Bilgisayar Bilimleri ve Mühendisliği Dergisi

Cilt-16
Sayı-1
2023

Derin ve Sığ Makine Öğrenmesi Yöntemleri ile Türkçe Tweet'lerden Saldırgan Dil Tespiti

Offensive Language Detection from Turkish Tweets with Deep and Shallow Machine Learning Methods

Pelin CANBAY
Kahramanmaraş Sütçü İmam Üniversitesi
Bilgisayar Mühendisliği Bölümü
Kahramanmaraş, Türkiye
pelincanbay@ksu.edu.tr
ORCID: 0000-0002-8067-3365

Ekin EKİNCİ
Sakarya Uygulamalı Bilimler Üniversitesi
Bilgisayar Mühendisliği Bölümü
Sakarya, Türkiye
ekinekinci@subu.edu.tr
ORCID: 0000-0003-0658-592X

Öz

Nefret söylemi, bir kişiye veya bir gruba yönelik nefreti ifade eden veya şiddeti teşvik eden söylemlerin genel adıdır. Bu söylemler son zamanlarda dijital ortamlarda kontrol edilemez bir şekilde artmıştır. Özellikle Twitter gibi sosyal mecralardaki yazılı nefret söylemleri hem kişiler hem de topluluklar için tehlikeli boyutlara ulaşmıştır. Nefret söyleminin dijital ortamlarda kolaylıkla ve hızlıca yayılabilmesinin önüne geçebilmek için bu söylemleri otomatik tespit edebilecek sistemlere ihtiyaç vardır. Çalışmamızda, en yaygın nefret söylemlerinden biri olan 'saldırgan' söylemleri otomatik olarak tespit edebilen yapay zeka modelleri ele alınmıştır. Derin ve sığ makine öğrenmesi yöntemlerinin karşılaştırmalı olarak kullanıldığı çalışmamızda, Türkçe tweetler'deki söylemler saldırgan veya değil olmak üzere 2 kategoriye ayrılabilmektedir. Yaklaşık %75-%25 dengesizliğindeki bir veri kümesini kullanarak geliştirdiğimiz modellerde, doğruluk ölçeğinde 0,85, f-skor ölçeğinde 0,74 oranında başarılı sonuçlar elde edilmiştir. Veri kümesinde bulunan tweetler'in terim frekansı-ters doküman frekansı (tf-idf) vektörleri kullanılarak eğitilen sığ modeller ile sözcük yerleştirmeleri kullanılarak eğitilen derin modellerden elde edilen sınıflandırma sonuçları karşılaştırmalı olarak bu çalışmada sunulmuştur. Yapılan deneysel çalışmalar ile Çift-Yönlü Uzun Kısa Süreli Bellek (BiLSTM) tekniği kullanılarak geliştirilen saldırgan söylem tespit modelinin, sığ yöntemlerden ve diğer bazı derin öğrenme yöntemlerinden daha başarılı sonuçlar ürettiği gösterilmiştir.

Anahtar sözcükler: Derin öğrenme, Makine öğrenmesi, Nefret söylemi, Saldırgan söylem, BiLSTM

Abstract

Hate speech is the general name for speech that expresses hatred towards a person or a group or encourages violence. These discourses have recently increased uncontrollably in digital environments. Written hate speech, especially on social media such as Twitter, has reached dangerous dimensions for both individuals and communities. In order to prevent the spread of hate speech in digital environments easily and quickly, systems that can automatically detect these speeches are needed. In our study, artificial intelligence models that can automatically detect 'offensive' speech, which is one of the most common hate speeches, are discussed. In our study, in which deep and shallow machine learning methods are used comparatively, the discourses in Turkish tweets can be divided into 2 categories as offensive or not. In the models we developed using a dataset with an imbalance of approximately 75%-25%, successful results are obtained with a rate of 0.85 on the accuracy and 0.74 on the f-score. The classification results obtained from shallow models trained using term frequency-inverse document frequency (tf-idf) vectors of tweets in the dataset and deep models trained using word embeddings are presented comparatively in this

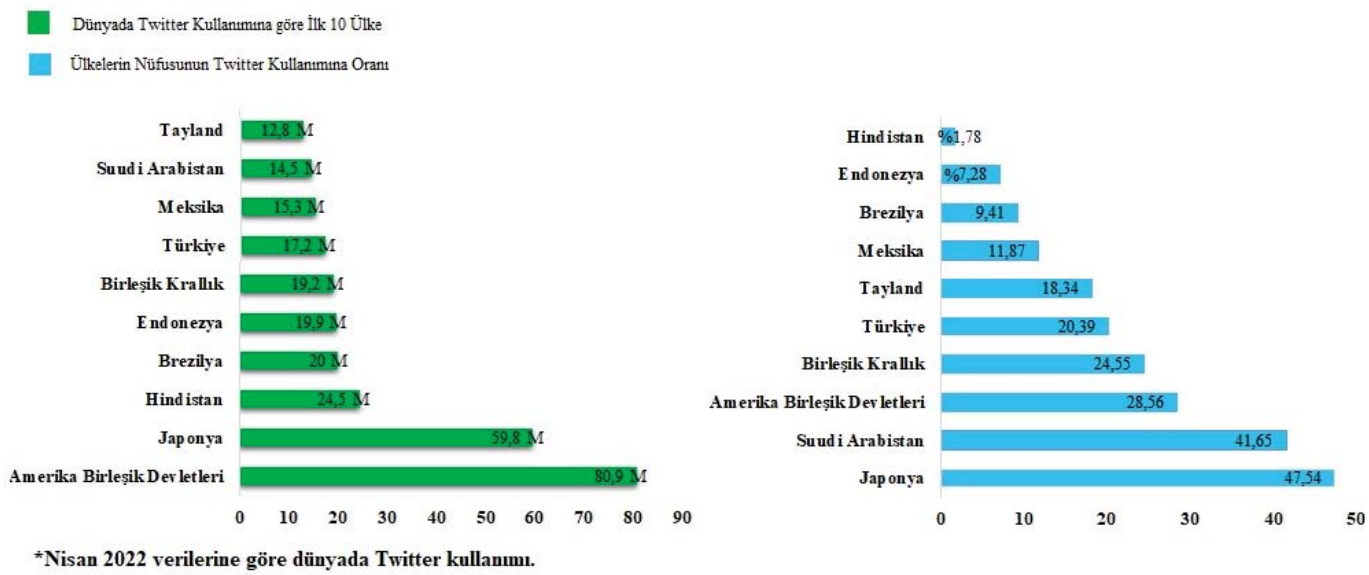
Gönderme, düzeltme ve kabul tarihi: 31.08.2022 - 19.10.2022 – 25.10.2022
Makale türü: Araştırma

study. Experimental studies have shown that the hate speech detection model developed using Bidirectional Long Short-Term Memory (BiLSTM) technique produces more successful results than shallow methods and some other deep learning methods.

Keywords: Deep learning, Machine learning, Hate speech, Offensive speech, BiLSTM

1. Giriş

Günümüzde sosyal hayat artık dijital ortamlara taşınmış durumdadır. Sosyal medya kullanımı dünya genelinde her gün hızla artarak büyümesini sürdürmektedir. İnsanlar sosyal medya üzerinden düşüncelerini, duygularını kolaylıkla dile getirebilmekte ve bunları kısa sürede geniş kitlelere duyurabilmektedir. Bu kolaylık ve hız çoğu zaman bireyleri ve toplumları olumsuz yönde etkileyebilecek sonuçlara da sebebiyet vermektedir. Günümüzün en yaygın kullanılan sosyal medya platformlarından biri olan Twitter, Türkiye genelinde de yoğun olarak kullanılmakta ve sosyal hayatın akışını önemli ölçüde etkilemektedir. Nisan 2022 tarihli, dünya genelinde Twitter kullanıcı sayısı en yüksek olan ilk on ülkedeki kullanıcı sayıları ve bu sayıların ülke nüfusuna oranları Şekil-1'de gösterilmektedir [1].



Şekil-1: Twitter kullanıcısının en çok olduğu ilk 10 ülke (sağda) ve ülkelerin Twitter kullanım oranları (solda)

Şekil-1'de görüldüğü üzere Türkiye, 17,2 milyon kullanıcısı ile Twitter uygulamasını en çok kullanan 7. ülke durumundadır. Bu oran ülke nüfusu ile karşılaştırıldığında, Türkiye, nüfusunun yüzde 20,39'unun Twitter kullanması ile bu uygulamayı en yaygın kullanan 5. ülke durumuna gelmektedir. Bu oran her geçen gün artmaktadır. Bu artış, bir doğal afetin kısa sürede geniş kitlelere duyurulması gibi önemli bir konuda insanlığa faydayı artırırken, bir nefret dalgasının veya bir yanlış haberin hızlıca geniş kitleler tarafından benimsenmesi gibi istenmeyen durumların artmasına da sebebiyet vermektedir.

Nefret söylemi son yıllarda hem yüz yüze hem de çevrimiçi iletişimde hızla büyüyen bir suçtur [2]. Bu çalışmada, Türkiye'de özellikle Twitter ortamında son yıllarda hızla yaygınlaşan ve önüne geçilmesi acil bir ihtiyaç olan nefret söyleminin yapay zeka ile otomatik tespiti amaçlanmıştır. Nefret söylemi, dilin saldırganlığından oluşan, insanların başkaları hakkındaki olumsuz düşünceleri, küfürleri, cinsiyetçilik, ırkçılık gibi toplumu ayrıştıran söylemleridir. Nefret, güncel Türkçe sözlüğüne göre ikiye ayrılarak tanımlanmıştır. İlk tanımında "bir kimsenin mutsuzluğunu istemeye yönelik duygu", ikinci tanımında ise "tiksinme, tiksinti" olarak açıklanmıştır [3]. Avrupa Konseyi'ne göre nefret söylemi, ırkçı nefreti, yabancı düşmanlığını veya hoşgörüsüzlüğe dayalı diğer her türlü nefret biçimini yayan, kışkırtan, teşvik eden veya meşrulaştıran her türlü ifade biçimini kapsar [4]. Bu tip ifadelerin toplumlarda yaygınlaşması ile işlenen suç miktarı giderek artmaktadır. Özellikle sosyal ortamın sağladığı anonim olabilme olanakları sebebi ile birçok durumda yüz yüze işlenemeyecek nefret suçları sosyal ortamlarda kolaylıkla işlenebilmektedir [5]. Diğer taraftan dijital teknolojilerin ve Doğal Dil İşleme (DDİ) alanındaki çalışmaların da hızla gelişmesi ve ilerlemesi ile sosyal medyanın anonimliğinden kaynaklı sorunların bir kısmı aşılmıştır. Siber zorbalık, ayrımcılık, küfür, kaba konuşma, hakaret, kışkırtma, öfkeli söylemler, aşırıcılık propagandası ve belirli bir kesime yönelik saldırgan söylemler gibi nefret söylemi ile yakından ilgili kötü niyetli söylemlerin dijital ortamlarda otomatik tespitine yönelik birçok başarılı DDİ çalışması mevcuttur [2]. Yapılan çalışmaların büyük bir çoğunluğu İngilizce veya diğer popüler dillerde olup Türkçe dilinde bu alanda sayılı çalışma bulunmaktadır. Twitter verileri kullanılarak kadına yönelik nefret söyleminin otomatik tespiti [6], sosyal medyadaki dini azınlıklara, LGBT ve kadına yönelik nefret suçu olayları üzerinden sosyal medyada nefret söyleminin otomatik olarak tanımlanması [7], nefret söylemi içeren haberler taranarak oluşturulan veri kümesi üzerinden nefret söylemi tespiti [8] ve nefret söylemi içeren tweet'lerden nefret söylemi tespiti [9] gibi sayılı güncel çalışmalar Türkçe literatürde mevcuttur.

Bu çalışmada, özellikle Türkçe sosyal ortamlarda son zamanlarda hızla yaygınlaşan "saldırgan" dilin otomatik tespitine yönelik bir araştırma ve uygulama faaliyeti yürütülmüştür. Çalışmamızda ele aldığımız, günümüzde en yaygın nefret söylemlerinden biri olan saldırgan dilin otomatik tespiti için, Çöltekin'in derlemiş olduğu, uluslararası bir DDİ yarışması olan SemEval-2020 Task 12: Sosyal Medyadaki çok dilli metinlerin saldırgan dil tespiti için sunulan, Türkçe tweet'lerden oluşan veri kümesi kullanılmıştır [10]. Kullanılan veri kümesi iki sınıflı, yaklaşık %75-%25 oranında dengesizliğe sahip bir veri kümesidir. Bu veri kümesinin uluslararası çalışmalarda da olduğu gibi kullanılarak kabul görmüş olması sebebi ile ek bir veri dengeleme işlemi yapılmamış, veri kümesi olduğu gibi kullanılmıştır. Literatürdeki çalışmalardan farklı olarak çalışmamızda, hem metinlerin tf-idf ağırlıkları kullanılarak sığ/geleneksel makine öğrenmesi yöntemleri ile sınıflandırmalar yapılmış hem de sözcük yerleştirmeleri kullanılarak derin makine öğrenmesi yöntemleri ile sınıflandırmalar yapılmıştır. Çalışmamızda sığ yöntemlerden Lojistik Regresyon (LR), Destek Vektör Makinaları (SVM), Karar Ağaçları, Çok Terimli Naive Bayes (MNB), Rastgele Orman (RF) ve K-En Yakın Komşuluk (kNN) algoritmaları kullanılmıştır. Derin yöntemlerden ise Evrişimli Sinir Ağları (CNN), Uzun Kısa Süreli Bellek (LSTM), Kapılı Yinelemeli Ağlar (GRU), Çift Yönlü Kapılı Yinelemeli Ağlar (BiGRU) ve BiLSTM algoritmaları kullanılmıştır. Toplamda 11 farklı yapay zeka modelinin uygulaması ve sonuçların karşılaştırması gerçekleştirilmiştir. Kullanılan veri kümesinin dengesiz olması sebebi ile sonuçlar arası karşılaştırma f-skor ölçeği ile yapılmış olup en yüksek başarı sözcük yerleştirme yöntemi temsili ile BiLSTM tekniğinin kullanıldığı modelde 0,74 olarak elde edilmiştir.

Çalışmanın ikinci bölümünde ilgili çalışmalar, üçüncü bölümünde çalışmada kullanılan veri kümesi, önişleme adımları ve kullanılan metotlar ayrıntıları ile verilmiştir. Dördüncü bölümde deneysel çalışmalar ve elde edilen sonuçlar, son bölümde ise çalışmanın genel bir değerlendirmesi yapılmış olup gelecek çalışmalardan bahsedilmiştir.

2. İlgili Çalışmalar

Metinlerden, özellikle sosyal ortamdaki dijital metinlerden nefret söylemi çıkarımı/tespiti günümüz popüler problemlerinden biri olması sebebi ile bu alanında son yıllarda birçok çalışma yapılmıştır. Çalışmalar hem farklı dilleri kapsamakta hem de farklı nefret söylemi çeşitlerini değerlendirmektedir. Aşağıda ele almış olduğumuz nispeten popüler çalışmalarda başta kullanılan nefret söylemi türüne göre daha sonra da kullanılan metin diline göre çalışmaların dağılımı ve güncel yaklaşımların derlemesi sunulmuş, son olarak Türkçe dili özelinde yapılan nefret söylemi tespiti çalışmaları incelenmiş ve ayrıntıları sunulmuştur.

Twitter'da gazetecilere yönelik nefret söylemini, Charitidis ve ark. İngilizce, Fransızca, Almanca, İspanyolca ve Yunanca dillerinde incelemişlerdir [11]. Çalışmada elde edilen veri kümelerine CNN, Skipped CNN (sCNN), CNN-GRU, BiLSTM, LSTM + Attention (aLSTM) ve gradient boosting karar ağaçları uygulanarak başarılı sonuçlar elde edilmiştir. Guellil ve ark. sosyal medyada Arap toplumundaki politikacılara yönelik nefret söyleminin tespitine yönelik bir yaklaşım önermişlerdir [12]. Öncelikle veri kümesindeki öznitelikler word2vec ve fasttext'in hem skip-gram hem de CBOW özellikleri ile çıkarılmış ve bu özellikler makine öğrenmesi ve derin öğrenme algoritmaları ile sınıflandırmada kullanılmıştır. Skip-gram üzerinden Lineer Destek Vektör Sınıflandırıcı (LSVC), BiLSTM ve Çok Katmanlı Algılayıcı (ÇKA) ile en iyi başarı elde edilmiştir. ABD, Birleşik Krallık ve Kanada İngilizcesinde göçmenlikle ilgili toplanan ve manuel etiketlenen tweet veri kümesi üzerinde göçmelik karşıtı twetlerin tespiti gerçekleştirilmiştir [13]. Sözcük n-gramları ve karakter n-gramları ile elde edilen bir dizi özellik NB, SVM ve LR kullanılarak sınıflandırılmış ve sözcük n-gramlarının karakter n-gramlarından daha iyi performans gösterdiği görülmüştür. Rus sosyal medya sitelerindeki metinlerinden oluşturulan RuEthnoHate isimli veri kümesinden etnik kökene dayalı nefret söylemini tespit etmek için üç sınıflı örnek tabanlı bir yaklaşım sunulmuştur [14]. Veri kümesi temsili uni-gramlar, Word2vec-Ethno ve Word2vec-RNC, RuBERT-emb ile gerçekleştirilmiştir. Sınıflandırma aşamasında NB, LR, SVM, oylamalı sınıflandırıcı, LSTM-GRU ve Convers-RuBERT modelleri kullanılmıştır. Convers-RuBERT ile tüm diğer yöntemlere göre daha yüksek bir sınıflandırma başarısı elde edildiği sonucuna varılmıştır. Cinsiyet üzerinden iki sınıflı ve çok sınıflı farklı dillerde nefret söylemi tespiti üzerine çalışmalar da literatürde yer almaktadır [15]–[17]. Kullanılan yöntemler içerisinden BERT'in sınıflandırma başarısının oldukça yüksek olduğu da görülmüştür.

Literatürde en yaygın nefret söylemi tespit çalışmaları dil bakımından İngilizce dili üzerine gerçekleştirilmiştir. İngilizce dilindeki 5 farklı veri kümesi üzerinden nefret söylemi sınıflandırmasını, Wullach ve ark. karakter seviyesindeki HyperNetworks mimarileri kullanarak gerçeklemişlerdir [18]. Önerilen mimari metinleri klasik mimarilerdeki sözcük düzeyinin aksine karakter düzeyinde işlemektedir. Çalışmada önerilen bu yöntem klasik yöntemler ile de karşılaştırılmış ve üstün performans sağlandığı görülmüştür. Kan ve ark. üç farklı tweet kümesi üzerinden nefret söylemi tespit etmek amacıyla BERT ile birlikte derin CNN, BiLSTM ve hiyerarşik dikkat mekanizmalarının gücünü kullandıkları BiCHAT modelini önermişlerdir [19]. Önerilen bu model temel yöntemlere kıyasla oldukça başarılı sonuçlar elde etmiştir. İspanyolca dilinde yazılmış iki veri kümesi üzerinden sosyal medyadaki nefret söylemlerini sınıflandırmada farklı makine öğrenmesi tekniklerinin karşılaştırmalı bir analizi Plaza-del-Arco ve ark. tarafından sunulmuştur [20]. Transfer Öğrenmeye dayalı modellerin, özellikle İspanyolca diliyle eğitilmemiş olsalar bile en iyi sonuçları elde ettiğini göstermişlerdir. Bir başka çalışmada İspanyolcada nefret söylemi tanımlamasına ilişkin çeşitli veri kümeleri incelenerek nefret söylemi tespiti için en iyi özellikleri bulmak, bu özelliklerin nasıl birleştirilebileceğini, dilsel özelliklerin nefret söyleminin tanımlanmasına ilişkin katkı sağlayıp sağlamayacağı araştırılmıştır [21]. Kullanılan özellikler kümesi dilsel özellikler, sözcük yerleştirme, cümle yerleştirme ve BERT şeklinde olup dilsel özellikler ve BERT'in LR'ye dayalı topluluk öğrenme stratejisini önermişlerdir.

Dil ailesi bakımından İngilizceden farklı olarak diğer yaygın nefret söylemi çalışmaları Arapça dili üzerine yapılmıştır. Arapça dili üzerine, Duwairi ve ark. Arapça veri kümesi olan ArHS ve birden fazla veri kümesinin birleşimi üzerinden CNN, CNN-LSTM, ve BiLSTM-CNN derin öğrenme yöntemlerinin nefret söylemi sınıflandırmadaki başarımını ölçmeyi hedeflemişlerdir [22]. Çalışma kapsamında ikili, üçlü ve çoklu olmak üzere üç farklı sınıflandırma gerçekleştirilmiştir. ArHS veri kümesinde en yüksek başarının CNN ile ikili sınıflandırmada %81 doğruluk ile sağlandığı görülmüşken, birleştirilerek oluşturulan veri kümesinde de yine ikili sınıflandırmada en yüksek başarı BiLSTM-CNN ile %73 olarak elde edilmiştir. Bir başka Arapça dili üzerine yapılan çalışmada Arapça tweet'ler beş farklı nefret söylemi olan dini, ırkçı, cinsiyetçi, genel nefret veya hiçbiri kategorilerine ayrılmaktadır [23]. On bir bin tweet'ten oluşan bir veri kümesinin toplandığı ve etiketlendiği çalışmada SVM modeli, LTSM, CNN + LTSM, GRU ve CNN + GRU derin öğrenme modelleri ile karşılaştırılmaktadır. Bulgular, nefret tweet'lerinin sınıflandırılmasında, dört derin öğrenme algoritmasının hepsinin SVM modeline üstünlük sağladığını göstermiştir. Kalra ve ark. Urduca twett'lerden nefret söylemini tespit edebilmek için dönüştürücü tabanlı bir yöntem olan Roberta'yı kullanmışlardır [24]. Bir diğer Urdu dilindeki çalışmada [25] tweet'lerden oluşan çok sınıflı veri kümesi üzerinden nefret söylemi tespiti amaçlanmıştır. Çalışma kapsamında ilk olarak nefret söylemi sözlüğü oluşturulmuş, daha sonra bu sözlük kullanılarak veri kümesi etiketlenmiştir ve makine öğrenmesi algoritmaları ile sınıflandırma yapılmıştır. Bir diğer deneyde ise FastText ve çok dilli BERT modelleri kullanılmıştır. Son olarak dört farklı BERT ile eğitim gerçekleştirilmiş ve BERT, xlm-roberta ve distil-BERT'in çoklu sınıflandırmada başarılı olduğu görülmüştür.

Türkçe için yapılan çalışmalar da son yıllarda önem kazanmaya başlamıştır. Hüsnübeyi çalışmasında 18.318 ulusal ve yerel haberden oluşan metinlerin nefret söylemi veya nefret söylemi olmayan olarak etiketlenmiş bir veri kümesini kullanarak sınıflandırma yapmıştır [8] Çalışmada, farklı sözcük temsilleri ile metnin hiyerarşik yapısını kullanarak ifadelerin değişen anlamlarını yakalamayı amaçlayan Hiyerarşik Dikkat Ağı (HAN) modeli kullanılmıştır. Önerilen yöntemin makine öğrenmesi ve CNN'e üstünlük sağladığı görülmüştür. Şahi ve ark. #kiyafetimekarisma hastagini kullanarak kadınlara yönelik nefret söylemi ile ilgili tweet'leri toplayıp sınıflandırma görevini gerçeklemişlerdir [6]. Özellik olarak tf-idf değerlerine göre ağırlıklandırılmış bi-gram, karakter, sözcük, cümle ve hece sayısını ve bu özelliklerin alt kümelerini kullanarak 5 farklı makine öğrenme algoritması ile sınıflandırma yapmışlardır. Dağaşan tarafından hazırlanan yüksek lisans tezinde Twitter'daki etnik/dini azınlıklara, LGBT'lere ve kadınlara yönelik nefret suçu söylemlerinin otomatik olarak sınıflandırılabilmesi amaçlanmıştır [7]. Bu bağlamda sözcük n-gramlar (1-3 gramlar), POS ve karakter n-gramlar (2-5 gramlar) üzerinden NB, RF, Lasso LR (LLR), LR ve SVM ile sınıflandırma gerçekleştirilmiştir. Mayda ve ark. toplayıp etiketledikleri 1000 adet tweet üzerinden nefret söylemi, saldırgan ifade, hiçbiri olmak üzere üç sınıflı nefret söylemi tespiti gerçekleştirmişlerdir [9]. Toplanan tweet'ler üzerinden karakteri iki-gram ve üç-gram, sözcük tek-gram, iki-gram ve tweet'lere özgü özellikler çıkartılmış ve sonrasında özellik seçimi yapılmıştır. Elde edilen özellikler üzerinden J48, sıralı minimum optimizasyon (SMO), NB ve RF ile sınıflandırma adımı gerçekleştirilmiştir. Yapılan deneyler sonucunda en başarılı yöntemin sözcük tek-gram, sözcük iki-gram ve karakter üç-gramlar üzerinden seçilen 600 özellik üzerinden SMO olduğu görülmüştür. Karayiğit ve ark. Instagram üzerinden topladıkları Homofobik-İstismarcı Türkçe Yorumları (HATC) kullanarak nefret söylemi sınıflandırmasını gerçeklemişlerdir [26]. Veri kümesinde nefret söylemi, homofobik ve tarafsız olmak üzere üç sınıf bulunmaktadır. HATC veri kümesinden yeniden örnekleme ile resHATc veri kümesi oluşturulmuş ve her iki veri kümesi de dönüştürücüler, derin öğrenme algoritmaları ve makine öğrenmesi algoritmaları ile sınıflandırılmıştır. Eldeki veri kümeleri için en başarılı yöntemin dönüştürücüler içerisindeki 104 dil ile önceden eğitilmiş M-BERT yöntemi olduğu görülmüştür. Toraman ve ark. ise hem Türkçe hem de İngilizce veri kümelerini kullanmışlardır [27]. Türkçe için 100k'lık bir tweet kümesi toplanmıştır. Toplanan veri kümesi din, cinsiyet, ırk, siyaset ve spor üzerinden beş farklı sınıftaki nefret söylemlerinden oluşmaktadır. Yapılan deneyler sonucunda dönüştürücü tabanlı dil modellerinin geleneksel modellerden daha iyi olduğu görülmüştür.

Çalışma kapsamında kullanılan veri kümesi üzerinden gerçekleşen yarışmada toplamda 46 takım yarışmıştır ve bu takımlar içerisinde en yüksek f-skor %82,58 iken en düşük f-skor %31,09 olarak elde edilmiştir [10]. Çalışmamızdan elde edilen sonuçlar uluslararası müsabakada elde edilen sonuçlara kıyasla kabul edilebilir ortalama bir başarıda olup, hem Türkçe olması hem de 11 farklı yöntemi karşılaştırması bakımından konu ile ilgili gerçekleştirilecek ileriki çalışmalara katkı sağlayacağı öngörülmektedir.

3. Yöntem

Bu bölümde, öncelikle çalışmada kullanılan veri kümesi ve uygulanacak makine öğrenmesi yöntemlerinde kullanılacak olan veri temsilleri açıklanmıştır. Daha sonra, uygulamalarda kullanılan sığ ve derin makine öğrenmesi yöntemlerinin özellikleri ve kullanım nedenleri sunulmuştur.

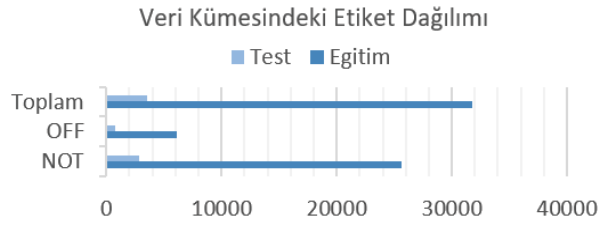
3.1 Veri Kümesi ve Ön İşleme

Çalışmamızda kullanılan veri kümesi, Çağrı Çöltekin'in, çalışmasında, Türkçe saldırgan dil tespiti amacı ile derlemiş olduğu ve SemEval-2020 Task 12 (OffensEval 2020) da kullanılan veri kümesidir [10]. Veri kümesi eğitim ve sınama olarak iki dosyadan oluşmaktadır. Toplanan verilerdeki kullanıcı isimleri yerine @user ve bağlantı adresleri yerine url eklenerek güncellemeler yapılmıştır. Eğitim kümesinde id, tweet ve tweet'in etiketi (OFF/NOT) yer almaktadır. Veri kümesi incelendiğinde bazı örneklerde birden fazla tweet'in tek örnek olarak eklendiği tespit edilmiştir. Bu tweet'ler ayrıştırılarak veri kümesine eklenmiştir. Sınama kümesi ise, id ve tweet ikililerinin olduğu bir csv dosyası ve id ile etiket ikililerinin olduğu bir etiket dosyasından oluşmaktadır. Sınama kümesine, uygulamalarda kullanılabilmesi için öncelikle id eşleşmeleri üzerinden tweet etiketlerinin ataması yapılmıştır. Kullanacağımız veri kümesi temizlenmemiş

metinlerden oluştuğu için bir ön işleme adımından geçirilerek kullanılacak temiz bir veri haline dönüştürülmüştür. Ön işlem adımında, veri kümesini oluşturan tüm tweet'ler için aşağıdaki işlemler gerçekleştirilmiştir:

- Özel karakterlerin silinmesi (@, #, …),
- Noktalama işaretlerinin temizlenmesi,
- Tüm sözcüklerin küçük harfe çevrilmesi
- Türkçe dilsiz sözcükler (stopwords) kaldırılması

Veri kümesi iki sınıf etiketinden oluşmaktadır, OFF (saldırgan) – NOT (saldırgan değil). Şekil-2' de sınıf etiketlerinin eğitim ve sınama kümelerindeki dağılımları bulunmaktadır.

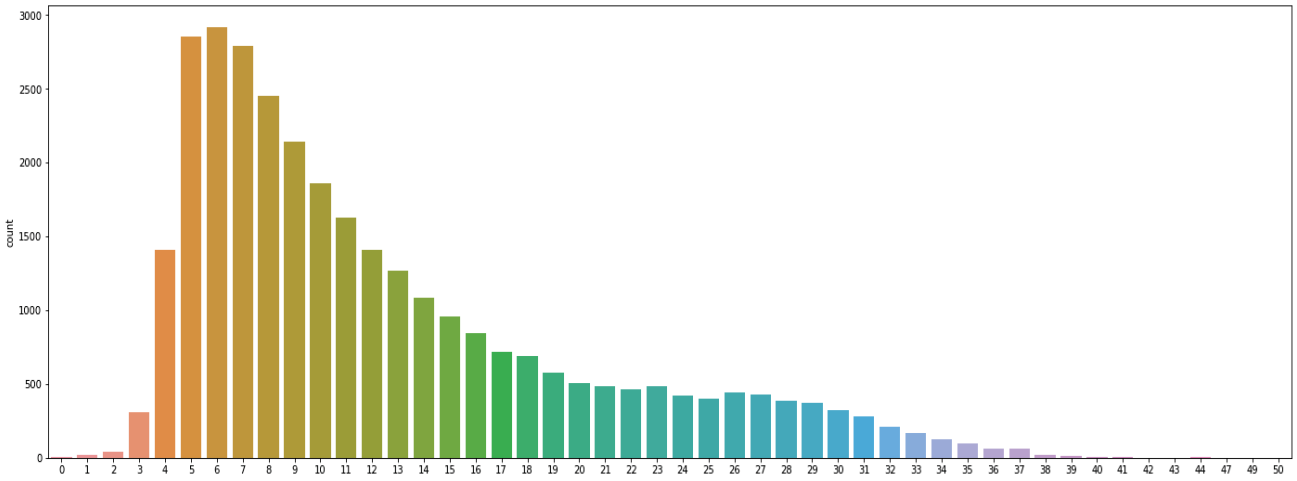


	NOT	OFF	Toplam
Test	2812	716	3528
Egitim	25625	6131	31756

Şekil-2: Eğitim ve sınama kümesindeki etiket dağılımları

Şekil 2'de görülebildiği üzere, hem eğitim kümesinde hem de sınama kümesinde iki sınıfın (OFF/NOT) dağılımı dengesizdir. Bu aşamada yapılacak sınıflandırma işlemlerinin büyük oranda çoğunluğun olduğu sınıfı ezberleme eğiliminde olacağı öngörülmektedir. Hem verinin sosyal medyadaki mevcut oranları hem de aynı veri kümesinin literatürde de bire bir kullanımı göz önüne alındığında, bu çalışmada da veri dengeleme tekniklerinin kullanımı tercih edilmemiştir. Özellikle mevcut dengesizliğe rağmen başarılı sonuçlar üretebilecek modellerin üzerinde durulmuştur.

Eğitim kümesinde toplamda 31.756 tweet bulunmakta. Ortalama tweet uzunluğu 12,5 sözcüktür. Eğitim kümesinde bulunan tweet – sözcük dağılımı Şekil-3'te gösterilmektedir.



Şekil-3: Eğitim kümesinde bulunan tweet'lerdeki sözcük sayısı dağılımı

Şekil-3 incelendiğinde tweet'lerde kullanılan sözcük sayısının 50'den az olduğu görülebilmektedir. Eğitim kümesi üzerinden elde edilen bu bilgiler ile kullanılan makine öğrenmesi yöntemlerinin bazı parametreleri belirlenebilmektedir.

3.2 Sözcük Temsilleri

3.2.1 TF-IDF

Tf-idf, DDİ çalışmalarında sıklıkla kullanılan bir sözcük ağırlıklandırma yöntemidir [28]. İstatistiksel bir ölçüm yöntemi olan tf-idf bir derlemdeki dokümanlarda bulunan sözcüklerin, dokümanların ayırt edilmesinde ne kadar etkili olduklarını hesaplayan bir ölçümdür. Bu ölçüm; t sözcük, d doküman, D derlem ve N derlemdeki toplam faklı sözcük sayısı olmak üzere; denklem (1), (2) ve (3)'teki gibi tf ve idf değerlerinin çarpımından elde edilir.

$$tf(t,d) = log\ (1 + frekans(t,d)) \quad (1)$$

$$idf(t,D) = \log(\frac{N}{|(d \in D : t \in d)|}) \quad (2)$$

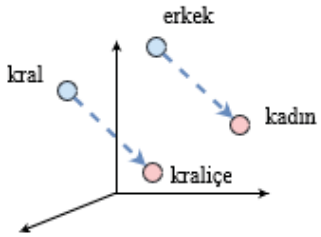
$$tf - idf(t,d,D) = tf(t,d) * idf(t,D) \quad (3)$$

TF-idf gibi farklı sözcük ağırlıklandırma yöntemleri özellikle metin sınıflandırma çalışmalarında sıklıkla kullanılmaktadır [29]. Veri kümesinde bulunan her dokümanın, barındırdığı sözcüklerin tf-idf ağırlıklarına göre tf-idf vektörüne

dönüştürüldüğü bu işlem sonucunda elde edilen vektörler, sığ makine öğrenmesi algoritmalarıyla çalışmamızda saldırgan söylem tespitinde kullanılmıştır.

3.2.2 Sözcük Yerleştirme

Sözcük yerleştirme (word embeddings), sözcüklerin sabit uzunluktaki vektörel temsillerine verilen addır [30]. Sözcük yerleştirmeleri, özellikle büyük veri kümelerinde bulunan dokümanlardaki sözcükler arası kullanım ilişkileri üzerinden soyut anlamsal bir yapı elde eder ev bu anlamsal yapı her sözcük için n boyutlu sayısal bir temsil sunar [30]. Veri işleme araçlarındaki gelişmeler ile artan derin öğrenme çalışmalarında sözcük yerleştirmeleri, metin sınıflandırmadan özet çıkarımına, soru cevap sistemlerinden otomatik çeviri sistemlerine kadar birçok NLP çalışmasında yaygın olarak kullanılmaktadır. Şekil-4'te sözcük yerleştirmeleri kullanımının n boyutlu uzayda anlamsal bağlamı koruduğunu göstermek amacı ile literatürde kullanılan bir temsil sunulmuştur.



Şekil-4: Sözcük yerleştirme yerleştirmeleri ile anlamsal ilişkinin korunması

Şekil-4'te de gösterildiği gibi sözcük yerleştirme yöntemi, yeterli ve uygun bir veri kümesi kullanıldığında, benzer anlamdaki sözcükleri benzer gösterimde sunabilme kabiliyetine sahiptir. Ayrıca, sözcük yerleştirme yönteminin tf-idf veya sözcük kesesi (bag of words) gibi temsil yöntemlerine oranla yüksek boyutluluk ve seyreklik sorunları da yoktur. Literatürde Word2vec [31], Glove [32], ve Fasttext [33] gibi hazır (pre-trained) sözcük yerleştirme bulunmaktadır. Bu çalışmada, hazır sözcük yerleştirme yerine kendi eğitim verisinden sözcük yerleştirmeleri eğitilmiştir. Veri kümesinden eğitilecek sözcük yerleştirmeleri aşağıdaki parametreler ile işleme sokulmuştur.

- Sözlük boyutu = 100.000
- Maximum tweet uzunluğu = 50
- Sözcük yerleştirme boyu = 200

Verilen değerler hem Türkçede yaygın kullanılan sözcük oranları, hem veri kümesindeki bir tweet'teki en fazla sözcük sayısı, hem de uygulamalarda yüksek performans verecek değerler göz önüne alınarak optimize edilmiştir.

3.3 Sığ Makine Öğrenmesi Modelleri

3.3.1 LR

LR, bir veya daha fazla değişkenine dayalı olarak kategorik bir değişkeni tahmin etmek için regresyon analizini kullanan istatistiksel bir ikili sınıflandırma yöntemidir [34]. LR, bağımlı değişkenin değerinin bağımsız değişkenler kullanılarak tahmin edilmesi ilkesi üzerine kurulmuştur. Modelde sınıf etiketi Y bağımlı değişken iken, X (x_1, x_2, … , x_n) özniteliklere karşılık gelen ve Y tahmininde kullanılan bağımsız değişkenler kümesidir.

3.3.2 SVM

SVM, hem doğrusal hem de doğrusal olmayan verileri sınıflandırabilen istatistiksel öğrenme teorisi temelli bir sınıflandırma algoritmasıdır [35]. Bu algoritma optimal ayırma hiperdüzlemini, yani karar sınırını bularak eldeki veri kümesini iki sınıfa ayırmaktadır. SVM, bu hiper düzlemi bulmak için destek vektörlerini ve kenar boşluklarını kullanır. Veriler doğrusal olarak ayrılabilir ise, bir ayırma hiperdüzlemi iki boyutlu uzaydaki verileri kolaylıkla ayırabilir. Ancak doğrusal olmayan verileri ayırmak için bu verileri daha yüksek bir boyuta dönüştürmek gereklidir. Bu aşamada, SVM bu amaç için doğrusal olmayan çekirdek fonksiyonlarını kullanır.

3.3.3 Karar Ağaçları

Aç-gözlü bir algoritma olan karar ağaçlarında ağaç yukarıdan aşağıya yinelemeli olarak böl-yönet mantığı ile kurulur. Öznitelikler kategoriktir eğer sürekli-değerli ise, önceden ayrıklaştırılması gerekmektedir. Karar ağacı oluşturulurken belirli bir yaprak olmayan düğümde öznitelikler içerisinden optimal bölünmeyi sağlayan seçilmektedir. Optimal bölünme, bilgi kazancı, Gini gibi istatistiksel ölçütlere göre belirlenmektedir. Her bir öznitelik bir düğüm ile temsil edilmektedir. Ağacın en üstünde yer alan düğüm kök düğümdür. Düğümler özniteliklerin sınanmasında kullanılırken dallar öznitelik değerlerini tutmaktadır. Yapraklar ise sınıf etiketlerini tutmaktadır.

3.3.4 MNB

Bayes teoremine dayalı olan Naive Bayes (NB) istatistiksel bir sınıflandırıcı olup sınama kümesindeki verilerin sınıf üyelik olasılıklarını tahmin eder. NB sınıf koşullu bağımsızlık varsayımına dayanmaktadır. Bu varsayım, bir veri kümesindeki belirli bir özniteliğin varlığı ile diğer özniteliklerin varlığı arasında bir ilişki olmadığını söylemektedir. MNB, NB'nin metin formatındaki veriler için tasarlanmış çok değişkenli dağılımı kullanan özel bir versiyonudur [36].

3.3.5 RF

RF torbalama ve rastgele alt uzay kavramlarına dayanan çok sayıda budanmış karar ağacından oluşan bir topluluk sınıflandırıcıdır [37]. Hem torbalama hem de rastgele alt uzay yöntemleri karar ağaçlarında çeşitliliği sağlamaktadır. Her bir karar ağacı rastgele seçilen öznitelik alt kümeleri ile oluşturulur. Karar ağaçları oluşturulurken belirli bir yaprak olmayan düğümde öznitelik alt kümeleri arasından optimal bölünmeyi seçer. Optimal bölünme, bilgi kazancı, Gini gibi ölçütler kullanılarak seçilir.

3.3.6 kNN

kNN, Yakowitz tarafından önerilen parametrik olmayan, gerçeklenmesi kolay ancak bellek gereksinimi yüksek olan bir sınıflandırma algoritmasıdır [38]. Öğrenme süreci sırasında mevcut tüm eğitim verilerini saklar. Yeni bir sınama verisi geldiğinde, bu yeni sınama verisini sınıflandırmak için sınama verisi ile tüm eğitim verisi arasındaki benzerlik hesaplanır.

Hesaplamalar sonucunda sınama verisi kendisine en yakın k komşudaki en yaygın sınıf etiketi ile etiketlenir.

3.4 Derin Makine Öğrenmesi Modelleri

3.4.1 CNN

CNN, girdiyi daha kullanışlı bir temsile dönüştürmek için tasarlanmış matematiksel bir modeldir [39]. CNN, evrişim, havuzlama ve tam bağlantılı katmanlar olmak üzere üç katmandan oluşur [40]. Evrişimsel katman, matematiksel ve doğrusal işlemlerin uygulanması nedeniyle katmanlar arasında anahtardır. Bu katmanda özellik çıkarımı gerçekleştirilir. Havuzlama katmanı, özellik haritasının boyutunu küçültmek için aşağı örnekleme gerçekleştirir. Böylece öğrenilecek parametre sayısı azaltılır. Tam bağlantılı katman, evrişim katmanının ve havuz katmanının özellik haritalarını tek boyutlu bir özellik vektörüne dönüştürür.

3.4.2 LSTM

Tekrarlı Sinir Ağı (RNN), sıralı verileri işlemek için özelleşmiş sinir ağlarına verilen addır. Geleneksel yapay sinir ağlarında çıktı, doğrudan girdilerden elde edilmektedir. RNN, gizli nöronlara döngüsel bağlantılar ekleyerek bu dezavantajın üstesinden gelmek için tasarlanmıştır [41]. Bu şekilde, girdi ve çıktı arasında diziden diziye eşleme gerçekleştirilir, yani önceki hesaplamaya dayalı olarak çıktı elde edilir. Ancak, RNN'nin uzun zaman adımları için yetersiz bellek kapasitesi, güçlü ezberleme yeteneğine sahip yeni bir mimari gerektirmektedir. Dört işlemi (3 sigmoid ve 1 tanh) olan tekrarlı modüle sahip LSTM, bu uzun zaman adımları için bilgilerin ezberlenmesini sağlar. LSTM, bir bellek hücresinden oluşan gizli nöronları ve unutma kapısı, giriş kapısı ve çıkış kapısı olmak üzere üç kapısı olan özel bir RNN mimarisidir. Unutma kapısı, önceki gizli duruma ve mevcut girdiye dayalı olarak hücre durumundan (bellek) hangi bilgilerin alınacağını seçer. Giriş kapısı, hangi bilgilerin güncellendiğini ve hücre durumuna eklendiğini belirler. Çıkış kapısı ile hücre durumunda hangi bilgilerin kullanılacağı belirlenir.

3.4.3 BiLSTM

Hem RNN hem de LSTM modeli, bilginin yalnızca zaman içinde ileriye doğru yayılmasına izin verir [42]. Bağlam bilgisini geçmiş ve gelecek zamandan aynı anda yakalamak için çift yönlü RNN (BRNN) geliştirilmiştir. Daha sonra Graves ve Schmidhuber geçmiş ve gelecek bilgilerin etkin kullanımını sağlamak için BRNN ve LSTM birimlerini birleştiren BiLSTM'i önermiştir [43].

3.4.4 GRU

GRU, LSTM'in biraz daha basitleştirilmiş bir çeşididir ve değişken uzunuklu diziler ile çalışmaktadır. GRU, LSTM'nin unutma kapısını, giriş kapısını ve çıkış kapısını basitleştiren bir modeldir. GRU'daki kapıların aktivasyonları yalnızca mevcut girişe ve önceki çıkışa bağlıdır. Ağın basitleşmesi ve daha az parametre sayesinde GRU kullanan problemler daha hızlı yakınsama eğilimindedir ve LSTM'den daha başarılı olabilmektedir [44].

3.4.5 BiGRU

GRU'nun tek yönlü doğası nedeniyle, bilgileri arkadan öne kodlamak mümkün değildir. BiGRU, ileri GRU ve geri GRU olmak üzere iki GRU'dan oluşan çift yönlü bir modeldir. İleri GRU bilgileri önden arkaya işlerken, geri GRU bilgileri arkadan öne işlemektedir [44].

4. Deneysel Çalışma

4.1 Performans Metrikleri

Önerilen modelleri değerlendirmek için temel değerlendirme metriği olarak f-skorunu kullanılmıştır. F-skoru, kesinlik ve duyarlılığın harmonik ortalamasıdır. Kesinlik (p) doğru sınıflandırılmış saldırgan olmayan tweetlerin (tp) tüm saldırgan olmayan tweetlere (tp+fp) oranıdır. Duyarlılık (r), doğru olarak sınıflandırılan saldırgan olmayan tweetlerin (tp) saldırgan olmayan olarak sınıflandırılan tüm tweetlere (tp+fn) oranıdır. F-skoru denklem 4 ile verilmiştir.

$$F - skoru = 2 \times \frac{p \times r}{p + r} \quad (4)$$

4.2 Deneysel Sonuçlar

Çalışma kapsamında ele alınan tüm uygulamalar Google Colab platformunda, Python 3 ile Tensorflow 2 ve scikit-learn kütüphaneleri kullanılarak gerçekleştirilmiştir. Sığ yöntemlerin uygulamasında CPU, derin yöntemlerin uygulamasında GPU kullanılmıştır.

Çalışma kapsamında ele aldığımız saldırgan söylem tespiti probleminde öncelikle, sığ makine öğrenmesi yöntemlerinden LR, SVM, Karar Ağaçları, ,MNB, RF, LR ve kNN sınıflandırma algoritmalarının ürettiği başarıların değerlendirmesi yapılmıştır. Girdi verisi olarak veri kümesinde bulunan tweetlerin tf-idf vektörel temsillerinin kullanıldığı makine öğrenmesi yöntemlerinden elde edilen sınama kümesi sonuçlarının doğruluk, kesinlik, duyarlılık ve f-skor değerleri Çizelge-1'de sunulmuştur.

Çizelge-1: Sığ makine öğrenmesi yöntemlerinden elde edilen sonuçlar

Yöntem	Doğruluk	Kesinlik	Duyarlılık	F-skor
LR	0,82	0,84	0,57	0,58
SVM	0,82	0,88	0,57	0,57
Karar Ağaçları	0,79	0,67	0,63	**0,65**
MNB	0,80	0,90	0,51	0,46
RF	0,83	0,83	0,60	0,62
kNN	0,80	0,90	0,50	0,45

Çizelge-1'de görüleceği üzere kullanılan yöntemlerden elde edilen doğruluk değerleri yüksek olmasına rağmen F-skor değerleri düşüktür. Bu durum veri kümesindeki dengesizliğin modellerin eğitimini etkilediği, sınıflandırıcı modellerin örnek sayısı daha fazla olan sınıfı ezberleme eğilimine gittiğinin göstergesidir. Değerlendirilen algoritmalar python 3 ile uyumlu scikit-learn kütüphanesindeki algoritmaların varsayılan parametreleri [45] ile kullanılmıştır. Kullanılan veri kümesinde eğitim ve sınama verisi sabit olduğundan bu aşamada çapraz doğrulama yapılmamıştır. Varsayılan parametrelere ek olarak kaba kuvvet yöntemi kullanılarak

güncellenen parametreler ile de deneyler yapılmış fakat elde edilen sonuçlarda bir artış gözlemlenmemiştir.

Yapılan deneyler sonucunda, Karar Ağacının performansının diğer beş algoritmanın performansına kıyasla göre eldeki veri kümesi için daha üstün olduğu görülmüştür. Karar ağacının LR, SVM, MNB ve kNN'ye kıyasla daha üstün olması kural tabanlı olması ile ilişkilendirilebilir. RF'nin karar ağaçlarından daha düşük performans göstermesi ise RF'nin büyük veri kümeleri için daha uygun olmasıdır, çünkü karar ağaçları rastgele oluşturulur ve sonuçlar arasında oylamaya bağlıdır.

Ele aldığımız problemin ve veri kümesinin zorluğunun üstesinden gelebilmek için, sığ makine öğrenmesi yöntemlerinin yanı sıra derin öğrenme teknikleri de kullanılarak deneyler yapılmıştır. Bu aşamada çalışmamızda, CNN, LSTM, BiLSTM, GRU ve BiGRU derin öğrenme teknikleri benzer derin ağ mimarilerinde kullanılmıştır. Veri kümesinde bulunan dokümanlardaki her sözcüğün vektörel temsillerinin çıkarılıp söz konusu modellere eşit boyutlarda dokümanların değerleri girdi olarak verilmiştir. Kullanılan modellerin ayrıntıları Çizelge- 2'de sunulmuştur.

Kullanılan her derin öğrenme modelinde epoch =10 olarak girilmiş fakat, sınama kaybının artmasına bağlı olarak erken durdurma (EarlyStopping) yöntemi ile öğrenme süreci yeterli görüldüğü yerde sonlandırılmıştır. CNN mimarisinde tüm ara katmanlarda "ReLU" aktivasyon fonksiyonu olan denklem (5) kullanılırken, diğer modellerde "hiperbolic tangent (tanh)" aktivasyon fonksiyonu olan denklem (6) kullanılmıştır. Modellerde ikili sınıflandırma gerçekleştirildiği için tüm çıktı katmanlarında 'sigmoid' aktivasyon fonksiyonu olan denklem (7) kullanılmıştır.

Çizelge-2: Kullanılan derin öğrenme mimarilerinin katmanları

Katmanlar	CNN	LSTM	BiLSTM	GRU	BiGRU
1. Katman	Embedding (200)	Embedding (200)	Embedding (200)	Embedding (200)	Embedding (200)
2. Katman	Convolution(128)	LSTM (128)	BiLSTM (128)	GRU(128)	BiGRU (128)
3. Katman	MaxPooling (2)	Dropout (0.2)	Dropout (0.2)	Dropout (0.2)	Dropout (0.2)
4. Katman	Dropout(0.2)	Dense (64)	BiLSTM (64)	Dense (64)	BiGRU(64)
5. Katman	Convolution(64)	Dropout(0.2)	Dense (64)	Dropout(0.2)	Dense (64)
6. Katman	MaxPooling (2)	Dense (64)	Dropout(0.2)	Dense (64)	Dropout(0.2)
7. Katman	Flatten	Dense (1)	Dense (64)	Dense (1)	Dense (64)
8. Katman	Dense (64)		Dense (1)		Dense (1)
9. Katman	Dropout(0.2)				
10. Katman	Dense (64)				
11. Katman	Dense (1)				

$$Relu(x) = \max\{0, x\}, \quad (5)$$

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}, \quad (6)$$

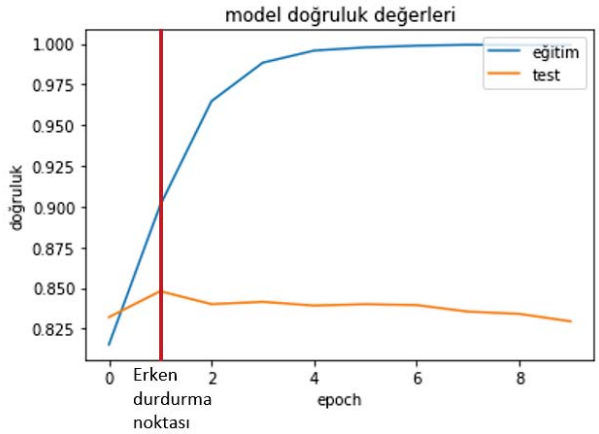
$$sigmoid(x) = \sigma(x) = \frac{1}{1 + e^{-x}}. \quad (7)$$

Tüm modellerde optimizasyon algoritması olarak "Adam", batch size için 32 değeri, kayıp fonksiyonu olarak da "binary_crossentropy" kullanılmıştır. Kullanılan mimarilerden elde edilen sınama kümesi doğruluk, kesinlik, hassasiyet ve makro ortalamalarda f-skor değerleri Çizelge-3'te sunulmuştur.

Çizelge-3: Geleneksel makine öğrenmesi yöntemlerinden elde edilen sonuçlar

Yöntem	Doğruluk	Kesinlik	Duyarlılık	F-skor
CNN	0,74	0,71	0,83	0,73
LSTM	0,40	0,50	0,80	0,44
BiLSTM	0,75	0,73	0,85	0,74
GRU	0,40	0,50	0,77	0,44
BiGRU	0,76	0,71	0,84	0,73

Çizelge-3'te sunulan sonuçlar dikkate alındığında, CNN, BiLSTM ve BiGRU teknikleri kullanılarak elde edilen sonuçlar en başarılı sonuçlardır. Yapılan uygulamalar arasında en yüksek başarı BiLSTM tekniğinin kullanıldığı modelden elde edilmiştir. Modelin öğrenme sürecindeki doğruluk değerleri ve erken durdurma noktası Şekil-5'te gösterilmektedir.



Şekil-5: BiLSTM modelinde elde edilen doğruluk değerleri

Şekil-5'te görüldüğü üzere, BiLSTM tekniği kullanılarak geliştirilen modelde eğitim sürecinin ilk aşamalarında model en yüksek başarıya ulaşmakta ve eğitim arttırıldıkça modelin sınama başarısı düşmektedir. Bu durum fazla eğitim işleminin modeli, eğitim kümesini ezberlemesine ve dolayısı ile sınama kümesinde başarı düşüklüğüne yol açmaktadır. Bu durumlarda en uygun teknik olarak erken durdurma kullanılarak modelden alınabilecek en yüksek performans elde edilmiştir.

BiLSTM'in başarılı olmasının nedeni çift yönlü anlamsal bilgiyi dikkate alması böylece modelin daha fazla anlamsal özellik ile eğitilmesidir. BiGRU'da çift yönlü bir model olduğu için metin sınıflandırmada diğer yöntemlere kıyasla başarılıdır. Yine CNN'in başarılı olması yerel özellik yaklaşımını kullanması ile sağlanmaktadır. Çalışmamızda elde ettiğimiz sonuçlar, veri kümesinin temin edildiği uluslararası SemEval-2020 Task 12 müsabakasında elde edilen sonuçlar ile karşılaştırıldığında kabul edilebilir, ortalama başarıda sonuçlardır. Söz konusu müsabaka kapsamında daha başarılı sonuçların, daha karmaşık mimarilere sahip ön-eğitimli hazır modeller ile elde edilmesi sağlanabilmiştir. Çalışmamızda ise hazır modeller kullanılmamış, mevcut veri kümesine özgü sözcük yerleştirmeleri elde edilerek en uygun derin mimarilerde çözümün performansı farklı teknikler kullanılarak karşılaştırılmıştır.

5. Sonuçlar ve Gelecek Çalışmalar

Dijital metinlerden nefret söylemi tespiti çalışmaları son yıllarda araştırmacılar tarafından yoğun olarak ele alınmaktadır. Türkçe dili özelinde yapılan nefret söylemi tespiti çalışmaları diğer dillerde yapılan çalışmalar ile karşılaştırıldığında hala yeterli değildir. Bu çalışmada, Türkçe dilinde yapılan nefret söylemi tespitine yönelik çalışmalarının arttırılması ve hem sığ hem de derin makine öğrenmesi modellerinin bu tür çalışmalarda kullanımının başarılı bir şekilde sunulması amaçlanmıştır. Bu amaç doğrultusunda, en yaygın nefret söylemlerinden biri olan 'saldırgan' söylemleri, Türkçe tweet'lerden oluşan dengesiz bir veri kümesi üzerinden tespit edebilmek üzere sığ ve derin makine öğrenmesi modelleri kullanılmış, uygulama çeşitliliği ile karşılaştırmalı bir çalışma gerçekleştirilmiştir. Veri kümesi temsili için tf-idf ve sözcük yerleştirmelerinin kullanıldığı çalışmamızda, sınıflandırıcıları eğitmek için sığ makine öğrenmesi yöntemlerinden LR, SVM, Karar Ağaçları, MNB, RF, kNN ve derin makine öğrenmesi yöntemlerinden CNN, LSTM, BiLSTM, GRU ve BiGRU teknikleri farklı mimarilerde kullanılmıştır. Deneyler sonucunda en yüksek başarı0,74 F-skor ile BiLSTM tekniğinin kullanıldığı modelden elde edilmiştir. Geliştirilen derin öğrenme mimarileri ile elde edilen sonuçlar, literatürdeki yüksek başarılı, hazır ve karmaşık modeller ile karşılaştırılabilir düzeyde başarılı sonuçlara sahiptir.

Gelecek çalışmalarda, nefret söylemi içeren dengesiz veri kümelerini çeşitli derin makine öğrenmesi yöntemleri ile dengeli hale getirerek nefret söylemi tespiti için daha yüksek başarılı modellerin kurulması hedeflenmektedir.

Kaynakça

[1] Statista. Number of social network users in selected countries in 2017 and 2022 (in millions), 2018.

[2] Fortuna, P., Nunes, S. A survey on automatic detection of hate speech in text, ACM Computing Surveys, 2018, 51(4), pp. 1-30.

[3] T.D.K., Türk Dil Kurumu, Türk Tarih Kurumu Basımevi, 1954.

[4] Evans, M., Weber, A., Council of Europe Manuals - Human Rights in Culturally Diverse Societies (2 vols.), 2010

[5] Burnap, P., Williams, M.L. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making, Policy Internet, 2015, 7(2), pp. 223-242.

[6] Sahi, H., Kilic, Y., Saglam R.B. Automated Detection of Hate Speech towards Woman on Twitter, In 3rd International Conference on Computer Science and Engineering (UBMK 2018), 2018, pp. 533-536. IEEE.

[7] Dağaşan, T. Automatic hate speech detection on social media: Turkish tweets as an example. M.Sc. Thesis, Ankara Yildirim Beyazit University, 2019, Ankara.

[8] Hüsünbeyi, Z.M. Detecting hate speech in Turkish texts. M.Sc. Thesis, Boğaziçi University, 2020, İstanbul.

[9] Mayda, İ., Diri, B., Yıldız, T. Türkçe Tweetler üzerinde Makine Öğrenmesi ile Nefret Söylemi Tespiti, European Journal of Science and Technology, 2021, 24, pp. 328-334.

[10] Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., Çöltekin, Ç. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020), In 14th International Workshop on Semantic Evaluation, 2020, pp. 1425–1447.

[11] Charitidis, P., Doropoulos, S., Vologiannidis, S., Papastergiou, I., Karakeva, S. Towards countering hate speech against journalists on social media, Online Social Networks and Media, 2020, 17, pp. 100071.

[12] Guellil, I., Adeel, A., Azouaou, F., Chennoufi, S., Maafi, H., Hamitouche, T. Detecting hate speech against politicians in Arabic community on social media, International Journal of Web Information Systems, 2020, 16(3), pp. 295-313.

[13] Pitropakis, N., Kokot, K., Gkatzia, D., Ludwiniak, R., Mylonas, A., Kandias, M. Monitoring Users' Behavior: Anti-Immigration Speech Detection on Twitter, Machine Learning & Knowledge Extraction, 2020, 2(3), pp. 192-215.

[14] Pronoza, E., Panicheva, P., Koltsova, O., Rosso, P. Detecting ethnicity-targeted hate speech in Russian social media texts, Information Processing & Management, 2021, 58(6), pp. 102674.

[15] Jiang A., Yang X., Liu Y., Zubiaga A., SWSR: A Chinese dataset and lexicon for online sexism detection, Online Social Networks and Media , 2022, 27, pp. 100182.

[16] Chiril, P., Moriceau, V., Benamara, F., Mari, A., Origgi, G., Coulomb-Gully, M. An annotated corpus for sexism detection in French tweets, In 12th International Conference on Language Resources and Evaluation (LREC 2020), 2020, pp. 1397–1403.

[17] Parikh, P., Abburi, H., Badjatiya, P., Krishnan, R., Chhaya, N., Gupta, M., Varma, V. Multi-label categorization of accounts of sexism using a neural framework, In 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference (EMNLP-IJCNLP 2019), 2019, pp. 1642–1652.

[18] Wullach, T., Adler, A., Minkov, E. Character-level HyperNetworks for Hate Speech Detection, Expert Systems with Applications, 2022, 205, pp. 117571.

[19] Wu, X.-K., Zhao, T.-F., Lu, L., Chen, W.-N. Predicting the Hate: A GSTM Model based on COVID-19 Hate Speech Datasets, Information Processing & Management, 2022, 59, pp. 102998.

[20] Plaza-del-Arco, F.M., Molina-González, M.D., Ureña-López, L.A., Martín-Valdivia, M.T. Comparing pre-trained language models for Spanish hate speech detection, Expert Systems with Applications, 2021, 166, pp. 114120.

[21] García-Díaz, J.A., Jiménez-Zafra, S.M., García-Cumbreras, M.A., Valencia-García, R. Evaluating feature combination strategies

for hate-speech detection in Spanish using linguistic features and transformers, Complex & Intelligent Systems, 2022, pp. 1-22.

[22] Duwairi, R., Hayajneh, A., Quwaider, M. A Deep Learning Framework for Automatic Detection of Hate Speech Embedded in Arabic Tweets, Arabian Journal for Science and Engineering , 2021, 46, pp. 4001-4016.

[23] Al-Hassan, A., Al-Dossari, H. Detection of hate speech in Arabic tweets using deep learning, Multimedia Systems, 2021, pp. 1-12.

[24] Kalra, S., Agrawal, M., Sharma, Y. Detection of Threat Records by Analyzing the Tweets in Urdu Language Exploring Deep Learning Transformer, Forum for Information Retrieval Evaluation , 2021, pp. 813-819.

[25] Ali, R., Farooq, U., Arshad, U., Shahzad, W., Beg, M.O. Hate speech detection on Twitter using transfer learning, Computer Speech & Language , 2022, 74, pp. 101365.

[26] Karayiğit, H., Akdagli, A., Aci, Ç.İ. Homophobic and Hate Speech Detection Using Multilingual-BERT Model on Turkish Social Media, Information Technology and Control, 2022, 51, pp. 356-375.

[27] Toraman, Ç., Şahinuç, F., Yılmaz, E.H. Large-Scale Hate Speech Detection with Cross-Domain Transfer, In 13th Conference on Language Resources and Evaluation (LREC 2022), 2022, pp. 2215–2225.

[28] Aizawa, A. An information-theoretic perspective of tf–idf measures, Information Processing & Management, 2003, 39, pp. 45-65.

[29] Canbay, P., Sezer, E.A., Sezer, H. Detection of Stylometric Writeprint from the Turkish Texts, In 28th Signal Processing and Communications Applications Conference (SIU 2020), 2020, pp. 1-4. IEEE.

[30] Wang, S., Zhou, W., Jiang, C. A survey of word embeddings based on deep learning, Computing, 2020, 102, pp. 717–740.

[31] Mikolov, T., Chen, K., Corrado, G., Dean, J. Efficient estimation of word representations in vector space, In 1st International Conference on Learning Representations, Workshop Track Proceedings, International Conference on Learning Representations (ICLR 2013), arXiv:1301.3781v1, 2013.

[32] Pennington, J., Socher, R., Manning, C.D. GloVe: Global vectors for word representation, In 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), 2014, pp. 1532–1543.

[33] Bojanowski, P., Grave, E., Joulin, A., Mikolov, T. Enriching Word Vectors with Subword Information, Transactions of the association for computational linguistics, 2017, 5, pp. 135-146.

[34] Ekinci, E. Classification of Imbalanced Offensive Dataset – Sentence Generation for Minority Class with LSTM, Sakarya University Journal of Computer and Information Sciences, 2022, 5(1), pp. 121-133.

[35] Küçüksille, E.U., Ateş, N. Destek Vektör Makineleri ile Yaramaz Elektronik Postaların Filtrelenmesi, Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, 2013, 6(1).

[36] Soygazi, F., Mostafapour, V., Inan, E. TurkiS: A Turkish Sentiment Analyzer Using Domain-specific Automatic Labelled Dataset, International Journal of Intelligent Systems and Applications in Engineering, 2019, 7(2), pp. 99-103.

[37] Ganaie, M.A., Tanveer, M., Suganthan, P.N., Snasel, V. Oblique and rotation double random forest, Neural Networks, 2022, 153, pp. 496–517.

[38] Yakowitz, S. Nearest-Neighbour Methods for Time Series Analysis, Journal of Time Series Analysis , 2008, 8(2), pp. 235-247.

[39] Ekinci, E., Takcı, H., Alagöz, S. Poet Classification Using ANN and DNN, Electronic Letters on Science and Engineering, 2022, 18(1), pp. 10-20.

[40] Albawi, S., Mohammed, T.A., Al-Zawi, S. Understanding of a convolutional neural network, In 2017 International Conference on Engineering and Technology (ICET 2017), 2017, pp.1-6. IEEE.

[41] Siami-Namini, S., Tavakoli, N., Namin, A.S. The Performance of LSTM and BiLSTM in Forecasting Time Series, In 2019 IEEE International Conference on Big Data (Big Data 2019), 2019, pp. 3285-3292. IEEE.

[42] Ekinci, E., İlhan Omurca, S., Özbay, B. Comparative assessment of modeling deep learning networks for modeling ground-level ozone concentrations of pandemic lock-down period, Ecological Modelling, 2021, 457, pp. 109676.

[43] Graves, A., Schmidhuber, J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures, Neural Networks, 2005, 18(5-6), pp. 602-610.

[44] Zhang, X., Li, R., Dai, H., Liu, Y., Zhou, B., Wang, Z. Localization of myocardial infarction with multi-lead bidirectional gated recurrent unit neural network, IEEE Access, 2019, 7, pp. 161152-161166.

[45] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.