Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı 48, S. 19-22, Şubat 2023

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue 48, pp. 19-22, February 2023

Research Article

Mültecilere Yönelik Nefret Söyleminin Tespitinde Makine Öğrenmesi Modellerinin Kullanılması

Figen Eğin[1*], Vahide Bulut[2]

[1*] Katip Çelebi Üniversitesi,Mühendislik Fakültesi,Yazılım Mühendisliği, İzmir, Türkiye, (ORCID: 0000-0003-4865-5789), figenkayamail.com
2 Katip Çelebi Üniversitesi,Mühendislik Fakültesi,Yazılım Mühendisliği, İzmir, Türkiye, Türkiye (ORCID: 0000-0002-0786-8860),vahide.bulut@ikcu.edu.tr

(1st International Conference on Frontiers in Academic Research ICFAR, February 18-21, 2023)

(**DOI:** 10.31590/ejosat.1253132)

ATIF/REFERENCE: Eğin, F. & Bulut, V. (2023). Mültecilere Yönelik Nefret Söyleminin Tespitinde Makine Öğrenmesi Modellerinin Kullanılması. *Avrupa Bilim ve Teknoloji Dergisi*, (48), 19-22.

Öz

Sosyal medya kullanımının yaygınlaşması ile birlikte sosyal ağlar üzerinden çeşitli gruplara yönelik nefret söylemi gibi olumsuz paylaşımların kontrolsüzce yayılabildiği görülmektedir. Suriye İç Savaşı'nı takiben Türkiye'ye yaşanan göç, mültecilere yönelik nefret söylemini gündeme getirmiştir. Nefret söylemi, toplumsal huzurun sağlanabilmesi için önüne geçilmesi gereken önemli bir hastalık olarak betimlenmektedir. Nefret söyleminin tespiti konusunda Türkçe dilinde yapılan çalışmaların ve nefret söyleminin tespitinde kullanılabilecek kapsamlı bir veri setinin eksikliği göz önüne alınarak bu araştırmada sosyal ağlarda Türkçe dilinde yapılan paylaşımlarda mültecilere yönelik nefret söyleminin makine öğrenmesi yöntemleri ile tespiti üzerine çalışılmıştır. Lojistik regresyon (LR), Yapay Sinir Ağı (YSA), Destek Vektör Makineleri (DVM), Karar Ağaçları ve Rastgele Orman modelleri uygulanarak deneysel sonuçlar karşılaştırmalı olarak sunulmuştur. Rastgele Orman, YSA ve LR ile elde edilen performans değerlerinin DVM ve Karar Ağaçları modellerinden daha yüksek olduğu ortaya konmuştur.

Anahtar Kelimeler: Nefret söylemi, mülteciler, makine öğrenmesi

Using Machine Learning Models to Detect Hate Speech Against Refugees

Abstract

With the widespread use of social media, it is seen that negative posts such as hate speech towards various groups can spread uncontrollably through social networks. The migration to Turkey following the Syrian Civil War has brought hate speech towards refugees to the agenda. Hate speech is described as an important disease that must be prevented in order to ensure social peace. Considering the lack of studies conducted in Turkish on the detection of hate speech and the lack of a comprehensive data set that can be used in the detection of hate speech, this research has been studied on the detection of hate speech against refugees using machine learning methods in Turkish language posts on social networks. Experimental results are presented comparatively by applying logistic regression (LR), Artificial Neural Network (ANN), Support Vector Machines (SVM), Decision Trees and Random Forest models. It has been revealed that the performance values obtained with Random Forest, ANN and LR are higher than the SVM and Decision Tree models.

Keywords: Hate speech, refugees, machine learning

[*] Sorumlu Yazar: figenkaya@gmail.com

1. Giriş

Sosyal medya, insanların kendilerini özgürce ifade edebilecekleri bir platform olarak betimlenmektedir. İnsanların hızlıca iletişim kurabildikleri ve daha büyük kitlelere seslerini duyurabildikleri bu platform büyük miktarlarda verinin çok hızlı bir şekilde paylaşılabilmesi nedeniyle toplumu belirli durumlara karşı kışkırtan söylemlere de ev sahipliği yapabilmektedir. Bu söylemlerin başında önüne geçilmesi gereken mühim bir sorun olarak betimlenen nefret söylemi gelmektedir [1]. Nefret söylemi, bir grup veya bireyi ırk, din veya cinsiyet gibi doğuştan gelen özelliklere dayalı olarak hedef alan ve toplumsal barışı tehdit edebilecek saldırgan söylemleri kapsamaktadır [2]. Suriye İç Savaşı nedeniyle 2011 yılından bu yana yoğun bir mülteci göçü ile karşı karşıya olan Türkiye'de, nefret söyleminin hedeflerinden biri de göçmenler olmuştur. Nefret söyleminin yaygınlaşmasının kolaylıkla nefret suçlarına yol açabileceği göz önüne alındığında toplumsal huzur ve barışın sağlanması için nefret söyleminin önüne geçilmesinin önemi ortaya çıkmaktadır [3].

Sosyal medya platformlarında nefret söyleminin önüne geçilmesinin, paylaşım miktarı ve hızının çok yüksek olması nedeniyle, insan gözetimiyle gerçekleştirilmesinin mümkün olmadığı görülmektedir. Bu noktada, otomatik bir şekilde tespit yapabilecek bir sisteminin gerekliliği ortaya çıkmaktadır [4]. Bu alanda yapılan çalışmaların son yıllarda artış gösterdiği ve çalışmalarda çok farklı yöntemlerin kullanıldığı görülmektedir. Nefret söyleminin metin madenciliği yöntemleriyle tespiti için kullanılan yöntemler incelendiğinde; makine öğrenmesi yöntemlerinin TF-IDF (%12), BERT (%12), CNN (%15), RNN (%17), leksikon tabanlı modeller (%15) ve hibrit algoritmalar (%29) olduğu; ayrıca makine öğrenmesi modelleri olarak destek vektör makineleri (DVM), Naive Bayes (NB), Lojistik Regresyon (LR), Karar Ağaçları ve K-En Yakın Komşu (KNN) algoritmalarının kullanıldığı ortaya konmuştur [5].

Nefret söyleminin tespitinde, özellikle metin tabanlı paylaşımların ağırlık kazandığı Twitter sosyal ağ hizmetinin gerekli verinin elde edilmesi için sıklıkla kullanıldığı görülmektedir. Twitter'dan çekilen verilerle yürütülen bir çalışmada Yunanca tweet'ler ile oluşturulan ve 1040 tanesi "toksik" ve 2964 tanesi "toksit değil" şeklinde etiketlenmiş 4004 tweet içeren bir veri seti kullanılmıştır. Yapılan bu çalışmada, mülteci ve sığınmacılara yönelik yabancı düşmanlığı, ırkçılık ve nefret söylemlerinin doğal dil işleme yöntemleri ile tespiti üzerine yoğunlaşılmıştır. BERT (Bidirectional Encoder Representations from Transformers) ve Resnet (Residual Neural Networks) algoritması uygulanmıştır. Model 0,97 doğruluk puanı ve 0,947 f1 puanı yakalamıştır. Yine Twitter'daki nefret ifadelerini tespit etmek için yürütülen başka bir çalışmada, eğitim setinden otomatik olarak toplanan unigramlara ve kalıplara dayanan bir yaklaşım önerilmiştir. Bu çalışmada bir tweet'in saldırgan olup olmadığı %87; bir tweet'in nefret dolu, saldırgan veya temiz olup olmadığını tespit etmede %78.4 doğruluk değerine ulaşılmıştır [6]. Djuric ve ark. [7] ise , nefret söyleminin tespitinde iki aşamalı bir model önermişlerdir. Öncelikle CBOW (Continous Bag of Words) NLP modelini kullandıkları çalışmalarında, yorumların ve kelimelerin ortak modellemesi için Paragraf2vec kullanmışlardır. En yüksek AUC değerine Paragraph2Vec ve LR algoritmasını birlikte kullandıkları model ile ulaşmışlardır.

Literatür incelendiğinde, Türkçe diline ilişkin veri seti ve yapılan çalışmaların ise yeterli olmadığı görülmektedir. Bu noktadan hareketle bu araştırmanın amacı, mültecilere yönelik Türkçe nefret söyleminin metin madenciliği yöntemleriyle tespitinin sağlanması olarak belirlenmiştir. Bu kapsamda öncelikle mültecilere yönelik nefret söylemi içeren Türkçe bir veri setinin oluşturulması, ardından bu veri seti kullanılarak çeşitli makine öğrenmesi modellerinin başarımlarının ortaya konması hedeflenmiştir.

2. Materyal ve Metot

2.1. Veri Setinin Oluşturulması

Bu çalışmada ilk olarak Türkçe dilinde mültecilere yönelik nefret söylemi veri seti oluşturulmuştur. Veri setinin oluşturulabilmesi için öncelikle verinin sağlanacağı sosyal medya ortamına karar verilmiştir. Özellikle metin şeklinde iletilerin paylaşıldığı ve yoğunlukla siyasi tartışmalar için kullanıldığı görülen Twitter araştırma kapsamında tercih edilmiştir. Tweet'lerin çekilmesi sürecinde sosyal ağ platformlarından veri çekmek için kullanılan SNScrape kullanılmıştır. SNScrape ile "mülteci, mülteciler, göçmen, göçmenler, ülkemdemülteciistemiyorum" anahtar kelimeleri kullanılarak gerçekleştirilen taramalar sonucunda, Twitter'dan 12200 tweet çekilmiştir. Sonrasında çekilen 12200 tweet incelenerek tekrarlayan ve alakasız olan tweet'ler veri setinden silinmiştir. Ardından kalan 10659 tweet'in, konuyla ilgili uzman 2 ayrı kodlayıcı tarafından "Nefret söylemi değil" (0) ve "Nefret Söylemi" (1) etiketleri ile etiketlenmesine geçilmiştir.

Nefret söyleminin tespit edilmesinde belirleyici öğelerin neler olacağına Papcunova ve ark. [8] tarafından yapılan çalışma temel alınarak karar verilmiştir. Bu çalışma kapsamında Papcunova ve ark. nefret söylemine yönelik kuramsal bir tanım ortaya koymak için birinde her yaş, cinsiyetten ve eğitim durumundan bireylerin diğerinde ise psikologların yer aldığı 2 odak grup oluşturmuşlardır. 2 ay boyunca yapılan 8 görüşmede göç ve mülteci konularına yoğunlaşılmıştır. Bu çalışma sonucunda, araştırmacılar nefret söylemini göçmenler bağlamında "şiddet içeren davranışları teşvik eden, insan haklarını reddeden, karalamalar, kaba sözler veya ad hominem saldırıları içeren, olumsuz klişeler kullanan, gerçeği veya tarihi gerçekleri kasıtlı olarak manipüle eden herhangi bir metin" olarak tanımlamışlardır. Etiketlemeler literatürdeki benzer çalışmalar da gözetilerek ve bu tanım temel alınarak yapılmıştır.

Kodlayıcılar tarafından etiketlenen veri setine ilişkin bir kesit Tablo.1'de sunulmuştur.

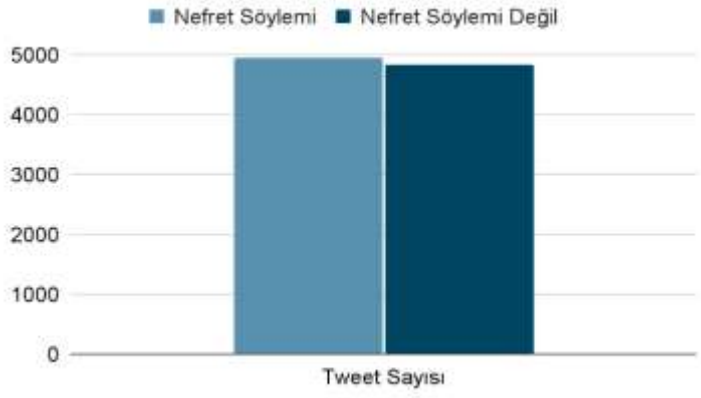
Verilerin etiketlenmesi tamamlandıktan sonra iki kodlayıcının 10659 tweet için %91,73 oranında benzer kodlama yaptığı görülmüşür. İki kodlayıcı arasındaki uyumun ölçülmesi için gerçekleştirilen Cohen's Kappa testi sonucunda Cohen's K değeri 0,835 ($p<0,01$) olarak hesaplanmıştır [9]. McHugh (2012)'a göre iki kodlamacı arasında "Güçlü düzeyde uyum" bulunduğu görülmüştür [10].

Tablo 1. Veri Setinden Bir Kesit

Tweet	Sınıf Etileti
Sığınmacılar, mülteciler adına ne denirse densin Türkiye için her zaman tehlikedir!	1
PKK nın 40 yılda veremediği zararı silahsız mülteciler verdi, kaç para harcadık belli değil.	1
İstanbul'da bayramda mülteciler için ücretsiz ulaşım imkanı tanınmayacak.	0
Avrupaya göç eden Türkler ile mülteciler aynı sebeple ülkelerini terk etmedi	0

Kodlayıcılar tarafından farklı etiketlenen 881 tweet silindikten sonra 4831 adet "Nefret söylemi değil" (NSD) ve 4947 adet "Nefret söylemi" (NS) etiketli toplam 9778 tweet içeren bir veri seti elde edilmiştir. Şekil 1'de yer alan veri setine ilişkin sınıf dağılımı incelendiğinde dengeli bir dağılım olduğu görülmektedir.

Şekil 1. Veri setinin sınıf dağılımı



2.2. Uygulanan Veri Ön İşleme Adımları

Verinin ön işlemesi, beş adımda tamamlanmıştır. Twitter üzerinden paylaşılan metinlerin birçok yazım hatası içerdiği, bazı tweet'lerde Türkçe karakterlerin kullanılmamasından kaynaklı yazım farklılıklarının olduğu görülmüştür. Öncelikle yazım hatalarının düzeltilmesi için Google Dökümanlar tarafından sunulan yazım denetimi servisi kullanılmıştır. Veri seti içerisindeki metinlerin tamamı küçük harfe çevrilerek fazla boşluklar, bağlantılar, karakterler, kullanıcı isimleri ve rakamlar silinmiştir. Metinler, sınıflandırmaya olumlu etkisi olmayan ve modelin başarımını düşüreceği tahmin edilen Türkçe dolgu kelimelerden arındırılmıştır. Son olarak metinde geçen kelimeler ek ve köklerine ayrılmıştır.

2.3. Kelime Temsil ve Sınıflandırma Yöntemleri

Bu çalışmada w2v kelime temsil yöntemi kullanılmıştır. Lojistik regresyon, Karar ağaçları, Destek Vektör Makineleri, Rastgele Orman ve Yapay Sinir Ağı makine öğrenmesi modelleri denenmiş, deneysel sonuçlar f1 ölçütü, kesinlik, geri çağırma ve doğruluk değerleriyle modellerin performansları değerlendirilmiştir. Tüm bu işlemler Python programlama dili kullanılarak gerçekleştirilmiştir.

3. Araştırma Sonuçları ve Tartışma

Veri seti üzerinde uygulanan modellerin performanslarının değerlendirilmesinde kullanılan ölçütler ve alınan sonuçlar Tablo 2'de sunulmuştur. Sonuçlar incelendiğinde doğruluk değerlerinin LR, YSA ve Rastgele Orman modellerinde eşit ve 0,68 olduğu; Karar Ağaçları modeli ile 0,65 doğruluk değerine ulaşıldığı ve en düşük doğruluk değerinin 0,65 ile DVM modeliyle elde edildiği görülmektedir (Tablo 3).

Nefret söyleminin tespiti halen üzerinde çalışılan bir doğal dil işleme problemi olarak karşımıza çıkmaktadır. Özellikle Türkçe dili üzerinde yapılan çalışmaların kısıtlılığı ve kapsamlı bir veri setinin eksikliği göze çarpmaktadır. Literatürdeki çalışmalar incelendiğinde 1000 adet veri ile gerçekleştirilen bir çalışmada, en yüksek performansın %79 f-ölçütü ile Sıralı Minimal Optimizasyon algoritması ile elde edildiği görülmüştür [4]. Başka bir çalışmada ise 40.623 tweet içeren ve sentetik olarak oluşturulmuş bir veri seti üzerinde Naive Bayes, Destek Vektör Makineleri, Karar Ağaçları ve Çok Katmanlı Algılayıcı modelleri uygulanmış ve en yüksek başarımın Karar Ağaçları ve Çok Katmanlı Algılayıcılar ile %80 olarak elde edildiği tespit edilmiştir [11]. DVM modelinin ve TF-IDF kelime temsil yönteminin nefret söyleminin tespitinde sıklıkla tercih edilen yöntemler olduğunu görülmektedir [12]. DVM modeli kullanılarak 14,509 tweet içeren bir veri seti ile yapılan çalışmada %78 doğruluk oranına ulaşılmıştır [13]. Bu çalışmada DVM ile elde edilen performansın diğer modellere göre daha düşük olmasının nedenlerinden birinin kullanılan kelime temsil yöntemi olabileceği düşünülmektedir.

Tablo 3. Uygulanan Modellere İlişkin Sonuçlar

Model	Sınıf Etiketi	Kesinlik	Duyarlılık	f1-ölçütü	Doğruluk
LR	0	0.65	0.81	0.72	0.68
	1	0.75	0.56	0.64	
Karar Ağaçları	0	0.66	0.65	0.65	0.65
	1	0.65	0.66	0.66	
DVM	0	0.59	0.95	0.72	0.64
	1	0.86	0.33	0.48	
Rastgele Orman	0	0.66	0.74	0.72	0.68
	1	0.75	0.56	0.64	
YSA	0	0.66	0.74	0.70	0.68
	1	0.71	0.62	0.66	

Bu çalışma ile literatüre en önemli katkının nefret söylemine yönelik kapsamlı bir Türkçe veri setinin oluşturulması olduğu düşünülmektedir.

4. Sonuç

Suriye İç Savaşını takiben Türkiye'ye Suriye'den yoğun bir göç yaşanmış ve yaşanan bu göç ile mültecilere yönelik olumsuz tepkilerin arttığı görülmüştür. Özellikle sosyal ağlar üzerinden yapılan yorumlarda nefret söyleminin engellenmesi, nefret suçlarının önlenmesi ve toplumsal huzurun korunması açısından önem taşımaktadır. Bu çalışmada nefret söyleminin Twitter sosyal ağı üzerinden Türkçe yapılan paylaşımlarda tespiti üzerinde çalışılmıştır. Öncelikle 9778 adet Türkçe tweet içeren ve "Nefret Söylemi" ve "Nefret Söylemi Değil" şeklinde etiketlenmiş bir veri seti oluşturulmuştur. Bu veri seti üzerinde uygulanan ön işleme adımlarından sonra W2V kelime temsil yaklaşımı kullanılarak çeşitli makine öğrenmesi modelleri uygulanmış ve sonuçlar kesinlik, doğruluk, f1-ölçütü duyarlılık ölçütleri temel alınarak karşılaştırılmıştır. Elde edilen sonuçlar incelendiğinde denenen 5 model içerisinde, Rastgele orman, YSA ve LR ile elde edilen doğruluk değerinin eşit ve DVM ve Karar Ağaçları modellerinden daha yüksek olduğu görülmektedir.

İleriki çalışmalarda veri seti üzerinde farklı kelime temsil yöntemleri, sınıflandırma algoritmaları ve derin öğrenme modelleri denenecektir

Kaynakça

1. Yanık, A. (2017). Sosyal medyada yükselen nefret söyleminin temelleri. *Global Media Journal TR Edition*, *8*(15), 364-383.
2. United Nations. (2022, Ağustos). Hate speech. Erişim Adresi: https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech
3. Aydos, S. S., ve Aydos, O. S. (2019). Yeni medyada nefret söylemi ve nefret söyleminden doğan hukukî sorumluluk. *Ankara Hacı Bayram Veli Üniversitesi Hukuk Fakültesi Dergisi*, *23*(2), 3-35.
4. Mayda, İ., Diri, B. ve Dalyan, T. (2021). Türkçe Tweetler üzerinde Makine Öğrenmesi ile Nefret Söylemi Tespiti. Avrupa Bilim ve Teknoloji Dergisi, (24), 328-334.
5. Ozkaya, U., Melgani, F., Bejiga, M. B., Seyfi, L., & Donelli, M. (2020). GPR B scan image analysis with deep learning methods. *Measurement*, *165*, 107770.
6. Watanabe, H., Bouazizi, M., & Ohtsuki, T. (2018). Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE access*, *6*, 13825-13835.
7. Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., Bhamidipati, N. (2015). Hate speech detection with comment embeddings, In Proceedings of the 24th international conference on world wide web, ss. 29-30
8. Papcunová, J., Martončik, M., Fedáková, D., Kentoš, M., Bozogáňová, M., Srba, I. & Adamkovič, M. (2021). Hate speech operationalization: a preliminary examination of hate speech indicators and their structure. *Complex & Intelligent Systems*, 1-16.
9. Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37-46.
10. McHugh, M. L. (2012). Interrater reliability: the kappa statistic. Biochemia Medica, 22(3), 276-282.
11. Baydoğan, V. C., ve Alatas, B. (2021). Çevrimiçi Sosyal Ağlarda Nefret Söylemi Tespiti için Yapay Zeka Temelli Algoritmaların Performans Değerlendirmesi. *Fırat Üniversitesi Mühendislik Bilimleri Dergisi*, *33*(2), 745-754.
12. Simon, H., Baha, B. Y., & Garba, E. J. (2022). Trends in machine learning on automatic detection of hate speech on social media platforms: A Systematic review. *FUW Trends in Science & Technology Journal*, *7*(1), 001-016.
13. Malmasi, S., & Zampieri, M. (2017). Detecting hate speech in social media. *arXiv preprint arXiv:1712.06427*.