Avrupa Bilim ve Teknoloji Dergisi
Sayı 31 (Ek Sayı 1), S. 77-84, Aralık 2021

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
No. 31 (Supp. 1), pp. 77-84, December 2021

Research Article

Türkçe Sosyal Medya Yorumlarındaki Siber Zorbalığın Derin Öğrenme ile Tespiti

Gözde Nergiz[1*], Erdinç Avaroğlu[2]

[1*] Mersin Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Mersin, Türkiye, (ORCID: 0000-0001-5018-2031), gozdenergiz01@gmail.com
[2] Mersin Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Mersin, Türkiye (ORCID: 0000-0003-1976-2526), eavaroglu@gmail.com

(İlk Geliş Tarihi 27 Ağustos 2021 ve Kabul Tarihi 6 Aralık 2021)

(**DOI:** 10.31590/ejosat.987259)

ATIF/REFERENCE: Nergiz, G., Avaroğlu, E. (2021). Türkçe Sosyal Medya Yorumlarındaki Siber Zorbalığın Derin Öğrenme ile Tespiti. *Avrupa Bilim ve Teknoloji Dergisi*, (31), 77-84.

Öz

Siber zorbalık, internet teknolojisinin gelişimi ve sosyal ağlara erişim kolaylığı ile birlikte büyük bir problem haline dönüşmüştür. Bir kişi veya grup tarafından gerçekleştirilen siber zorbalık, başkalarını taciz etmek için bilgi ve iletişim teknolojilerinin kullanılması anlamına gelir. İntihar ile sonuçlanan siber zorbalık vakaları siber zorbalık tespitini önemli hale getirmiştir. Bu çalışmada günümüzde yaygın olarak kullanılan Twitter, Instagram ve Youtube sosyal ağlarından toplanan Türkçe yorumlar üzerinde siber zorbalık tespiti yapılmıştır. Derin öğrenme tabanlı kelime gömme modelleri kullanılarak sınıflandırma modelleri oluşturulup başarı oranları karşılaştırılmıştır. % 93,15 başarı oranı ile en başarılı sonucu veren model Fasttext modeli olmuştur. Fasttext modeli ile LSTM sinir ağı kullanılarak sosyal medya yorumlarının sınıflandırılması sağlanmıştır.

Anahtar Kelimeler: Sosyal Ağlar, Derin Öğrenme, Fasttext, LSTM, Siber Zorbalık.

Detection of Cyberbullying in Turkish Social Media Comments with Deep Learning

Abstract

Cyberbullying has become a big problem with the development of internet technology and ease of access to social networks. Cyberbullying is by a person or group refers to the use of information and communication technologies to harass others. Cyberbullying cases resulting in suicide have made the detection of cyberbullying important. In this study, cyberbullying was detected on Turkish comments collected from Twitter, Instagram and Youtube social networks, which are widely used today. Classification models were created using deep learning-based word embedding models and success rates were compared. The model that gave the most successful result with a success rate of 93.15% was the Fasttext model. Classification of social media comments is provided by using fasttext model and LSTM neural network.

Keywords: Social Networks, Deep Learning, Fasttext, LSTM, Cyberbullying.

[*] Sorumlu Yazar: gozdenergiz01@gmail.com

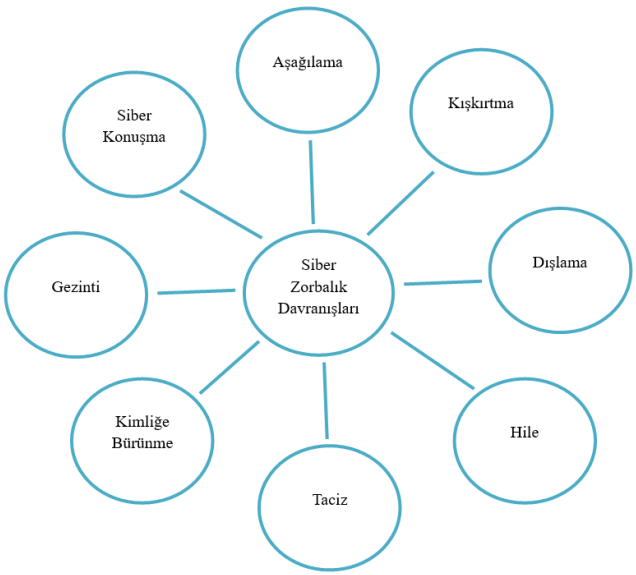
1. Giriş

Sosyal medya platformları, insanların birbirleriyle iletişim kurmaları, güncel olaylardan haberdar olmaları ve sosyalleşmeleri açısından faydalı araçlardır. Fakat kötü niyetli kullanıcıların elinde bazen tehlikeli bir araç haline dönüşebilmektedir. Dalga geçmek, tehdit etmek, küfür etmek, taciz etmek gibi siber zorbalık davranışları sosyal medyanın kötü niyetli kullanımına örnek olarak gösterilebilir [1].

Siber zorbalık, bilgisayarlar, akıllı telefonlar gibi çeşitli teknolojik araçları kullanarak bir kişiye veya gruba zarar vermek amacıyla kasıtlı ve sürekli olarak gerçekleştirilen davranışlardır [2]. Siber zorbalık, sosyal ağlar üzerinde 8 farklı yolla gerçekleştirilebilmektedir.

- **Aşağılama:** Bir kişi hakkında internette asılsız söylentiler yayarak onu küçük düşürmek.
- **Dışlama:** Bir kişiyi çevrimiçi gruplardan kasıtlı olarak dışlamak.
- **Kışkırtma:** Bir kişiyi çevrimiçi tartışmalara teşvik etmek için kırıcı, kaba, saldırgan bir dil içeren mesajlarla kışkırtmak.
- **Hile:** Bir kişiyi kişisel bilgilerini veya fotoğraflarını almak için kandırmak ve bu bilgileri sosyal medyada paylaşmak.
- **Taciz:** Hedef kişiye sürekli olarak saldırgan, kötü niyetli mesajlar göndermek veya bu mesajları çevrimiçi olarak paylaşmak.
- **Kimliğe Bürünme:** Hedef kişi adına sahte profil oluşturarak veya hesaplarını hackleyerek onun itibarını zedeleyecek zararlı paylaşımlarda bulunmak.
- **Gezinti:** Bir kişiye ait özel bilgileri, fotoğrafları kişiden habersiz bir şekilde onu aşağılamak, utandırmak için sosyal ağlarda paylaşmak.
- **Siber Konuşma:** Hedef kişiye şiddet uygulayacağını ifade eden, tehdit içerikli, korkutucu mesajlar göndermek.

Siber zorbalık davranışları Şekil 1'de gösterilmektedir.



Şekil 1. Siber Zorbalık Davranışları

Sosyal medya kullanımının yaygınlaşması ve kullanıcı yaşının giderek düşmesi ile birlikte özellikle çocuklar ve gençler Twitter, Instagram, Youtube gibi sosyal medya platformlarında siber zorbalığa çok sık maruz kalabilmekte ve siber zorba olma eğilimi göstermektedirler [3]. İnsanlarla dalga geçmek, eğlenmek, popüler olmak, siber zorbalığa maruz kaldığı için intikam almak vb. sebeplerden dolayı çocuklar siber zorbalık yapabilmektedirler.

Siber zorbalıkta, klasik zorbalıkta olduğu gibi zorba ile mağdur arasında fiziksel bir temas yoktur. Fakat mağdur depresyon, özgüven kaybı, insanlara güvenememe, okul derslerinde başarısızlık, intihar eğilimi gösterme gibi psikolojik problemler yaşayabilmektedir.

Siber zorbalığın sanal ortamda anonim olarak gerçekleştirilebiliyor olması siber zorbayı tespit etmede ve siber zorbalığı önlemede büyük bir dezavantajdır. Çünkü suçu işleyen belli olmadığından bu suçun önüne geçmek için caydırıcı cezalar vermekte mümkün olmamaktadır. Böylece zorba farklı isimlerle gerçek kimliğini gizleyerek insanlara zarar vermeye rahatlıkla devam edebilmektedir. Zorba, insanların birbirleriyle olan iletişimlerine, sosyal ilişkilerine, psikolojik durumlarına zarar verebilmektedir. Bu durumun önüne geçebilmek zararları en aza indirebilmek için siber zorbalığın tespit edilmesi gerekir. Bu çalışmada, derin öğrenme yöntemleri ile sosyal ağlardaki siber zorbalık içeren Türkçe yorumları tespit etmek amaçlanmıştır.

Derin öğrenme, insan müdahalesine ihtiyaç duymadan kendi kendine öğrenme yeteneğine sahip bir makine öğrenmesi tekniğidir. Derin öğrenme verilerden özellik çıkarımını otomatik olarak gerçekleştirmesi ve yüksek boyutlu verilerde daha iyi performans göstermesi gibi avantajları sayesinde makine öğrenmesinin önüne geçmektedir.

Siber zorbalık tespiti işlemi veri kümesinin oluşturulması, veri ön işleme, derin öğrenme yöntemleri ile modeller oluşturma, sınıflandırma ve sonuç adımlarından oluşmaktadır.

Veri kümesi, Twitter, Instagram ve Youtube sosyal medya platformlarından Türkçe yorumlar toplanarak oluşturulmuştur. Pozitif ve negatif olmak üzere veri kümesi iki kategoriye ayrılmıştır. Pozitif, siber zorbalık içermeyen verileri negatif ise siber zorbalık içeren verileri ifade etmektedir. Veri kümesi her bir kategori için 90.000 olmak üzere toplamda 180.000 yorumdan oluşmaktadır.

Sosyal ağlar, kullanıcılar yorum paylaşırken belirli bir karakter uzunluğuna kadar paylaşıma izin vermektedir. Bu nedenle kullanıcılar bazı kelimeleri kısaltarak paylaşım yapabilmektedir. Ayrıca kullanıcılar yazım kurallarına dikkat etmeden günlük hayatta kullandıkları dil ile paylaşım yapabilmektedir. Ancak bu tür gürültülü veriler içeren metinler üzerinde sınıflandırma işlemi gerçekleştirmek zordur. Gürültülü verileri temizlemek ve verileri uygun formata getirmek için doğal dil işleme yöntemleri kullanılmaktadır. Veri setindeki gürültüler ön işleme adımları uygulanarak temizlenmiştir.

Verilere ön işleme uygulandıktan sonra kelimelerin sayısal ifadelerini elde etmek için kelime gömme yöntemleri kullanılmıştır. Fasttext, Word2Vec ve Doc2Vec kelime gömme yöntemleri ile 3 model oluşturulmuştur. Geliştirilen bu derin öğrenme tabanlı modeller üzerinde LSTM ile sınıflandırma işlemi gerçekleştirilmiştir.

Çalışmanın bölümleri aşağıdaki gibi belirlenmiştir. İkinci bölümde literatürde bulunan siber zorbalık tespiti çalışmaları,

üçüncü bölümde materyal ve metot, dördüncü bölümde araştırma sonuçları ve tartışma, son bölümde ise sonuçlar açıklanmıştır.

2. İlgili Çalışmalar

Son zamanlarda siber zorbalık suçunun artışı ile birlikte siber zorbalığın tespitine yönelik çalışmaların sayısında da artış görülmektedir. Bu bölümde literatürde bulunan siber zorbalık tespiti ile ilgili çalışmalar hakkında bilgiler verilmiştir.

Zhang ve arkadaşları [4] yaptıkları çalışmada siber zorbalığı tespit etmek için telaffuz tabanlı evrişimli sinir ağını önermişlerdir. Çalışmada kullanılan Twitter ve Formspring.me veri setlerine önce ön işleme adımları sonra kelimeden telaffuza dönüştürme işlemi uygulanmıştır. Kelimelerin fonetik temsilleri elde edildikten sonra bu temsiller evrişimli sinir ağlarında özellik olarak kullanılmıştır. Çalışmanın sonucunda, önerilen yaklaşımın veri setlerinde yüksek performans sağladığı gözlemlenmiştir.

Soni ve Singh [5] yorum yapma davranışının zamansal dinamiklerini modellemişlerdir. Sundukları çalışmada siber zorbalığın tek bir yoruma bağlı olmadığını, çevrimiçi bir yorum dizisinde veya oturumda bireyler arasında tekrarlanan etkileşimlerin birleşik etkisi olduğunu düşündüklerini ifade ederek farklı bir yaklaşım önermişlerdir. Siber zorbalık içeren ve içermeyen sosyal medya oturumları arasındaki zamansal farklar hesaplanıp özellik olarak kullanılmıştır. Sonuç olarak, önerilen yaklaşımın siber zorbalık tespitinde performans artışı sağladığı görülmüştür.

Özel ve arkadaşları [6] sundukları çalışmada Twitter ve Instagram platformlarından manuel olarak topladıkları 900 tane Türkçe mesajdan oluşan bir veri seti kullanmışlardır. Destek Vektör Makineleri (SVM), Karar Ağacı (C4.5), Naive Bayes Multinomial (NBM), K En Yakın Komşu (KNN) sınıflandırıcılarından oluşan makine öğrenmesi tekniklerini kullanarak siber zorbalığı tespit etmişlerdir. Sınıflandırma doğruluğunu artırmak için bilgi kazanımı ve ki-kare özellik seçim yöntemlerini uygulamışlardır. Özellik seçim yöntemleri uygulandıktan sonra sınıflandırma doğruluğunu % 84'e kadar artırmışlardır.

Bozyiğit ve arkadaşları [7] çalışmalarında 3000 adet Türkçe Twitter mesajlarını içeren bir veri seti oluşturmuşlardır. Veri setindeki Türkçe yazım kurallarına uymayan sanal zorbalık ifadelerini düzeltmek için bir algoritma geliştirmişlerdir. Veri setine geliştirdikleri algoritma ve ön işleme adımlarını uyguladıktan sonra makine öğrenmesi algoritmalarını kullanarak sınıflandırma işlemini gerçekleştirmişlerdir. Sınıflandırma sonucunda Destek Vektör Makineleri, Multinomial Naive Bayes, ve K En Yakın Komşu yöntemlerinin en iyi sınıflandırma performansına sahip olduğu gözlemlenmiştir.

Ön ve Yeniterzi [8] çalışmalarında Bozyiğit ve arkadaşlarının [7] paylaştığı 3000 Türkçe Tweet içeren veri seti üzerinde derin öğrenme teknikleri kullanarak siber zorbalık tespiti gerçekleştirmişlerdir. Yaptıkları çalışmada önceden eğitilmiş 3 farklı kelime vektör temsillerini kullanarak bunların siber zorbalık tespitine etkisini incelemişlerdir. Kelime vektör temsilleri ile Evrişimsel Sinir Ağı Modelleri (CNN) kullanarak 3 farklı model oluşturmuşlardır. Bu 3 modelin yanı sıra rastgele kelime vektör temsilleri kullanılarak oluşturulan bir CNN modelini de kullanmışlardır. Sınıflandırma sonucunda 0,937 F1 skoru elde eden rastgele kelime temsilleri ile oluşturulan CNN modeli en yüksek başarı oranını yakalamıştır.

Balakrishnan ve arkadaşları [9] siber zorbalık tespitini iyileştirmek için Twitter kullanıcılarının kişilikleri, duyguları ve duygusallıkları dahil olmak üzere psikolojik özelliklerinden yararlanan yeni bir siber zorbalık tespit modeli önermişlerdir. Kullanıcı kişiliklerinin belirlenmesinde Big Five(dışadönüklük, uyumluluk, dürüstlük, nevrotiklik ve açıklık) ve Dark Triad(makyavelcilik, psikopati ve narsisizm) modellerini kullanmışlardır. Temel algoritma olarak Twitter tabanlı özellikleri(metin, kullanıcı ve ağ tabanlı özellikler) tercih etmişlerdir. Temel algoritmanın yanı sıra psikolojik özellikleri dahil eden farklı modeller oluşturmuşlardır. 5453 tweet içeren veri kümesini zorba, saldırgan, spam gönderen ve normal olmak üzere dört kategoriye göre Naïve Bayes, Random Forest ve J48 algoritmalarını kullanarak sınıflandırmışlardır. Sınıflandırma sonuçları değerlendirildiğinde kişilik özellikleri ve duygusallık kullanıldığında siber zorbalık tespitinin arttığını ancak duygu kullanıldığında bir artış olmadığı gözlemlenmiştir. Kişilik özellikleri üzerinde yapılan analizde ise siber zorbalığı tespit etmede diğer özelliklere kıyasla dışadönüklük, uyumluluk, nevrotiklik ve psikopati özelliklerinin daha etkili olduğu görülmüştür.

Laxmi ve arkadaşları [10] yaptıkları çalışmada Endonezyadaki siber zorbalık eylemlerini önlemek amacıyla Endonezya dilindeki 1425 tweet'ten oluşan veri kümesini kullanmışlardır. Veri kümesine ön işleme adımları uyguladıktan sonra Doc2Vec yöntemini kullanarak kelimelerin vektörel temsillerini elde etmişlerdir. CNN, SVM ve RF sınıflandırıcılarını kullanarak dengeli ve dengesiz veri kümeleri üzerinde iki tür deney gerçekleştirmişlerdir. Sınıflandırma sonucunda sınıflandırıcıların dengeli bir veri kümesinde siber zorbalık içeren tweet'leri daha iyi tespit edebileceği gözlemlenmiştir. CNN ve Doc2Vec modeli % 65,08 F1 puanı ile diğer sınıflandırma yöntemlerinden daha başarılı olmuştur.

Alsubait ve Alfageh [11] çalışmalarında siber zorbalığı tespit etmek için Arapça Youtube yorumlarından oluşan herkese açık hazır bir veri seti kullanmışlardır. Sayım vektörü ve Tfidf vektörü yöntemleri kullanılarak özellik çıkarımı yapılmıştır. Makine öğrenmesi algoritmalarından Multinomial Naïve Bayes, Complement Naïve Bayes ve Logistic Regression kullanılarak sınıflandırma işlemi gerçekleştirilmiştir. Sonuç olarak, sayım vektörü kullanıldığında % 78,6 F1 puanı ile Logistic Regression modeli diğer modellerden daha başarılı olmuştur. Tfidf vektörü kullanıldığında ise Complement Naïve Bayes modeli % 78,6 F1 puanı ile daha yüksek performans göstermiştir.

Razvani ve Beheshti [12] 2021 yılında yaptıkları çalışmada dikkat temelli bağlam destekli bir siber zorbalık algılama yaklaşımı önermişlerdir. Çalışmada iki farklı veri seti kullanmışlardır. Instagram ve Twitter platformlarından elde edilen veri kümeleri beğenileri, arkadaşlık grafiklerini, görüntüleri ve yorumları içermektedir. Metinsel özelliklerin yanında görüntülerden çıkarılan özellikler, sosyal ağ meta verileri ve dış bilgi tabanları gibi bağlamsal özellikleri de ekleyerek metinsel özellikleri zenginleştiren bir yöntem sunmuşlardır. Geleneksel sinir ağı, LSTM ve ALBERT ağının kullanıldığı sınıflandırıcı modelleri oluşturmuşlardır. ALBERT ağının kullanıldığı modeller metinsel ve bağlamsal özellikleri birleştiren önerilen modellerdir. Sınıflandırma işlemi gerçekleştirildikten sonra önerilen yöntemin diğer yöntemlerden daha iyi performans sağladığı görülmüştür.

Luo ve arkadaşları [13] siber zorbalığı tespit etmek için BİGRU katmanı, dikkat mekanizması katmanı, CNN katmanı, tam bağlantılı katman ve sınıflandırma katmanından oluşan bir BİGRU-CNN duygu sınıflandırma modeli(GCA: BiGRU+CNN+ATTENTION) önermişlerdir. Modeli eğitmek ve test etmek için Kaggle platformundan alınan veri setini ve sosyal ağlardan toplanan emoji veri setini kullanmışlardır. Kelimeler arasındaki ilişkiyi öğrenmek için Glove kelime gömme yöntemini tercih etmişlerdir. GCA modelinde bulunan dikkat mekanizması katmanı kelimelerin dikkat ağırlıklarını hesaplamak ve bu ağırlıklardan özel anlamlı kelimeleri seçmek için kullanılmıştır. GCA modeli ile birlikte MLP, CNN, BİGRU, BİLSTM-ATTENTION, BİGRU-ATTENTION modelleri de eğitilerek sınıflandırma doğruluk oranları karşılaştırılmıştır. Önerilen GCA modeli % 91,07 doğruluk oranı ile diğer modellerden daha başarılı olmuştur.

Literatürdeki çalışmalar incelendiğinde bazı çalışmalar siber zorbalık tespitinin özellik çıkarımı aşamasında farklı yöntemler kullanırken sadece birkaç çalışma kelime gömme yöntemlerini kullanmıştır. Ayrıca, İngilizce dilinde yapılan çalışma sayısı oldukça fazla iken Türkçe dilinde hala yeteri kadar çalışma bulunmamaktadır. Literatüre katkıda bulunmak amacıyla bu çalışmada sosyal ağlardan Türkçe metinler toplanarak siber zorbalık tespiti için Türkçe dilinde en büyük veri kümesi oluşturulmuştur. Oluşturulan veri kümesi üzerinde kelime gömme yöntemleri uygulanarak sınıflandırma başarısına olan etkisi incelenmiştir.

3. Materyal ve Metot

3.1. Veri Kümesi

Veri kümesi Twitter, Instagram, Youtube sosyal ağlarından elde edilen Türkçe yorumlardan oluşturulmuştur. Yorumlar toplanırken özel hayatın gizliliğini ihlal etmemek adına herkese açık sayfalar ve hashtagler kullanılmıştır. Hashtagler ve herkese açık sayfalar belirlenirken siber zorbalık içerebilecek yorumlar bulundurmasına dikkat edilmiştir. Yorumlar toplanırken Python programlama dili ve Selenium aracı tercih edilmiştir.

Selenium, bir web tarayıcı otomasyon aracıdır. Öncelikli olarak web uygulamalarını test etmek için kullanılmaktadır. Fakat seçilen bir tarayıcıyı açmak, butonlara tıklamak, formlara bilgi girmek, web sayfalarında belirli bilgileri aramak gibi insanın yapacağı görevleri de gerçekleştirebilmektedir. Selenium yardımıyla belirlenen hashtagler veya herkese açık olan sayfaların isimleri aratılıp çıkan sonuçlar altındaki yorum bilgileri elde edilerek veri kümesi oluşturulmuştur. Veri setine ait kelime bulutu gösterimi Şekil 2'de verilmiştir.

Şekil 2. Veri Seti Kelime Bulutu Gösterimi

Veri kümesinde Twitter platformundan pozitif ve nagatif kategorilerin her biri için 42.178 toplamda 84.356, Instagram platformundan her bir kategori için 21.340 toplamda 42.680, Youtube platformundan her bir kategori için 26.482 toplamda 52.964 adet yorum bulunmaktadır. Veri kümesinin % 80'i eğitim adımında % 20'si ise test adımında kullanılmıştır.

3.2. Ön İşleme

Sosyal medya yorumlarında kelimelerin yanlış yazılması, uzatılarak veya kısaltılarak yazılması, özel ifadelerin kullanılması yorumların gürültülü veriler içerdiğini göstermektedir. Verileri Türkçe yazım kurallarına uygun hale getirmek, gürültülü verilerden temizlemek için ön işleme adımları uygulanmıştır.

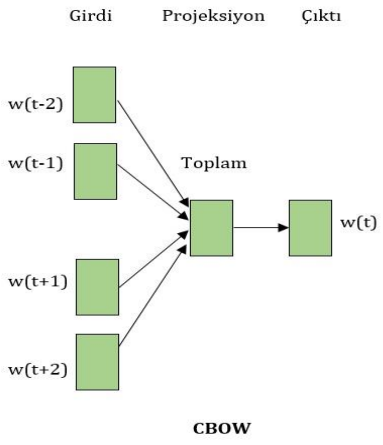
Veri setinden @ ile başlayan kullanıcı adları, # ile başlayan hashtagler, web adresleri, gereksiz boşluklar, emojiler, noktalama işaretleri ve rakamlar temizlenmiştir. Tüm harfler küçük harflere dönüştürülmüştür. "ve", "veya", "ile", "ki", "de" gibi cümlenin anlamını etkilemeyen kelimeler veri setinden temizlenmiştir. Uzatılarak yazılan yorumlar içerisinde ikiden fazla tekrar eden harfler Türkçe yazım kuralına uygun bir şekilde düzeltilmiştir. Türkçe dilinin yapısı gereği kök bulma işlemi kelimelerin anlamını değiştireceğinden uygulanmamıştır. Ön işleme aşamaları Türkçe dili için geliştirilmiş olan Zemberek doğal dil işleme kütüphanesi yardımıyla gerçekleştirilmiştir. Veriler üzerinde ön işleme adımları uygulanarak kelimelerin sayısal temsillerini elde etmek için uygun formata getirilmiştir.

3.3. Kelime Gömme Modelleri

Kelime gömme, kelimeleri sayısal vektörlere dönüştürmek için kullanılan bir yöntemdir. Benzer anlamlara sahip kelimelerin vektörleri birbirine yakındır. Her kelime bir vektör ile temsil edilmekte ve vektörler sinir ağları yardımıyla öğrenilmektedir. Bu çalışmada Word2Vec, Fasttext ve Doc2Vec modelleri kullanılmıştır.

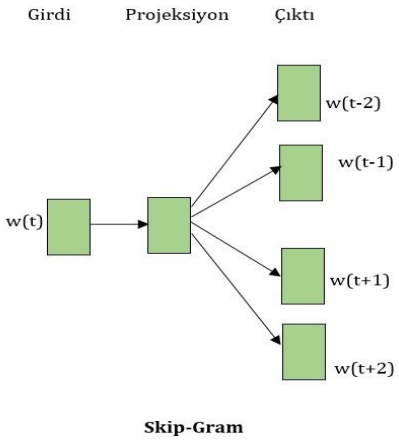
Word2Vec, kelimelerin vektörel temsillerini elde edip kelimeler arasındaki mesafeyi hesaplayarak aralarındaki anlamsal benzerliği tespit etmek için geliştirilmiş bir yöntemdir [14]. Temelinde yapay sinir ağları barındırmaktadır. CBOW ve Skip-Gram olmak üzere iki öğrenme modeli içermektedir.

CBOW, hedef kelimenin çevresindeki komşu kelimeleri girdi olarak almakta ve bu kelimelerden hedef kelimeyi tahmin etmeye çalışmaktadır. CBOW modeli Şekil 3'de gösterilmiştir.



Şekil 3. CBOW Modeli [15]

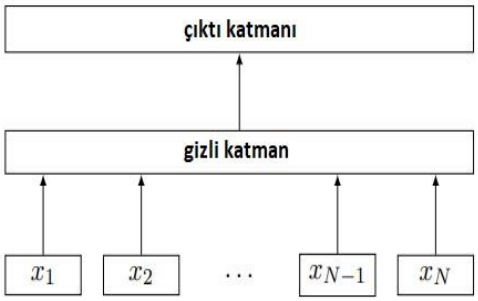
Skip-Gram ise hedef kelimeyi girdi olarak almakta ve çevresindeki komşu kelimeleri tahmin etmeye çalışmaktadır. Skip-Gram modeli Şekil 4'de gösterilmiştir.



Şekil 4. Skip-Gram Modeli [15]

CBOW, sık kullanılan kelimeleri daha iyi temsil etmektedir ve Skip-Gram'a göre daha hızlıdır. Skip-Gram ise nadir kullanılan kelimeleri daha iyi temsil etmekte ve veri miktarının az olduğu durumlarda daha iyi performans göstermektedir.

Fasttext, kelimelerin karakter n-gramlarını kullanarak sayısal vektörler oluşturan Word2Vec tabanlı bir yöntemdir [16]. Word2vec yönteminde kelimelerin vektörleri oluşturulup yapay sinir ağlarına girdi olarak verilirken Fasttext yönteminde her kelimenin karakter n-gramlarının toplamından sayısal vektörler elde edilerek yapay sinir ağlarına girdi olarak verilmektedir. Bu nedenle Fasttext, veri kümesinde bulunmayan veya cümlelerde nadir geçen kelimelerin vektör temsillerini oluşturmada daha başarılıdır. Fasttext modeli Şekil 5'de gösterilmiştir.



Şekil 5. Fasttext Modeli [17]

Doc2Vec, Word2Vec mimarisine paragraf vektörlerinin eklenmesi ile oluşturulmuş bir yöntemdir. Doc2Vec ile üretilen vektörler belgeler arasındaki benzerlikleri bulmak için kullanılabilmektedir. Belgenin uzunluğunun bir önemi olmadan sayısal vektörler oluşturabilmektedir [18]. PV-DM ve PV-DBOW olmak üzere iki öğrenme modeline sahiptir.

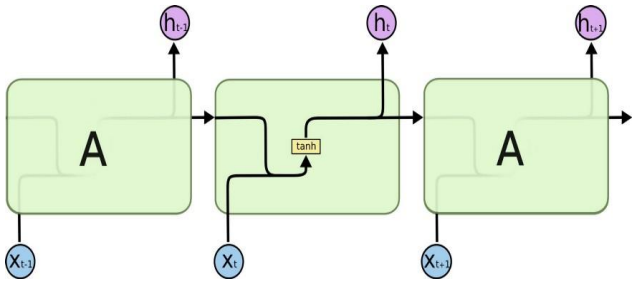
PV-DM, Word2Vec mimarisindeki CBOW modeline karşılık gelmektedir. CBOW modelinde tahmin yapmak için kelimelerin vektörlerine ihtiyaç duyulurken, PV-DM modelinde kelimelerin vektörlerinin yanı sıra paragraf vektörüne de ihtiyaç duyulmaktadır.

PV-DBOW, PV-DM'den farklı olarak bir sonraki kelimeyi tahmin etmek yerine paragraf vektörünü kullanarak belgedeki kelimeleri sınıflandırmaktadır. Word2Vec mimarisindeki Skip-Gram modeline karşılık gelmektedir.

Bu çalışmada kelimelerin vektör temsillerini elde etmek için Word2Vec, Fasttext ve Doc2Vec modelleri oluşturulmuştur. Word2Vec modelini oluşturmak için Skip-Gram, Fasttext için n-gram, Doc2Vec için PV-DBOW yöntemi kullanılmıştır. Modellerin oluşturulması için gerekli en uygun parametre değerleri belirlenmiştir. Vektör boyutu 100, pencere boyutu 5, minimum kelime sayısı 5, workers değeri 400 olarak belirlenmiştir. 3 farklı modelin ürettiği vektör matrisleri sınıflandırma aşamasında oluşturulacak modelin embedding katmanında kullanılacaktır.

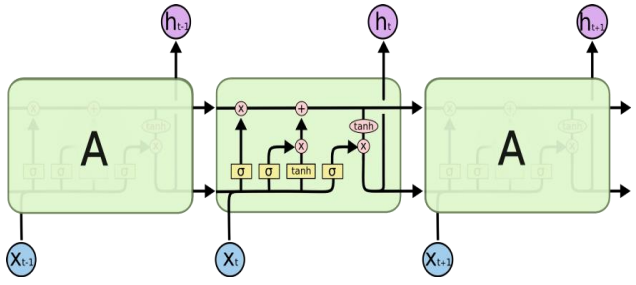
3.4. Uzun Kısa Süreli Bellek (LSTM)

Tekrarlayan sinir ağları (RNN), sıralı bilgileri işlemek için kullanılan yapay sinir ağlarıdır. İleri beslemeli sinir ağlarında girdiler ve çıktılar birbirinden bağımsızken RNN'lerde bir önceki adımın çıktısı mevcut adımın girdisi ile birlikte beslenir [19]. İleri beslemeli sinir ağlarında sonraki adımda tahmin yapabilmek için geçmişteki bilgileri hatırlayabilme özelliği yoktur. RNN'ler bu probleme çözüm bulmak için geliştirilmiş sinir ağlarıdır. RNN'lerin kendilerine ait hafızaları vardır ve bu hafızalar sayesinde geçmişteki bilgileri hatırlayabilmektedirler. RNN'ler kısa vadeli hafızalara sahiptirler. Bu nedenle uzun zaman önceki bilgileri hatırlamakta zorluk çekmektedirler. RNN mimarisi Şekil 6'da gösterilmiştir.



Şekil 6. RNN Mimarisi [20]

Uzun kısa süreli bellek (LSTM), uzun vadeli hafızaları sayesinde RNN'lerin hafıza problemine çözüm olmuştur. RNN'den farklı olarak LSTM'lerde uzun zaman önceki bilgileri hatırlamak için kullanılan hafıza hücreleri bulunmaktadır. LSTM mimarisi giriş kapısı, çıkış kapısı, unutma kapısı ve hafıza hücreleri olmak üzere 4 katmana sahiptir [21]. LSTM mimarisi Şekil 7'de gösterilmiştir.



Şekil 7. LSTM Mimarisi [20]

Sınıflandırma yapabilmek için Word2Vec ile elde edilen vektör matrisleri kullanılarak Sınıflandırma Modeli 1, Fasttext ile Sınıflandırma Modeli 2 ve Doc2Vec ile Sınıflandırma Modeli 3 isimli sınıflandırma modelleri oluşturulmuştur. Sınıflandırma modelleri giriş katmanı, gizli katman, dense katmanı ve çıkış katmanından oluşmaktadır.

Vektör matrisleri yardımıyla oluşturulan embedding katmanı giriş katmanı olarak kullanılmıştır.

Gizli katmanda Sınıflandırma Modeli 1, Sınıflandırma Modeli 2 ve Sınıflandırma Modeli 3 isimli modellere LSTM sinir ağı parametre olarak verilmiştir. Sigmoid fonksiyonu aktivasyon fonksiyonu olarak kullanılmıştır. Dropout katmanı kullanılarak modellerin aşırı öğrenme probleminin önüne geçilmiştir.

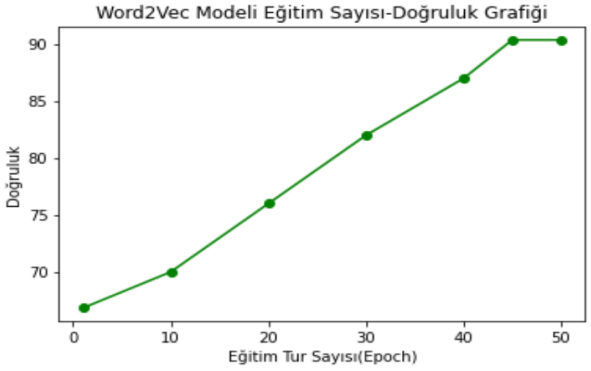
Dense katmanına LSTM ile elde edilen çıktı verilmiştir. Aktivasyon fonksiyonu olarak relu fonksiyonu kullanılmıştır.

Çıkış katmanında softmax aktivasyon fonksiyonu kullanılmıştır. Çıkış katmanı pozitif ve negatif sınıfları için 2 nörona sahiptir. Bu nöronlar sınıflandırma işlemi sonucunda her iki sınıf için de olasılıksal olarak değerler elde etmeyi sağlar.

4. Araştırma Sonuçları ve Tartışma

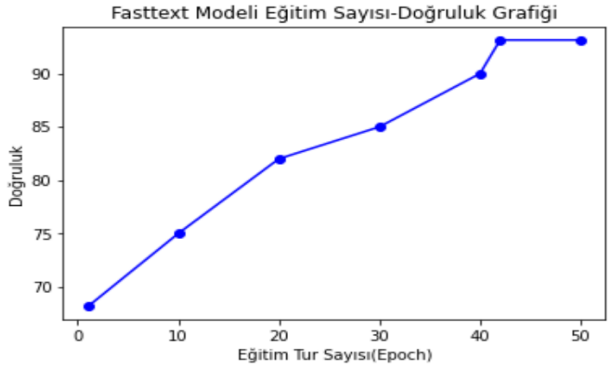
3 farklı sosyal medya platformundan toplanan 180.000 yorumdan oluşan veri kümesinin % 80'i eğitim % 20'si test işlemi için kullanılmıştır. Oluşturulan sınıflandırma modelleri üzerinde eğitim ve test aşamaları uygulanmıştır. Modellerin eğitimi sırasında veriler belirli bölümler halinde eğitilmiştir. Eğitim işlemi tek bir adım yerine başarı oranına göre her seferinde ağırlıklar güncellenerek gerçekleştirilen birden fazla eğitim adımından oluşur. Her bir eğitim adımı epoch olarak isimlendirilmektedir. Sınıflandırma modellerinin eğitimi sırasında epoch değeri başlangıçta 10 olarak belirlenmiştir ve en iyi başarı oranı elde edilene kadar epoch değeri artırılmıştır.

Word2Vec ile tasarlanan sınıflandırma modelinin 45. eğitim adımından sonra başarı oranının artmadığı gözlemlenmiştir. Şekil 8'de modele ait epoch değerleri ve doğruluk değerleri gösterilmiştir.



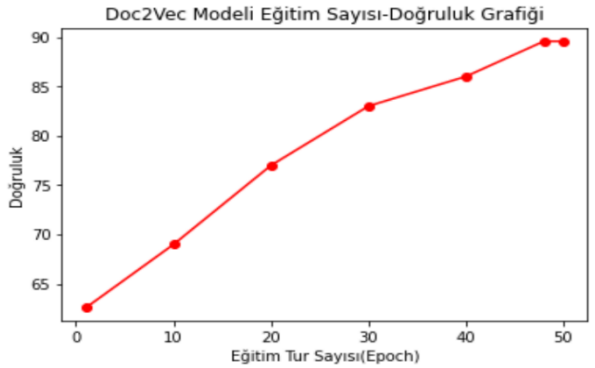
Şekil 8. Word2Vec Modeli Doğruluk Grafiği

Fasttext ile oluşturulan 2. Sınıflandırma modeli eğitilirken başarı oranı 42. adımdan sonra sabit kalmıştır. Modelin epoch değerlerine göre oluşan başarı oranları Şekil 9'da gösterilmiştir.



Şekil 9. Fasttext Modeli Doğruluk Grafiği

Doc2Vec kullanılarak tasarlanan sınıflandırma modelinin başarı oranının 48. eğitim adımından sonra sabit kaldığı ve eğitim işlemini tamamladığı gözlemlenmiştir. Modelin epoch değerleri ve başarı oranları Şekil 10'da gösterilmiştir.



Şekil 10. Doc2Vec Modeli Doğruluk Grafiği

Loss (kayıp) ve accuracy (doğruluk) değerleri, modellerin başarı oranlarını ölçmeyi sağlayan değerlerdir. Sınıflandırma modellerinin başarı oranları bu değerler dikkate alınarak değerlendirilmektedir.

Loss değeri, modelin tahmin ettiği değerin gerçek değerden ne kadar farklı olduğunu göstermektedir [22]. Bu değerin 0'a yakın bir değer olması modelin başarılı olduğunu, 0'dan büyük bir değer ise modelin başarız olduğunu, 0'a eşit ise modelin aşırı öğrenerek ezberleme yaptığını göstermektedir.

Accuracy değeri, doğru tahmin edilen değerlerin oranını göstermektedir [22]. Eğer bir modelin loss değeri düşük, accuracy değeri büyük ise o model başarılı bir modeldir.

Bu çalışmada accuracy ve loss değerlerini hesaplamak için Binary Cross-Entropy loss fonksiyonu kullanılmıştır. Binary Cross-Entropy loss fonksiyonu, ikili sınıflandırmada kullanılan gerçek değer ile tahmin edilen değer arasındaki çapraz entropi kaybını hesaplayan bir fonksiyondur. Bu fonksiyona ait matematiksel formül denklem (1)'de gösterilmiştir.

$$L(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}(y_i * \log(\hat{y}_i) + (1 - y_i) * \log(1 - \hat{y}_i)) \quad (1)$$

Denklemdeki y gerçek değeri, ŷ tahmin edilen değeri, N ise toplam veri sayısını temsil etmektedir.

3 sınıflandırma modeline ait sonuçlar Tablo 1'de verilmiştir.

Tablo 1. Sınıflandırma Modellerinin Sonuçları

	Sınıflandırma Modelleri		
	1	**2**	**3**
Kelime Gömme Modeli	Word2Vec	Fasttext	Doc2Vec
Eğitim Verisi	144.000	144.000	144.000
Test Verisi	36.000	36.000	36.000
Eğitim Adımı/ Epoch Değeri	50/45	50/42	50/48

Accuracy (Doğruluk) Değeri	0,9138	0,9315	0,8956
Loss (Kayıp) Değeri	0,1803	0,1642	0,1927

Tablo 1'de gösterilen sonuçlar epoch değeri arttırılarak gerçekleştirilen eğitim ve test işlemleri sonucunda elde edilen en iyi başarı oranlarıdır. Sonuçlar incelendiğinde Fasttext yönteminin kullanıldığı Sınıflandırma Modeli 2 isimli model % 93,15 başarı oranı sağlayarak en başarılı model olmuştur. Bu model eğitim adımında süre açısından diğer modellere kıyasla düşük performans göstermiş olsa da kelimelerin vektör temsillerini oluştururken n-gram yöntemini kullandığı için daha yüksek bir başarı oranına sahip olmuştur. Genel olarak sınıflandırma modelleri değerlendirildiğinde başarı oranına olumlu etki eden faktörler veri kümesinin pozitif ve negatif sınıflar için eşit sayıda oluşturulması, ön işleme adımları uygulanarak verilerin uygun formata getirilmesi ve modellerin oluşturulma aşamasında en uygun parametre değerlerinin seçilmesidir. Başarı oranının daha fazla artmamasına sebep olan faktörler ise sosyal medya verilerinin yazım kurallarına uymayan yapısı ve Instagram platformundan alınan yorumların içerik bakımından yetersiz olmasıdır.

5. Sonuç

Bu çalışmada 3 farklı sosyal ağdan toplanan Türkçe yorumların siber zorbalık içerip içermediği derin öğrenme teknikleri ile tespit edilmiştir. Veri setindeki kelimelerin sayısal temsillerini oluşturmak için Word2Vec, Fasttext, Doc2Vec kelime gömme modelleri kullanılmıştır. Kelime gömme modelleri sonucunda elde edilen vektör matrisleri ile sınıflandırma modelleri oluşturulmuştur. Modelleri test etmek için LSTM sinir ağı kullanılmıştır. Test sonucunda modeller birbirleri ile karşılaştırılarak en başarılı model tespit edilmiştir. Fasttext kullanılarak oluşturulan Sınıflandırma Modeli 2 isimli model % 93,15 başarı oranı ile diğer modellerden daha başarılı olmuştur.

Gelecekteki çalışmalarda, veri seti üzerinde farklı makine öğrenmesi yöntemleri uygulanarak bu çalışmada elde edilen sonuçlar ile karşılaştırılabilir. Ek olarak, daha doğru tahminler yapabilmek için günlük konuşma dilindeki jargonlar ile ilgili bir lookup table oluşturulup kelimelerdeki hatalar düzeltilip uygun bir standarta getirilebilir.

6. Teşekkür

Bu çalışma Mersin Üniversitesi Bilimsel Araştırma Projeleri Birimi tarafından desteklenmiştir. Proje: 2019-1-TP2-3339.

Kaynakça

[1] Kavuk, M. (2016). Ortaokul ve liselerin siber zorbalık farkındalık profillerinin oluşturulması ve okul paydaşlarına yönelik siber zorbalık farkındalık eğitimi etkililiğinin değerlendirilmesi.

[2] Aksaray, P. D. S. (2011). SİBER ZORBALIK. Journal of the Cukurova University Institute of Social Sciences, 20(2).

[3] Campbell, M. A. (2005). Cyber bullying: An old problem in a new guise?. *Journal of Psychologists and Counsellors in Schools*, *15*(1), 68-76.

[4] Zhang, X., Tong, J., Vishwamitra, N., Whittaker, E., Mazer, J. P., Kowalski, R., ... & Dillon, E. (2016, December). Cyberbullying detection with a pronunciation based convolutional neural network. In *2016 15th IEEE international conference on machine learning and applications (ICMLA)* (pp. 740-745). IEEE.

[5] Soni, D., & Singh, V. (2018, June). Time reveals all wounds: Modeling temporal characteristics of cyberbullying. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 12, No. 1).

[6] Özel, S. A., Saraç, E., Akdemir, S., & Aksu, H. (2017, October). Detection of cyberbullying on social media messages in Turkish. In *2017 International Conference on Computer Science and Engineering (UBMK)* (pp. 366-370). IEEE.

[7] Bozyiğit, A., Utku, S., & Nasiboğlu, E. (2018). Sanal zorbalık içeren sosyal medya mesajlarının tespiti. In *3rd International Conference on Computer Sciences and Engineering UBMK*.

[8] Ön, E. P., & Yeniterzi, R. (2020). Cyberbullying detection using deep learning and word embedding analysis [Derin öğrenme ile siber zorbalık tespiti ve kelime vektör temsili analizi].

[9] Balakrishnan, V., Khan, S., & Arabnia, H. R. (2020). Improving cyberbullying detection using Twitter users' psychological features and machine learning. *Computers & Security*, *90*, 101710.

[10] Laxmi, S. T., Rismala, R., & Nurrahmi, H. (2021, August). Cyberbullying Detection on Indonesian Twitter using Doc2Vec and Convolutional Neural Network. In *2021 9th International Conference on Information and Communication Technology (ICoICT)* (pp. 82-86). IEEE.

[11] Alsubait, T., & Alfageh, D. (2021). Comparison of Machine Learning Techniques for Cyberbullying Detection on YouTube Arabic Comments. *International Journal of Computer Science & Network Security*, *21*(1), 1-5.

[12] REZVANI, N., & BEHESHTI, A. (2021). TOWARDS ATTENTION-BASED CONTEXT-BOOSTED CYBERBULLYING DETECTION IN SOCIAL MEDIA. *Journal of Data Intelligence*, *2*(4), 418-433.

[13] Luo, Y., Zhang, X., Hua, J., & Shen, W. (2021, August). Multi-featured Cyberbullying Detection Based on Deep Learning. In *2021 16th International Conference on Computer Science & Education (ICCSE)* (pp. 746-751). IEEE.

[14] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[15] Erdinç, H. Y., & Güran, A. (2019, April). Semi-supervised turkish text categorization with word2vec, doc2vec and fasttext algorithms. In *2019 27th Signal Processing and Communications Applications Conference (SIU)* (pp. 1-4). IEEE.

[16] Alessa, A., Faezipour, M., & Alhassan, Z. (2018, June). Text classification of flu-related tweets using fasttext with sentiment and keyword features. In *2018 IEEE International Conference on Healthcare Informatics (ICHI)* (pp. 366-367). IEEE.

[17] Çelik, Ö., & Koç, B. C. TF-IDF, Word2vec ve Fasttext Vektör Model Yöntemleri ile Türkçe Haber Metinlerinin Sınıflandırılması. *Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi*, *23*(67), 121-127.

[18] Le, Q., & Mikolov, T. (2014, June). Distributed representations of sentences and documents. In *International*

conference on machine learning (pp. 1188-1196). PMLR.
[19] Arslan, M. (2020). *Bir İnsan Çiz Projektif Testi Yapay Zekâ tabanlı sistem tasarımı* (Master's thesis, Lisansüstü Eğitim Enstitüsü).
[20] Olah, C. (2021, 27 Mayıs), Understanding LSTM Networks, Erişim Adresi: http://colah.github.io/posts/2015-08-UnderstandingLSTMs/
[21] Karakoyun, E. Ş. (2018). *Derin öğrenme ile zaman serilerinin gerçek zamanlı tahmini* (Master's thesis, Necmettin Erbakan Üniversitesi Fen Bilimleri Enstitüsü).
[22] Zhang, Z., & Sabuncu, M. R. (2018, January). Generalized cross entropy loss for training deep neural networks with noisy labels. In *32nd Conference on Neural Information Processing Systems (NeurIPS)*.