkırtmak; Bir kişiyi yapmaması gereken durumlar için teşvik etmek,
- ✓ Gezinti ve Düzenbazlık; Bir kişi hakkında utandırıcı ve özel bilgileri yaymak yayınlamak,
- ✓ Ayırma; Bir kişiyi bir gruptan çıkarmak veya gruba dâhil etmemek gibi türleri bulunmaktadır".[1]

3. Materyal ve Yöntem (Material and Method)

Siber Zorbalığı Tespit Algoritmaları:
Naive Bayes: Naive Bayes algoritmaları, metinde kullanılan kelime ve kelime öbeklerini analiz ederek metnin siber zorbalık olup olmadığını sınıflandırmak için kullanılır[6].

Metin Sınıflandırması: Bu algoritma genellikle metni siber zorbalık veya değil olarak sınıflandırmak için kullanılır. Bu, sözdizimi, duyarlılık ve dil gibi çeşitli özellikler kullanılarak yapılır[7].

Kümeleme: Kümeleme algoritmaları, verilerdeki siber zorbalığa işaret eden kalıpları belirlemek için kullanılabilir. Örneğin, siber zorbalığı tanımlamak için benzer yorumları gruplamak için kümeleme kullanılabilir[8].

Anormallik Tespiti: Anormallik tespit algoritmaları, sıra dışı olan ve siber zorbalığa işaret edebilecek davranış veya yorumları belirlemek için kullanılabilir[8].

Sinir Ağları: Sinir ağları, daha sonra siber zorbalığı tespit etmek için kullanılabilecek verilerdeki kalıpları öğrenmek için kullanılabilir[8].

BERT: Transformers'tan Çift Yönlü Kodlayıcı Temsilleri, doğal dil işleme (NLP) ön eğitim tekniğine dayalı bir derin öğrenme algoritmasıdır. BERT, 2018 yılında Google tarafından geliştirildi ve o zamandan beri dünyanın en güçlü NLP modellerinden biri haline geldi. BERT'nin amacı, en alakalı anlamı belirlemek için bağlamı kullanarak bilgisayarların metindeki belirsiz dilin anlamını anlamalarına yardımcı olmaktır. Örneğin, BERT, sözcükler bağlam dışında kullanılsa bile bir cümlenin amacını belirleyebilir. BERT, bilgisayarların kullanıcı sorgularını yorumlamasına ve ilgili yanıtları üretmesine yardımcı olan doğal dil anlama (NLU) modelleri oluşturmak için de kullanılır. BERT, çok çeşitli dillerde soru yanıtlama, duygu analizi ve diğer doğal dil görevleri için kullanılmıştır [9].

GPT-3 (Generative Pre-trained Transformer 3): OpenAI tarafından geliştirilmiş gelişmiş bir Doğal Dil İşleme (NLP) modelidir. 175 milyar parametre kapasitesiyle şimdiye kadar oluşturulmuş en büyük ve en güçlü sinir ağıdır. GPT-3, milyonlarca web sayfasından oluşan devasa bir veri kümesi üzerinde eğitilmiştir ve metin oluşturma yetenekleriyle insan benzeri metinler üretebilir. Soru yanıtlama, çeviri, özetleme ve duygu analizi gibi çeşitli NLP görevlerinde dikkate değer bir performans göstermiştir. Gücü, sınırlı eğitim verilerinden genelleme yapma yeteneğinde ve çok çeşitli görevlere uygulanmasına izin veren ölçeklenebilirliğinde yatmaktadır [10].

3.1. BERT (Bidirectional Encoder Representations from Transformers)

Metin sınıflandırması için BERT kullanmanın ana avantajı, metinden bağlamsal bilgileri yakalama yeteneğidir. Bu, yalnızca metindeki tek tek sözcükleri dikkate alan sözcük torbası veya TF-IDF gibi diğer geleneksel yöntemlere göre çok büyük bir avantajdır. Ayrıca BERT, metindeki uzun vadeli bağımlılıkları yakalayabildiği için LSTM'ler veya CNN'ler gibi diğer derin öğrenme modellerinden önemli ölçüde daha güçlüdür. Ek olarak, BERT çeşitli metin sınıflandırma görevlerinde kullanılabilir. İncelemelerdeki duyarlılığı sınıflandırmak, haber makalelerindeki konuları tespit etmek veya müşteri sorgularının amacını sınıflandırmak için kullanılabilir. Belge özetleme, soru yanıtlama ve diğer birçok görev için de kullanılabilir. Genel olarak BERT, metin sınıflandırma görevleri için son derece güçlü ve çok yönlü bir araçtır. Metindeki uzun vadeli bağımlılıkları yakalamasına izin veren kelimelerin bağlamsal temsillerini öğrenme yeteneğine sahiptir. Ayrıca, duygu analizinden belge özetlemeye kadar çeşitli görevlerde kullanılabilir. İşte BERT önceden eğitilmiş ağının artıları ve eksileri.
Artıları:

- BERT, bağlamı diğer ağlardan daha iyi kodlayabilir ve dilin nüanslarını anlamasına olanak tanır.
- BERT, çeşitli doğal dil işleme görevlerinin üstesinden gelmek için kullanılabilen bir derin öğrenme mimarisidir.
- BERT, manuel özellik mühendisliği ihtiyacını azaltarak kendi kendine öğrenme yeteneğine sahiptir.
- BERT, belirli görevler için ince ayar yapılarak daha doğru sonuçlar üretebilir. Eksileri:
- BERT, hesaplama açısından pahalıdır ve eğitilmesi için çok sayıda GPU kaynağı gerektirir.
- BERT, uzun cümleleri doğru bir şekilde temsil etme yeteneğini sınırlayan uzun metin dizilerini

kodlamakta güçlük çekiyor.
- Bir kara kutu modeli olduğu için BERT'in sonuçlarını yorumlamak zor olabilir.
- BERT, metnin anlamını doğası gereği anlamıyor ve belirli görevler için ek manüel özellik mühendisliği gerektiriyor [9].

3.2. GPT-3 (Generative Pre-Trained Transformer 3)

GPT-3, metnin sınıflandırılma biçiminde devrim yaratan, çığır açan bir doğal dil işleme (NLP) teknolojisidir. GPT-3, metindeki kalıpları belirlemek ve içeriğine göre kategorilere ayırmak için derin öğrenmeyi kullanır. Bu, GPT-3'ün makaleler, blog gönderileri, sosyal medya gönderileri ve diğer yazılı materyaller gibi metinleri hızlı ve doğru bir şekilde sınıflandırabileceği anlamına gelir. GPT-3 güçlü bir araçtır çünkü daha doğru sınıflandırmalar yapmak için metnin konusu ve yazarın üslubu gibi bağlamsal bilgileri kullanabilir. Örneğin, GPT-3, sporla ilgili bir blog gönderisini, konular birbiriyle ilişkili olsa bile, siyasetle ilgili bir blog gönderisinden farklı olarak tanıyabilir. GPT-3, sağlık, eğitim, finans ve daha fazlası gibi çeşitli alanlarda metinleri tanımlamak ve sınıflandırmak için kullanılmaktadır. Ayrıca sohbet botları ve otomatik müşteri hizmetleri sistemleri oluşturmak için kullanılmaktadır [10].

GPT-3, çok sayıda eğitim verisinden öğrenme yeteneği nedeniyle güçlü bir metin sınıflandırma aracıdır. Bu verileri kalıpları algılamak ve metni geleneksel yöntemlerden daha doğru bir şekilde sınıflandırmak için kullanabilir. GPT-3, manüel olarak etiketlenmiş verilere ihtiyaç duymadan metni de sınıflandırabilir, bu da onu geleneksel yöntemlerden çok daha verimli hale getirmiştir. Bu, GPT-3'ün metni daha hızlı ve daha az hatayla sınıflandırabileceği anlamına gelmektedir. GPT-3, makaleler, blog gönderileri, sosyal medya gönderileri, müşteri hizmetleri konuşmaları ve daha fazlasını içeren çok çeşitli metinleri sınıflandırmak için yaygın bir şekilde kullanılmaktadır. Ayrıca, müşteri sorgularını anlayabilen ve otomatik yanıtlar sağlayabilen yapay zeka destekli sohbet botları ve müşteri hizmetleri sistemleri oluşturmak için de kullanılmaktadır [8]

GPT-3, insan müdahalesine ihtiyaç duymadan metni doğru bir şekilde sınıflandırma yeteneği nedeniyle kısa sürede metin sınıflandırma için en popüler araçlardan biri haline gelmiştir. GPT-3, daha doğru sınıflandırmalar yapmak için metnin konusu ve yazarın üslubu gibi bağlamsal bilgileri kullanır. GPT-3, makaleler, blog gönderileri, sosyal medya gönderileri, müşteri hizmetleri konuşmaları ve daha fazlasını içeren çok çeşitli metinleri sınıflandırmak için kullanılmaktadır. Ayrıca, müşteri sorgularını anlayabilen ve otomatik yanıtlar sağlayabilen yapay zeka destekli sohbet botları ve müşteri hizmetleri sistemleri oluşturmak için de sıklıkla kullanılmaktadır. GPT-3, metnin sınıflandırılma biçiminde devrim yaratmakta ve işletmeler, araştırmacılar ve daha fazlası için paha biçilmez bir araç haline gelmektedir *[11]*.
Artıları:
- GPT-3, doğru metin sınıflandırma sonuçları üretebilen gelişmiş bir doğal dil işleme (NLP) modelidir.
- GPT-3, karmaşık kalıpları tanımasını ve daha doğru sonuçlar üretmesini sağlayan geniş bir eğitim veri setine (175 milyar parametre) sahiptir.
- GPT-3, mevcut uygulamalara ve hizmetlere kolayca entegre edilebilir, bu da onu metin sınıflandırma görevleri için uygun maliyetli bir çözüm haline getirir.
- GPT-3, kelime dağarcığı dışındaki kelimeleri işlemek için yerleşik bir özelliğe sahiptir ve bu da onu yüksek doğruluk gerektiren metin sınıflandırma görevleri için güvenilir bir araç haline getirir.

Eksileri:
- GPT-3, çalışması için büyük bilgi işlem kaynakları gerektiren pahalı bir çözümdür.
- GPT-3, doğru sonuçlar elde etmek için modeli eğitmek ve ince ayar yapmak için çok zaman gerektirir.
- GPT-3, kısa metin kalıplarını tanımlayamadığı için kısa metin sınıflandırma görevleri için uygun değildir.
- GPT-3'ün İngilizce dışındaki dilleri tanıma yeteneği sınırlıdır.

3.3. BERT ve GPT-3'ün karşılaştırılması (Comparison of BERT and GPT-3)

Metin sınıflandırması için BERT ve GPT-3'ü karşılaştırırken dikkate alınması gereken birkaç faktör vardır. Birincisi, BERT, dildeki daha ince nüansları yakalamasına izin veren çift yönlü eğitim kullandığından GPT-3'ten daha sağlam bir modeldir. İkincisi, BERT, GPT-3'ten daha geniş bir metin külliyatı üzerinde önceden eğitilmiştir ve ona dil hakkında daha fazla bilgi verir. Üçüncüsü, BERT, metin sınıflandırma görevlerinde

GPT-3'ten çok daha hızlıdır ve bu da onu gerçek zamanlı uygulamalar için daha uygun hale getirmektedir. Son olarak BERT, GPT-3'e kıyasla uzun vadeli bağımlılıkları yakalamada daha iyi olduğu görülmektedir. Sonuç olarak, hem BERT hem de GPT-3, metin sınıflandırması için güçlü modellerdir. Ancak BERT, sağlamlığı, daha büyük önceden eğitilmiş veri topluluğu, daha hızlı performansı ve uzun vadeli bağımlılıkları yakalama yeteneği nedeniyle görev için daha uygunluğu tartışmasızdır.

3.4. Eğitim verileri (Training data)

Bu veri seti, üç yüksek lisans öğrencisi tarafından siber zorbalık içeren (1) veya içermeyen (0) olarak etiketlenen 7574 tweet'ten oluşmaktadır. Veriler Twitter'dan elde edilmiş ve bir tweet'in siber zorbalık içerip içermediğini doğru bir şekilde sınıflandırmak için makine öğrenimi algoritmalarının eğitiminde kullanılması amaçlanmaktadır. Etiketler, öğrenciler tarafından tweet'in içeriğini analiz edilerek belirlenmiştir. Ancak, ön işleme işlemlerinden sonra sadece 150 örnek seçilmiştir. 0. sınıf için 75, 1. sınıf için 75. 100 örnek eğitime, kalanı ise teste ayrılmıştır.

4. Sonuçlar (Results)

İlk adımda GPT-3 modelinde, Türkçe tweet'ler içeren orijinal verilerle eğitim yapılmıştır. Eğitimden sonra, orijinal test verileriyle test edilmiştir.

Table 1. Türkçe Tweetler için Gpt-3 performans sonuçları (Gpt-3 performance results for Turkish Tweets)

	Precision	**Recall**	**F1-score**	**Support**
0	0.60	0.35	0.44	25
1	0.53	0.76	0.62	25
Accuracy			0.55	
Macro avg	0.56	0.55	0.53	50
Weighted avg	0.56	0.55	0.53	50

Yukarıdaki performans sonuçları, modelin genel doğruluğu olan %55'lik bir doğruluk göstermektedir. Bu, modelin test edildiği 51 örnekten 55'inin sonucunu doğru bir şekilde tahmin edebildiği anlamına gelmektedir. Tablonun ilk satırı, 0,60 kesinlik, 0,35 geri çağırma ve 0,44 F1 puanı ile sınıf 0'a karşılık gelmektedir. Bu, modelin sınıf 0 örneklerinin %60'ını doğru bir şekilde tanımlayabildiği, ancak yalnızca %35'ini hatırladığı görülmüştür. F1 puanı kesinlik ve hatırlamanın harmonik ortalamasıdır ve bu durumda 0,44'tür. Tablonun ikinci satırı, 0,53 kesinlik, 0,76 hatırlama ve 0,62 F1 skoru ile sınıf 1'e karşılık gelmektedir. Bu, modelin 1. sınıf örneklerin %53'ünü doğru bir şekilde tanımlayabildiği ve %76'sını hatırlayabildiği görülmüştür F1 puanı 0,62'dir. Makro ortalama, her iki sınıf puanının ortalamasıdır ve bu durumda 0,53'tür. Ağırlıklı ortalama da 0,53'tür, bu da modelin her iki sınıfta da benzer performans gösterdiğini gösterir. Genel olarak, model, daha yüksek hatırlama ve F1 puanı ile gösterildiği gibi, sınıf 1'de sınıf 0'dan biraz daha iyi performans göstermektedir. Bununla birlikte, nispeten düşük puanların gösterdiği gibi, model hala her iki sınıfta da optimum performans göstermemektedir.

Nispeten başarısız olan sonuçların ardından Train ve test veri setlerindeki tüm tweetler İngilizceye çevrilerek tekrar train and test işlemleri gerçekleştirildi. İkinci modelin performans sonuçları Tablo 2'de görülebilir.

Table 2. İngilizce Tweetler için Gpt-3 performans sonuçları (Gpt-3 performance results for English Tweets)

	Precision	**Recall**	**F1-score**	**Support**
0	0.65	0.68	0.67	25
1	0.53	0.76	0.62	25
Accuracy			0.66	
Macro avg	0.66	0.66	0.66	50
Weighted avg	0.66	0.66	0.66	50

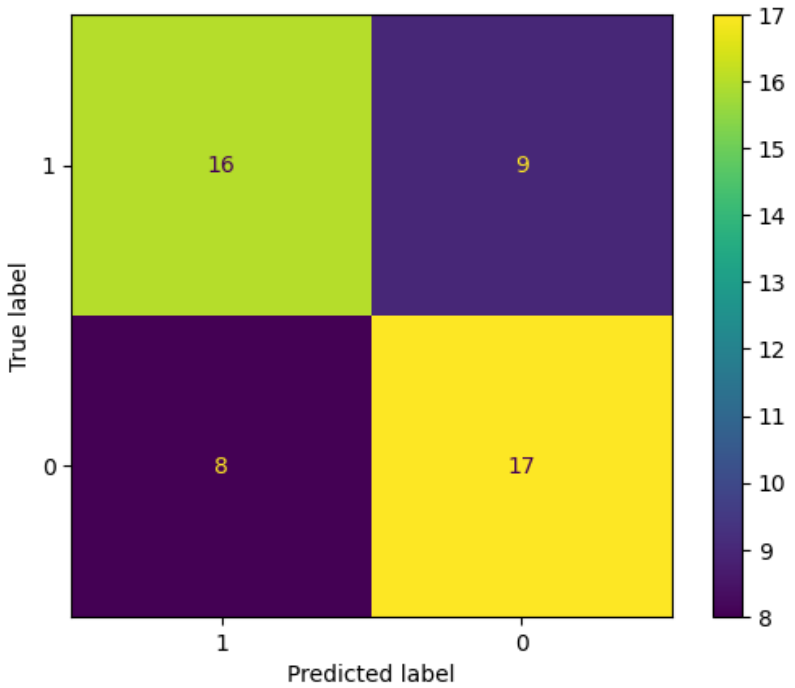
Bu GPT-3 sınıflandırma modeli, siber zorbalık içermeyen tweet'leri sınıflandırmak için 0,65 ve siber zorbalık içeren tweet'leri sınıflandırmak için 0,67 kesinliğe sahiptir. Bu, modelin bir tweet'in siber zorbalık içerip içermediğini doğru bir şekilde sınıflandırabildiğini göstermekte ve siber zorbalık içermeyen tweet'ler için ise geri çağırma puanı 0,68'dir. Buda, modelin siber zorbalık içermeyen tweet'lerin çoğunu doğru bir şekilde tanımlayabildiğini gösterirken, siber zorbalık içeren tweet'ler için hatırlama puanı 0,64'tür ve bu da modelin Siber zorbalık içeren tweetlerin çoğunu hatırlayabildiği anlamına gelmekteidr. Her iki sınıf için f1 puanı sırasıyla 0.67 ve 0.65'tir. f1 puanı, modelin hem kesinlik hem de hatırlama puanlarını hesaba katan bir

doğruluk ölçüsüdür. Modelin, her iki tweet sınıfını da oldukça yüksek bir doğruluk derecesi ile doğru bir şekilde sınıflandırabildiğini göstermektedir. Genel olarak, GPT-3 sınıflandırma modelinin doğruluğu 0,66'dır ve bu, tweet'leri siber zorbalık içeren ve siber zorbalık içermeyen tweet'leri makul bir doğruluk derecesi ile sınıflandırabildiğini göstermektedir.

Table 3. İngilizce Tweetler için Bert performans sonuçları (Bert performance results for English Tweets)

	Precision	Recall	F1-score	Support
0	0.33	0.04	0.07	25
1	0.49	0.92	0.64	25
Accuracy			0.48	
Macro avg	0.41	0.48	0.36	50
Weighted avg	0.41	0.48	0.36	50

Bu performans raporu ise bir sınıflandırma algoritması içindir. Kesinlik, geri çağırma ve f1 puanı her sınıf (0 ve 1) için değerlendirilmektedir. Destek değeri, her sınıfa ait örnek sayısını gösterir. 0 sınıfı için kesinlik 0,33, geri çağırma 0,04 ve f1 skoru 0,07'dir. Bu, modelin, sınıf 0'a ait olduğu tahmin edilen örneklerin yalnızca %33'ünü doğru bir şekilde sınıflandırabildiği ve sınıf 0'ın gerçek üyelerinin yalnızca %4'ünü tanımlayabildiği anlamına gelmektedir. 0.07'lik f1 puanı oldukça düşüktür ve modelin bu sınıfta iyi performans göstermediğini göstermektedir. 1. sınıf için kesinlik 0,49, geri çağırma 0,92 ve f1 puanı 0,64'tür. Bu, modelin 1. sınıfa ait olduğu tahmin edilen örneklerin %49'unu doğru bir şekilde sınıflandırabildiği ve 1. sınıfa ait gerçek üyelerin %92'sini tanımlayabildiği anlamına gelmektedir. 0.64'lük f1 puanı oldukça yüksektir ve model bu sınıfta iyi performans gösterdiği anlamına gelmektedir. Genel olarak, modelin doğruluğu 0,48'dir ve bu oldukça düşüktür. Makro ortalamalı ise f1 puanı 0,36'dır ve bu puanda oldukça düşüktür. Bu da genel olarak İngilizce'den Türkçe'ye çevrilen Tweet'lerin kullanıldığı siber zorbalık sınıflandırmasında bile modelin iyi performans göstermediğini göstermektedir.



Şekil 1. İngilizce Tweetler için Gpt-3 Hata Matriksi Sonuçları (Gpt-3 Confusion Matrix Results for English Tweets)

Tweetlerin siber zorbalık içerip içermediğini sınıflandırmaya çalışan bir GPT-3 sınıflandırıcısının karışıklık matrisi, tweetlerin 17'sinin siber zorbalık içeriyor olarak doğru bir şekilde sınıflandırıldığını, 8'inin ise yanlış bir şekilde siber zorbalık içermediğini göstermektedir. Öte yandan, 9 tweet yanlışlıkla siber zorbalık içeriyor olarak sınıflandırılırken, 16 tweet doğru bir şekilde siber zorbalık içermiyor olarak sınıflandırılmıştır.

5. Tartışma ve Sonuç (Results and Discussion)

Hem GPT-3 English hem de BERT English, tweet'leri siber zorbalık içeren ve içermeyen olarak sınıflandırma yetenekleri açısından değerlendirilmiştir. GPT-3, 0,66'lık bir doğruluk elde etmiş ve bu, her iki tweet sınıfını

da oldukça yüksek bir doğruluk derecesi ile doğru bir şekilde sınıflandırabildiğini göstermiştir. Öte yandan BERT, genel olarak İngilizce'den Türkçe'ye çevirilen tweet'leri kullanarak siber zorbalık sınıflandırmasında bile iyi performans göstermediğini belirten 0,48'lik bir doğruluk elde etmiştir. BERT için makro ortalama f1 puanı 0,36 olarak görüşmüş ve bu puan oldukça düşük olarak görülmektedir. Bu modellerin sonuçları, tweet'lerdeki siber zorbalığı sınıflandırmak için hem GPT-3'ün hem de BERT'nin doğruluğunu artırmak için daha fazla çalışma yapılması gerektiğini göstermektedir. Performansı GPT-3'ünkinden önemli ölçüde düşük olduğundan, özellikle daha fazla araştırma BERT'nin doğruluğunu artırmaya odaklanmalıdır. Ayrıca, diğer dillerden İngilizce'ye çevrilen tweet'leri sınıflandırmak için her iki modelin doğruluğunu artırmaya yönelik yöntemler üzerinde araştırma yapılmalıdır.

GPT-3 Türkçe, 0.60 kesinlik, 0.35 hatırlama ve 0.44 F1 puanı ile sınıf 0 ile %55 doğrulukla gerçekleştirilmiş olup, sınıf 1, 0.53 kesinliğe, 0.76 hatırlamaya ve 0.62 F1 puanına sahiptir. Her iki sınıfın makro ortalaması 0,53, ağırlıklı ortalaması ise 0,53'tür. GPT-3 English, siber zorbalık içermeyen tweet'leri sınıflandırmak için 0,65 ve siber zorbalık içeren tweet'ler için 0,67 kesinliğe sahip olduğu görülmektedir. Siber zorbalık içermeyen tweetlerin geri çağırma puanı 0,68, siber zorbalık içeren tweetlerin geri çağırma puanı ise 0,64 olarak sonuçlanmıştır. Her iki sınıf için f1 puanı sırasıyla 0.67 ve 0.65 ve modelin genel doğruluğu 0.66'dır. BERT İngilizce twettlerde, sınıf 0 için 0,33 kesinliğe, 0,04 hatırlamaya ve 0,07 f1 puanına sahiptir. 1. sınıf için kesinlik 0.49, geri çağırma 0.92 ve f1 skoru 0.64'dür. Modelin genel doğruluğu 0,48 ve makro ortalamalı f1 puanı 0,36 olarak görülmektedir. GPT-3 İngilizce karışıklık matrisi, tweet'lerin 17'sinin doğru bir şekilde siber zorbalık içeriyor olarak sınıflandırıldığını, 8'inin ise yanlışlıkla siber zorbalık içermiyor olarak göstermiştir. Genel olarak, modellerin hiçbiri optimum performans göstermemiş, GPT-3 ve BERT İngilizce modelleri en kötü performansı göstermiştir.

Tweetlerdeki siber zorbalığı sınıflandırmak için BERT'nin doğruluğunu artırmak gerekmektedir ve bunun için daha fazla araştırma yapılmalıdır. Bu, modelin ön eğitimi için yeni mimarileri ve farklı teknikleri keşfetmeye öncelik verilebilir. Diğer dillerden İngilizce'ye çevrilmiş tweet'leri sınıflandırırken hem GPT-3'ün hem de BERT'nin doğruluğunu artırmaya yönelik yöntemler hakkında daha fazla araştırma yapılmalıdır. Bu, dile özel ön eğitim için yöntemleri keşfetmeyi ve farklı diller için ince ayar yapmak gerekmektedir. İspanyolca, Fransızca ve Almanca gibi diğer dillerde ve Instagram ve Facebook gibi diğer sosyal medya türlerinde siber zorbalığı tespit etmeye yönelik yöntemler geliştirmek için araştırma yapılmalı, bilgiyi aktarmak için transfer öğrenme tekniklerini kullanmanın etkililiği araştırılmalıdır.

Çıkar Çatışması Beyanı (Conflict of Interest Statement)

Yazarlar tarafından herhangi bir çıkar çatışması bildirilmemiştir.

Kaynaklar (References)

[1] D. M. Gezgin, and C. Çuhadar, "Bilgisayar ve öğretim teknolojileri eğitimi bölümü öğrencilerinin siber zorbalığa ilişkin duyarlılık düzeylerinin incelenmesi." *Eğitim Bilimleri Araştırmaları Dergisi*, *2*(2), 93-104. December 2012 Doi:10.24106/kefdergi.702927

[2] M. Özdemir, and F. Akar, "Lise öğrencilerinin siber-zorbalığa ilişkin görüşlerinin bazı değişkenler bakımından incelenmesi." *Educational Administration: Theory and Practice* 2011, Vol. 17, Issue 4, pp: 605-626 Kuram ve Uygulamada Eğitim Yönetimi 2011, Cilt 17, Sayı 4, ss: 605-626. May 2011, Doi:10.14527/kuey.2014.005

[3] H. T. Tanrıkulu, Kınay, and O. T. Arıcak, "Siber Zorbalığa İlişkin Duyarlılık Ölçeği: Geçerlik ve Güvenirlik Çalışması", *Trakya Üniversitesi Eğitim Fakültesi Dergisi*, vol. 3, no. 1, 2013. March 2013, Doi:10.24315/trkefd.305449

[4] T. Aricak, S. Siyahhan, A. Uzunhasanoglu, S. Saribeyoglu, S. Ciplak, N. Yilmaz, and C. Memmedov, "Cyberbullying among Turkish adolescents." *Cyberpsychology & behavior*, *11*(3), 253-261. Jun 2008 Doi:10.1089/cpb.2007.0016

[5] T. Ayas, and M. B. Horzum, "Öğretmenlerin sanal zorbalık algılarının çeşitli değişkenlere göre incelenmesi." *International Online Journal of Educational Sciences*, *3*(2), 619-640. Doi:10.31805/acjes.433205

[6] S. Akhter, (2018, December). "Social media bullying detection using machine learning on Bangla text." In *2018 10th International Conference on Electrical and Computer Engineering (ICECE)* (pp. 385-388). IEEE. 10 February 2018, Doi:10.1109/ICECE.2018.8636797

[7] M. G. Hussain, T. Al Mahmud, and W. Akthar, "An approach to detect abusive bangla text." In *2018 International Conference on Innovation in Engineering and Technology (ICIET)* (pp. 1-5). IEEE. March 2019, Doi:10.1109/CIET.2018.8660863

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, and I. Polosukhin, "Attention is all you need."*31st Conferenceon Neural Information Processing Systems(NIPS2017),LongBeach,CA,USA.*

[9]J. J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding", *arxiv.org, Oct. 2018.* [Online]. Avaliable: https://arxiv.org/abs/1810.04805. [Accessed: Nov. 12, 2023].

[10] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, and D. Amodei, "Language models are few-shot learners." *Advances in Neural Information Processing Systems 33 (NeurIPS 2020) 33*, 1877-1901, arXiv:2005.14165

[11]*Models - OpenAI API.* (n.d.). [Online]. Available: https://beta.openai.com/docs/models/gpt-3 [Accessed: June 10, 2022,].

** This paper was presented at the 5th International Conference on Artificial Intelligence and Applied Mathematics in Engineering (ICAIAME 2023) and the abstract was published as an e-book.*