BERT-Base Modeli ile Türkçe Sosyal Medya Paylaşımlarında Nefret Söylemi Tespiti

Detection of Hate Speech in Turkish Social Media Posts with BERT-Base Model

Şengül BAYRAK, Alper KARACA, Ferhat TOSON, Aleyna KOCABEY, Fatma Begüm ARSLANOĞLU
Yazılım Mühendisliği Bölümü, Bilgisayar Mühendisliği Bölümü,
İstanbul Sabahattin Zaim Üniversitesi, Türkiye
bayraksengul@ieee.org, karaca.alper@std.izu.edu.tr, toson.alper@std.izu.edu.tr, kocabey.aleyna@std.izu.edu.tr, arslanoglu.begum@izu.edu.tr

2023 31st Signal Processing and Communications Applications Conference (SIU) | 979-8-3503-4355-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/SIU59756.2023.10224040

***Özetçe*— İnternetin gelişmesiyle birlikte dünya genelinde kullanıcılar tarafından ifade edilen veri miktarı artmıştır. Sosyal medyada nefret söylemi tespiti, zararlı dilin önlenmesi ve güvenli çevrimiçi toplulukların teşvik edebilmesi önemli bir görevdir. Dönüştürücülerden Çift Yönlü Kodlayıcı Temsilleri (Bidirectional Encoder Representations from Transformers – BERT), doğal dil işleme görevlerinde başarılı performans sergileyen popüler bir dil modelidir. Bu çalışmada, Türkiye'deki Twitter kullanıcılarının yorumlarında nefret söylemi tespiti için BERT-Base modeli kullanılmıştır. Zararlı ve zararsız olarak etiketlenen örneklerden oluşan bir veri setiyle eğitilerek %92,53 test doğruluk başarımı elde edilmiştir. Geliştirilen model, canlı ortamda kullanıcıların hizmetine sunulmuştur.**

Anahtar Kelimeler — nefret söylemi tespiti, siber zorbalık, BERT-Base modeli.

***Abstract*— With the development of the Internet, the amount of data expressed by users worldwide has increased. Detecting hate speech in social media, preventing harmful language and promoting safe online communities is an important task. Bidirectional Encoder Representations from Transformers (BERT) is a popular language model that performs well in natural language processing tasks. In this study, the BERT-Base model is used to detect hate speech in the comments of Twitter users in Turkey. It was trained with a dataset of examples labeled as harmful and harmless and achieved 92.53% test accuracy. The developed model was presented to users in a live environment.**

Keywords — hate speech detection, cyberbullying, BERT-Base model.

I. GİRİŞ

Farklı dünya görüşlerinin ve bireylerin mevcudiyeti, duygu analizini zorlaştırmaktadır. Twitter gibi sosyal medya platformları geniş kullanıcı kitlesine sahiptir ve insanlar arasında çeşitli konular hakkında tartışma ve etkileşim olanağı sunmaktadır.

Ancak, bazı kullanıcılar bu platformları nefret söylemi, yani kişisel özellikleri, etnik kökenleri, din, cinsiyet, cinsel yönelim, fikirleri veya diğer özellikleri nedeniyle ayrımcılık yaratmak için kullanabilmektedir. Nefret söylemi, insanları yaralayabilecek ve ön yargılı düşünceleri yayabilecek ciddi bir sorundur. Bu nedenle, sosyal medya kullanıcılarının aşağılayıcı söylemlerinin tespit edilmesi gerekmektedir. Literatürde, kullanıcılar arasındaki nefret söylemi içeren mesajları tespit etmede farklı yapay zekâ ve makine öğrenmesi yöntemleri kullanılmıştır. Bununla birlikte, Tablo 1'de ifade edildiği gibi son yıllarda Twitter'da nefret söylemini tespit etmek için BERT gibi makine öğrenimi modellerinin kullanılması giderek daha popüler hale gelmiştir.

TABLO I. BERT_BASE MODEL İLE İLGİLİ LİTERATÜR ÖZETİ

Yazarlar	*Veri seti*	*Modeller*	*Doğruluk*
Xu vd. (2022) [1]	11639 kamyona ait hasar şiddeti tahminlemesi	BERT tabanlı sinir ağı	%69,2
Prottasha vd. (2022) [2]	Bangladeş dilinde yapılan olumlu olumsuz sosyal medya ifadeleri	BERT ve Transfer Öğrenme	%94,10
Lin vd.(2022) [3]	Sosyal medya kullanıcılarının yorumlarını yararlı/zararlı olarak sınıflandırma	BERT	%99,00
Emon vd.(2022) [4]	Facebook gönderilerinden toplanan 44,001 Bangla yorumuyla bir Bangla metin veri setinin sınıflandırılması	XML-RoBERT	%86,00

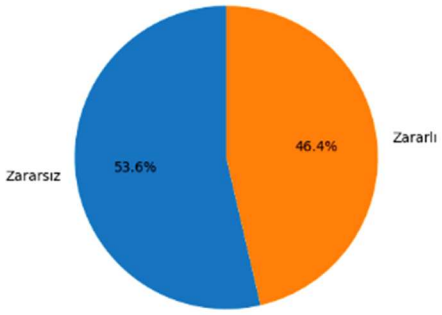
Karaahmetoğlu vd.(2021) [5]	Sosyal ağ tabanlı verilerden faydalanarak korona virüs konulu duygu analizi	BERT	%99,00
Souza vd. (2019) [6]	Finansal haberlerin sınıflandırılması	BERT	%72,50
Zhao vd.(2021) [7]	Çevrimiçi finansal metinsel verilerin sınıflandırılması	BERT	%95,42
Acikalin vd.(2020) [8]	Film ve otel yorumlarından oluşan, pozitif ve negatif olmak üzere 2 etikete sahip Türkçe veri kümelerinin sınıflandırılması	BERT	%88,00
Karimi vd. (2021) [9]	Müşteri memnuniyeti ile ilgili unsur çıkarma(Aspect Extraction – AE)	BERT-PT	%76,50
Guven vd.(2021) [10]	Türkçe tweetlerde duygu analizi	DistilBERT-Turkish	%98,63
Özkan vd.(2022) [11]	Türkçe dilinde yazılan bilimsel metinlerin çoklu sınıflandırılması	BERT	%95,00

Bu çalışmada, Türkçe nefret söylemi içeren tweetleri tespit etmek için dilin doğasını anlama yeteneği yüksek makine öğrenimi olan BERT-Base model kullanılmıştır. Çalışmada ilk olarak metni belirteçlerine ayırılmıştır. Belirteçler BERT-Base'in sözcük dağarcığı kullanılarak özel belirteç olarak eklenmiştir. İkinci aşamada, zararlı/zararsız olarak etiketli verileri kullanarak önceden eğitilmiş bir BERT modeline ince ayar yapılmıştır. İnce ayar, BERT-Base'in önceden eğitilmiş modelinde yararlanırken modelin nefret söylemi tespiti özel görevi üzerinde eğitilmesini içermektedir. Eğitim sırasında model, önceden eğitilmiş BERT katmanlarının ağırlıklarını güncellemiş ve öğrenmeye uygun sınıflandırma katmanlarını öğrenmiştir. Elde edilen model başarımıyla birlikte, test metin girdilerini zararlı/zararsız olarak %92,53 başarımla tespit edebilmiştir.

II. MALZEME

A. Veri Seti

Nefret söylemi tespiti için BERT-Base kullanmanın dengesiz verilerle başa çıkmak, sözcük dağarcığı dışındaki sözcükleri ele almak ve nefret söyleminin doğasında var olan öznelliği ele almak gibi çeşitli zorlukları vardır. Mevcut dengesiz veri setlerindeki modelleme maliyetlerindeki problemlerine çözüm olarak 2022 yılında Tanyel vd. [12] özgün veri kümesi oluşturmuşlardır. Türkçe tweetlerden oluşmuş veri seti hakaret ve saldırgan dil tespiti yapılabilmesi için toplam 53,005 sınıflandırılmış tweet içermektedir. Veri setindeki zararlı/zararsız tweet dağılımları Şekil 1'deki gibidir.



Şekil. 1. Veri Setindeki Zararlı ve Zararsız Tweet Dağılımları

B. Veri Ön İşleme Adımları

Bu çalışmada uygulanan veri ön işleme adımları Şekil 2'de verilmiştir.

Şekil. 2. Veri Setine Uygulanan Ön İşleme Adımları

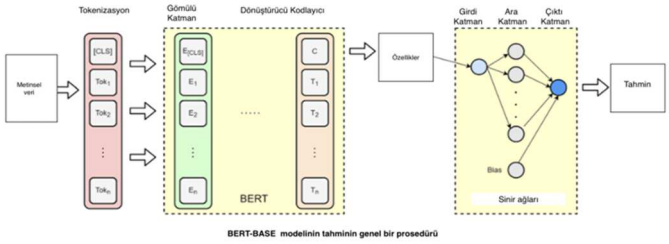
III. YÖNTEM

BERT modeli, çok sayıda yapay sinir ağı katmanının bir araya gelmesiyle oluşturulan dil anlama modelidir [13]. Bu yapay sinir ağı katmanları, dönüştürücü adı verilen bir mimariyi kullanarak önceden eğitilmiş bir model olarak çalışmaktadır. Bu nedenle, çeşitli doğal dil işleme görevleri için özelleştirilebilir. BERT modeli, ön eğitim ve ince ayar olmak üzere iki aşamalı çalışmaktadır. Ön eğitim aşamasında, büyük bir metin veri kümesindeki cümleleri, maskeleme ve segmentasyon adı verilen tekniklerle işlemektedir. Maskeleme işlemi, cümledeki bazı sözcüklerin rastgele olarak maskelenmesini ve modelin bu sözcükleri tahmin etmesini sağlamaktadır. Segmentasyon işlemi, iki farklı cümleyi birbirinden ayırmak için kullanılmaktadır ve modelin cümleler arasındaki ilişkiyi anlamasını sağlamaktadır. Bu işlemler, modelin dil anlama yeteneğini geliştirmektedir ve model, daha sonra ince ayar ile belirli bir doğal dil işleme görevi için özelleştirilmektedir. Bu aşamada, BERT modelinin çıktı katmanı, görevin gereksinimlerine göre değiştirilerek nefret söylemi, duygu analizi, cümle sınıflandırma gibi görevler için özelleştirilmektedir.

IV. DENEYSEL SONUÇLAR

A. BERT-Base Model Mimarisi

Bu çalışma için uygulanan BERT-Base mimarisi Şekil 3'te gösterilmiştir. Mimariye uygulanan eğitim, test ve doğrulama veri setleri Tablo 2'de verilmiştir. Veri setinin %60'ı eğitim, %20'si test, ve %20'si doğrulama veri seti olarak ayrılmıştır.



Şekil. 3. BERT-Base model mimarisi

TABLO II. UYGULANAN EĞİTİM, TEST, DOĞRULAMA VERİ SETİ

Sınıf	*Eğitim*	*Test*	*Doğrulama*
Zararlı (1)	13,895	4,646	4,557
Zararsız (0)	17,908	5,955	6,044

Şekil 3'e göre, BERT-Base mimarisi giriş katmanı, Tablo 2'de verilen veri setlerine göre zararlı/zararsız cümlelerin girdi olarak verildiği katmandır. Kodlayıcı katmanları, cümlelerin anlamını kodlamaktadır. Çıktı katmanı, verilen cümlelerin sonuçlarını zararlı/zararsız tespit edildiği katmandır. Maskeleme işlemi, belirli sözcüklerin gizlenmesi ve modelin bu sözcükleri tahmin etmesini sağlamaktadır.

B. Model Eğitimi

Maksimum dizi uzunluğu için kullanılan hiperparametre aralığı 256, 512 ve 1024 olarak belirlenmiştir. Öğrenme katsayısı içinse 1e-1, 1e-2, 1e-3 ve 1e-6 olarak belirlenmiştir. İyileştirme aşamasında uygulanan optimum parametre değerleri Tablo 3'te verilmiştir.

TABLO III. İYİLEŞTİRME İÇİN UYGULANAN OPTİMUM PARAMETRİK DEĞERLER

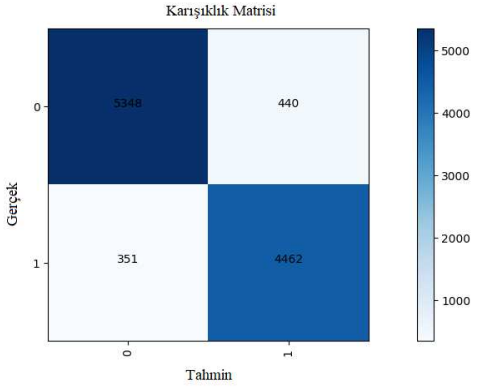
Parametre Adı	*Parametrik Değer*
Maksimum dizi uzunluğu	512
Eğitim grubu boyutu	2
Optimize edici	Adam
Model türü	BERT-Base
Öğrenme katsayısı	1e-6

C. Model Başarım Sonuçları

Optimum değerlerle elde edilen modelin eğitim veri seti için doğruluk oranı %89,19, doğrulama veri seti için doğruluk oranı %91,05 elde edilmiştir. Modelin test başarımı Tablo 4'te verilmiştir. Test veri setinde yer alan 5955 Zararsız(0), 4646 Zararlı (1) verinin model başarımı Şekil 4'teki gibi %92,53 olarak elde edilmiştir.

TABLO IV. MODEL TEST BAŞARIM SONUÇLARI

Sınıf	*Hassasiyet*	*Geri Çağırma*	*F1-skor*
Zararlı (1)	%93,00	%91,00	%92,00
Zararsız (0)	%92,00	%94,00	%93,00



Şekil. 4. Test veri seti için karışıklık matrisi

D. Modelin Canlıya Alınması

Bu çalışmada amaçlanan sosyal medya kullanıcılarının tweet içerikleri veri ön işleme adımları ile işlenerek BERT-Base modeli ile zararlı ve zararsız olarak tespit edilebilmektedir. Bireylerin kullandıkları sosyal medya dilinin tespiti için geliştirilen modele canlı ortamda Gradio Servisi ile ulaşılabilmektedir [14]. Gelecekte canlıya alınan bu sistemin Şekil 5'teki adımları karşılaması hedeflenmektedir.

Şekil. 5. Çalışmanın canlıya alınma adımları

V. SONUÇ VE GELECEK HEDEFLERİ

Twitter'da Türkçe yorumların nefret içerikli söylem tespitinde hazır veri seti kullanılarak zararlı/zararsız yorum tespiti BERT-Base ile modellenmiştir. Model test başarımı %92,53 olarak elde edilmiştir. Model canlı ortamda kullanıcıların hizmetine açılmıştır. Modelin iyileştirilmesinde farklı yöntemler kullanılarak test başarımının arttırılması hedeflenmektedir. Gelecekte daha kapsamlı özgün bir veri seti oluşturularak yeni modellemeler geliştirilecektir.

KAYNAKLAR

[1] Xu, S., Zhang, C., and Hong, D., "BERT-based NLP Techniques for Classification and Severity Modeling in Basic Warranty Data Study", *Insurance: Mathematics and Economics*, 107: 57-67, 2022.

[2] Prottasha, N. J., Sami, A. A., Kowsher, M., Murad, S. A., Bairagi, A. K., Masud, M., and Baz, M., "Transfer learning for sentiment analysis using BERT based supervised fine-tuning", *Sensors*, *22*(11): 4157, 2022.

[3] Lin, S. Y., Kung, Y. C., and Leu, F. Y., "Predictive Intelligence in Harmful News Identification by BERT-Based Ensemble Learning Model with Text Sentiment Analysis", *Information Processing & Management*, *59*(2):102872, 2022.

[4] Emon, M. I. H., Iqbal, K. N., Mehedi, M. H. K., Mahbub, M. J. A., and Rasel, A. A., "Detection of Bangla Hate Comments and Cyberbullying in Social Media Using NLP and Transformer Models", In *Advances in Computing and Data Sciences: 6th International Conference, ICACDS 2022, Kurnool, India, April 22–23, 2022, Revised Selected Papers, Part I* (pp. 86-96). Cham: Springer International Publishing.

[5] Karahmetoğlu, E., Ersöz, S., and Karahmetoğlu, O., "Sosyal Ağ Tabanlı Verilerden Faydalanarak Korona Virüs Konulu Duygu Analizi Çalışması", *Ergonomi*, *4*(1):47-54, 2021.

[6] Sousa, M. G., Sakiyama, K., de Souza Rodrigues, L., Moraes, P. H., Fernandes, E. R., and Matsubara, E. T., "BERT for Stock Market Sentiment Analysis", In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)* (pp. 1597-1601), IEEE.

[7] Zhao, L., Li, L., Zheng, X., and Zhang, J., "A BERT based Sentiment Analysis and Key Entity Detection Approach for Online Financial Texts", In *2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD)* (pp. 1233-1238), IEEE.

[8] Acikalin, U. U., Bardak, B., and Kutlu, M., "Turkish Sentiment Analysis Using BERT", In *2020 28th Signal Processing and Communications Applications Conference (SIU)* (pp. 1-4), IEEE.

[9] Karimi, A., Rossi, L., and Prati, A., "Adversarial Training for Aspect-Based Sentiment Analysis with BERT", In *2020 25th International Conference on Pattern Recognition (ICPR)* (pp. 8797-8803), IEEE.

[10] Guven, Z. A., "Türkçe Tweetlerde Duygu Analizi için BERT Modelleri ve Makine Öğrenme Yöntemlerinin Karşılaştırılması".

[11] Ozkan, M., and Görkem, K. A. R. "Türkçe Dilinde Yazılan Bilimsel Metinlerin Derin Öğrenme Tekniği Uygulayarak Çoklu Sınıflandırılması", *Mühendislik Bilimleri ve Tasarım Dergisi*, *10*(2): 504-519, 2022.

[12]T. Tanyel, B. Alkurdi and S. Ayvaz, "Linguistic-based Data Augmentation Approach for Offensive Language Detection," 2022 7th International Conference on Computer Science and Engineering (UBMK), 2022, 1-6, doi:0.1109/UBMK55850.2022.9919562.

[13] Gao, Z., Feng, A., Song, X., and Wu, X., "Target-dependent Sentiment Classification with BERT ", IEEE Access, 7: 154290-154299, 2019.

[14] https://huggingface.co/spaces/thealper2/turkish-hate-speech-streamlit