İZMİR KÂTİP ÇELEBİ ÜNİVERSİTESİ
2010
FEN BİLİMLERİ ENSTİTÜSÜ

Sosyal Ağlar Üzerinde Mültecilere Yönelik Nefret Söyleminin Metin Madenciliğine Dayalı Tespiti

Yazılım Mühendisliği Ana Bilim Dalı

Yüksek Lisans Tezi

Yazar Adı

Figen Eğin

ORCID 0000-0003-4865-5789

Tez Danışmanı: Doç. Dr Vahide Bulut

ORCID 0000-0002-0786-8860

Eş Danışman: Doç. Dr. Aytuğ Onan

ORCID 0000-0002-9434-5880

Mayıs 2022

İzmir Kâtip Çelebi Üniversitesi Fen Bilimleri Enstitüsü öğrencisi **Figen Eğin** tarafından hazırlanan **Sosyal Ağlar Üzerinde Mültecilere Yönelik Nefret Söyleminin Metin Madenciliğine Dayalı Tespiti** başlıklı bu çalışma tarafımızca okunmuş olup, yapılan savunma sınavı sonucunda kapsam ve nitelik açısından başarılı bulunarak jürimiz tarafından YÜKSEK LİSANS TEZİ olarak kabul edilmiştir.

ONAYLAYANLAR:

Tez Danışmanı: **Doç. Dr. Vahide Bulut**
İzmir Kâtip Çelebi Üniversitesi

Eş Danışman: **Doç Dr. Aytuğ Onan**
İzmir Kâtip Çelebi Üniversitesi

Jüri Üyeleri: **Dr. Öğr. Üyesi Emre Şatır**
İzmir Kâtip Çelebi Üniversitesi

Dr. Öğr. Üyesi Emin Borandağ
Manisa Celal Bayar Üniversitesi

Savunma Tarihi: 29.05.2023

Yazarlık Beyanı

Ben, **Figen Eğin** başlığı **Sosyal Ağlar Üzerinde Mültecilere Yönelik Nefret Söyleminin Metin Madenciliğine Dayalı Tespiti** olan bu tezimin ve tezin içinde sunulan bilgilerin şahsıma ait olduğunu beyan ederim. Ayrıca:

- Bu çalışmanın bütünü veya esası bu üniversitede Yüksek Lisans derecesi elde etmek üzere çalıştığım süre içinde gerçekleştirilmiştir.

- Daha önce bu tezin herhangi bir kısmı başka bir derece veya yeterlik almak üzere bu üniversiteye veya başka bir kuruma sunulduysa bu açık biçimde ifade edilmiştir.

- Başkalarının yayımlanmış çalışmalarına başvurduğum durumlarda bu çalışmalara açık biçimde atıfta bulundum.

- Başkalarının çalışmalarından alıntıladığımda kaynağı her zaman belirttim. Tezin bu alıntılar dışında kalan kısmı tümüyle benim kendi çalışmamdır.

- Kayda değer yardım aldığım bütün kaynaklara teşekkür ettim.

- Tezde başkalarıyla birlikte gerçekleştirilen çalışmalar varsa onların katkısını ve kendi yaptıklarımı tam olarak açıkladım.

Tarih: 09.05.2023

Sosyal Ağlar Üzerinde Mültecilere Yönelik Nefret Söyleminin Metin Madenciliğine Dayalı Tespiti

Öz

Nefret söylemi, belirli kişi ya da gruplara yönelik olarak ait oldukları kimlik nedeniyle aşağılama, ayrımcılık veya ötekileştirme içeren her türlü davranış, yazı veya konuşma olarak tanımlanmaktadır. Türkiye, Suriye iç savaşı sonrası yoğun bir göç almış ve bu süreçte mültecilere karşı nefret söyleminin yükseldiği gözlenmiştir. Nefret söylemi özellikle sosyal medyada hızla yayılabilmektedir. Nefret söyleminin kolaylıkla şiddet eylemlerine dönüşebileceği göz önüne alındığında, sosyal medya üzerinde tespit edilerek yayılımının engellenmesinin önemi ortaya çıkmaktadır. Bu çalışmada Twitter üzerindeki veriler kullanılmış ve araştırmanın ilk aşamasında mültecilere yönelik nefret söyleminin tespiti için 9778 tweet içeren bir veri seti oluşturulmuştur. Bu veri seti üzerinde öncelikle makine öğrenmesi modelleri (Lojistik Regresyon, Destek Vektör Makineleri, Karar Ağaçları, Rastgele Orman ve Yapay Sinir Ağları), Word2Vec ve TF-IDF kelime temsil yöntemleri ile uygulanmıştır. Makine öğrenmesi modelleri arasında en iyi performansın TF-IDF kelime temsil yöntemi ile uygulanan Lojistik Regresyon, Destek Vektör Makineleri ve Rastgele Orman ile elde edildiği ve 0.81 doğruluk değerine ulaşıldığı görülmüştür. Sonraki aşamada BERT tabanlı bir model olan BERTurk ile farklı hiperparametreler kullanılarak yürütülen deneysel çalışmalar sonucu, 0.85 doğruluk değerine ulaşılmıştır. Araştırmanın mültecilere yönelik nefret söyleminin tespiti konusunda yürütülen çalışmalara katkısının olacağı düşünülmektedir.

Anahtar Sözcükler: Metin madenciliği, nefret söyleminin tespiti, makine öğrenmesi, derin öğrenme, sosyal medya

Text Mining Detection of Hate Speech Against Refugees on Social Networks

Abstract

Hate speech is defined as any kind of behavior, writing or speech that includes humiliation, discrimination or marginalization towards certain people or groups because of their identity. Turkiye received an intense migration after the Syrian civil war and it was observed that hate speech against refugees increased in this process. Hate speech can spread rapidly, especially on social media. The importance of detecting and preventing the spread of hate speech on social media becomes clear when it is taken into account that hate speech can easily turn into acts of violence. In this study, the data on Twitter was used and in the first stage of the research, a data set containing 9778 tweets was created to detect hate speech against refugees. First of all, machine learning models (Logistic Regression, Support Vector Machines, Decision Trees, Random Forest and Artificial Neural Networks), Word2Vec and TF-IDF word embeddings methods were applied on this data set. It was seen that the best performance among machine learning models was obtained with Logistic Regression, Support Vector Machines and Random Forest applied with TF-IDF word embeddings method, and an accuracy value of 0.81 was reached. In the next stage, as a result of experimental studies carried out with BERTurk, a BERT-based model, using different hyperparameters, an accuracy value of 0.85 was reached. It is thought that the research will contribute to the studies carried out on the detection of hate speech against refugees.

Keywords: Text mining, detecting hate speech, machine learning, deep learning, social media

Beni her konuda ve her zaman destekleyen canım aileme…

Teşekkür

Bu araştırmanın gerçekleştirilmesinde değerli bilgilerini paylaşarak bana yol gösteren, karşılaştığım güçlüklerde her zaman tüm desteğiyle yanımda olan, zamanını ve ilgisini hiçbir zaman esirgemeyerek çalışmamla yakından ilgilenen saygıdeğer danışman hocalarım Doç. Dr. Vahide Bulut ve Doç. Dr. Aytuğ Onan'a teşekkürü bir borç bilirim.

Her konuda bana destek olan, beni yüreklendiren ve bu süreçte de en büyük yardımcım olan sevgili eşim Ercan EĞİN'e, yaşlarından beklenmeyecek bir sabır ve anlayış gösteren sevgili çocuklarım Yiğit Efe ve Ahmet Kemal'e, beni ilgi ve sevgiyle büyüten canım annem ve canım babama sonsuz teşekkürlerimi sunarım.

İçindekiler

Şekiller

Tablolar

Kısaltmalar Listesi

TÜİK	Türkiye İstatistik Kurumu
İKÇÜ	İzmir Kâtip Çelebi Üniversitesi
DVM	Destek Vektör Makineleri
NLP	Doğal Dil İşleme
NB	Naive Bayes
CS	Bilgisayar Bilimi
TF-IDF	Terim Sıklığı-Ters Belge Sıklığı
YSA	Yapay Sinir Ağları

Bölüm 1

Giriş

Sosyal ağlar, günümüzde toplumun büyük bir kesimine ulaşabilen bir iletişim aracı olarak karşımıza çıkmaktadır. İnternet kullanımının artması ve mobil cihazların yaygınlaşması ile birlikte hayatımızda oldukça geniş bir yer kaplamaya başlayan sosyal ağlar, haber alma ya da bir fikri paylaşma gibi farklı amaçlarla kullanılabilmektedir. Gelinen noktada, sosyal medya araçlarının toplumları harekete geçirebilecek bir güce sahip olduğu savunulabilir. Öyle ki Ortadoğu ve Kuzey Afrika'da 2011 yıllarında yaşanan halk hareketlerinde, sosyal medya araçlarının halkın örgütlenmesinde önemli rol oynadığı görülmektedir [1]. Ayrıca sahip olduğu özellikler sayesinde, büyük kitleleri etkileme imkânının olduğu görülen bu iletişim aracı, insanları manipüle etmek veya olumsuz söylemleri yaymak için de kullanılabilir [2]. Nitekim geniş bir kullanım oranına ulaşan sosyal medya araçlarının, nefret söyleminin yaygınlaşmasında önemli bir rol oynadığı görülmektedir [3]. Nefret söyleminin yaygınlaşması ise birey veya topluluklara yönelik nefret suçlarına ve şiddet olaylarının sebep olabilmesi nedeniyle önlem alınması gereken önemli bir sorundur.
Birçok boyutu olan nefret söylemi, bazen bir kişiyi hedef alabileceği gibi bazen de belirli bir grubu hedef alabilmektedir. Hedefi ne olursa olsun nefret söyleminin yaygınlaşmasının önüne geçebilmenin ilk adımı, nefret söyleminin tespitidir. Sosyal medya kullanımının artmasıyla birlikte önemli bir kriz haline gelen nefret söyleminin [4], yapay zekâ ile tespit edilebilmesi önemli bir iş gücü tasarrufu sağlayacaktır. Fakat dilin esnekliği, yazım hataları, dilin karmaşıklığı gibi durumlar bu tespiti zorlaştırmaktadır. Bu konunun, veri madenciliği ve yapay zekâ alanında çalışan araştırmacıların oldukça ilgisini çektiği ve literatürde bu konuda birçok çalışma olduğu görülmektedir. Bu araştırmalarda, başta Destek Vektör Makineleri (DVM) modeli olmak üzere makine öğrenmesi modellerinin sıklıkla kullanıldığı, ayrıca derin öğrenme modelleri ile de yüksek başarımlı sonuçlar elde edildiği görülmektedir [5].

Literatürdeki çalışmalar incelendiğinde, bu konunun yoğun olarak çalışıldığı görülse de, özellikle Türkçe diline yönelik çalışmaların azlığı ve bu alanda nitelikli bir veri

setinin eksikliği göze çarpmaktadır. Nefret söyleminin gerçek dünyaya etkisi göz önüne alındığında, bu konuda Türkçe nefret söyleminin tespitine yönelik bir çalışmanın literatüre önemli katkıları olacağı düşünülmektedir.

1.1. Tezin Amacı ve Kapsamı

Türkiye son dönemde, yakın coğrafyasında yaşanan savaş ve iç karışıklıklar neticesinde önemli oranda göç almıştır. Son yayınlanan TÜİK verilerine göre 2019 yılında Türkiye'ye göç eden yabancı uyruklu kişi sayısı %17 artış göstererek 677 bin 42 kişi olmuştur [6]. Türkiye'de geçici sığınma statüsünde olan Suriyeli sayısı ise 3 buçuk milyonu geçmiştir [7]. Yaşanan bu yoğun göç, toplumun bazı kesimlerinde huzursuzluğa neden olabilmekte ve çeşitli endişeler doğurabilmektedir. Böyle hassas bir ortamda, nefret söyleminin yaygınlaşmasının çok ciddi sonuçları olabilir. Nefret söyleminin en çok körüklendiği ortamların başında ise sosyal medya gelmektedir. Sosyal medya, nefret söyleminin toplumun geniş kesimlerine çok hızlı bir şekilde ulaşmasında önemli rol oynamaktadır. Özellikle siyasi içeriklerin ağırlıklı olarak paylaşıldığı bir sosyal ağ olan "Twitter" kötü niyetli kişiler tarafından, toplumu belirli gruplara karşı kışkırtmak için yalan haberlere sahne olabilmektedir. Mülteci ve göçmenlere karşı yürütülen nefret kampanyalarının kolaylıkla şiddet eylemlerine dönüşebileceği göz önünde bulundurulduğunda, nefret söyleminin yayılmadan kontrol altına alınmasının önemi daha net ortaya çıkmaktadır. Bu noktada nefret söyleminin hızlı bir şekilde tespiti için yapay zekânın kullanımının iyi bir çözüm olacağı düşünülmektedir. Literatürde nefret söyleminin tespitine yönelik birçok çalışma olmasına karşın mültecilere yönelik büyük miktarda veri içeren Türkçe bir veri setinin olmaması ise önemli bir eksiklik olarak karşımıza çıkmaktadır. Tüm bunlar göz önünde bulundurularak bu tez çalışmasında;

1. Literatüre açık erişimli nefret söylemine yönelik Türkçe bir veri setinin kazandırılması,
2. Mültecilere yönelik nefret söyleminin tespiti için farklı kelime temsil yöntemleri, makine öğrenmesi algoritmaları ve derin öğrenme algoritmalarının karşılaştırılarak etkin bir model önerisinin geliştirilmesi amaçlanmıştır.

Bu kapsamda Twitter üzerinden çekilen tweetler ile hazırlanan Türkçe veri seti üzerinde, makine öğrenmesi ve derin öğrenme modelleri ile nefret söyleminin tahminlemesi yapılmıştır.

1.2. Sınırlılıklar

Yüksek kullanım oranlarına sahip birçok sosyal medya aracı bulunmaktadır. Bu araştırmada ise sadece Twitter sosyal ağı üzerinden paylaşılan içeriklere odaklanılmıştır. Twitter, kullanıcıların 280 adet karakter içeren 'tweet' olarak isimlendirilen bir metin mesajı paylaşılmasına yönelik bir sosyal medya aracı olmasına karşılık kullanıcılar bu tweete metin, resim, bağlantı, başlık etiketleri (hashtag) ekleyebilmektedirler. Bu araştırma kullanıcıların paylaştıkları metin içerikli tweetler üzerinde nefret söyleminin tespit edilmesiyle sınırlı olup, resim, video veya bağlantıları takip eden içerikler incelenmemiştir. Kullanılan veri seti 9778 adet etiketlenmiş veri ile sınırlıdır.

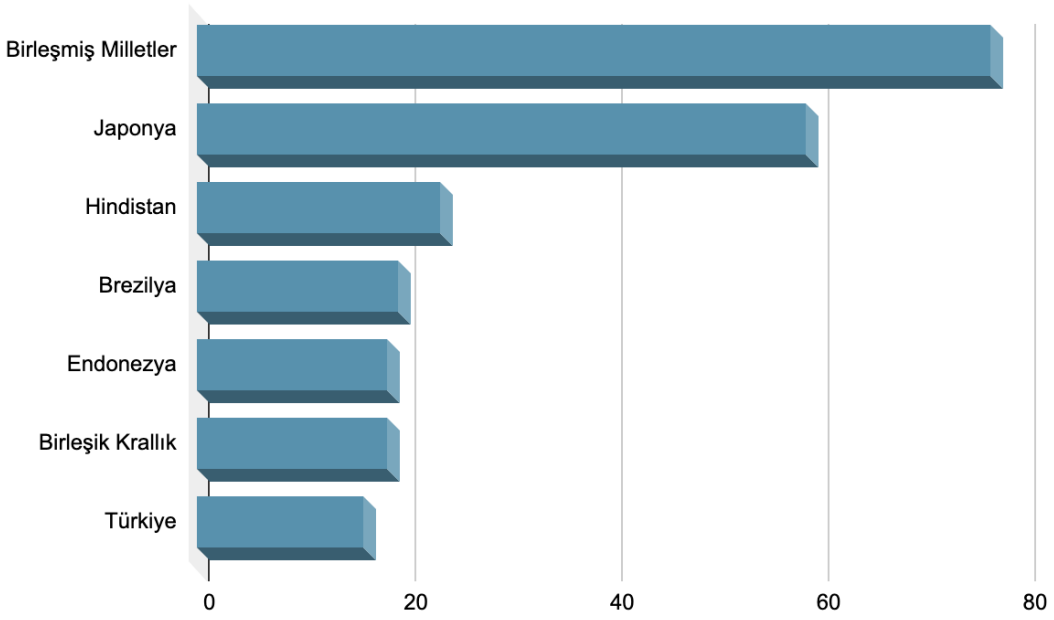
Tezin izleyen bölümleri şu şekilde organize edilmiştir; ikinci bölümde sosyal ağlarda nefret söyleminin tespitine yönelik literatürde yer alan çalışmalara yer verilmiş, nefret söyleminin kuramsal tanımı ve boyutları tartışılmıştır. Üçüncü bölümde metin madenciliğinde kullanılan yöntemler, modellerin değerlendirilmesi için kullanılan performans metrikleri irdelenmiştir. Dördüncü bölümde kullanılan materyal ve yöntem üzerinde durulmuştur. Beşinci bölümde ise elde edilen deneysel sonuçlar sunulmuştur. Beşinci bölümde bulgular tartışılmış ve altıncı bölümde araştırmanın sonuçları sunulmuştur.

Bölüm 2

Sosyal Ağlarda Nefret Söyleminin Tespiti

2.1 Sosyal Ağlar

Sosyal ağlar, kullanıcılar arasındaki etkileşimlerin algılanmasını kolaylaştıran, kullanıcı tarafından oluşturulan içerikten değer elde eden, kitlesel kişisel iletişimin internet tabanlı, kontrolsüz ve kalıcı kanalları olarak tanımlanmaktadır [8]. Bilgisayar aracılı bir iletişim ortamı olan sosyal ağlarda, bireyler benliklerine ilişkin bilgilerin sadece istedikleri kadarını paylaşarak var olabilirler. Bu özelliği ile sosyal ağ ortamında insanların yüz yüze iletişime nazaran daha dikkatsiz veya nezaketsiz davranabildikleri savunulabilir. Sosyal ağların kullanım oranlarına bakıldığında ise tüm dünyada geniş bir kullanıcı kitlesine sahip oldukları görülmektedir.



Şekil 2.1: Ülkelere göre Twitter kullanım miktarları (milyon kişi) [10]

Türkiye'de sosyal ağların 2022 yılı itibariyle 67 milyonun üzerinde kullanıcısı bulunurken bu rakamın 2027 yılına kadar 76 milyonun üzerine çıkacağı öngörülmektedir [9]. Türkiye 16.1 milyon Twitter kullanıcı sayısı ile ise Dünyada 7. sırada yer almaktadır [10] . Görüldüğü üzere sosyal ağlar günümüzde yoğun olarak kullanılmakta, gelecekte de kullanım oranlarının artacağı tahmin edilmektedir (Şekil 2.1).

2.2 Nefret Söylemi ve Boyutları

Toplumsal bir sorun olarak nefret söylemi, beşeri bilimlerde eski olmasına rağmen bilişim alanında halen yeni bir araştırma alanıdır [11]. Nefret söyleminin ne olduğuna yönelik tanımlara birçok kaynakta rastlanmaktadır. Castano-Pulgarín ve ark. [12] sosyal ağlar ve nefret söylemi üzerine yürüttükleri sistematik tarama çalışmalarında, 2015 ve 2019 yılları arasındaki, 67 makaleyi incelemişler ve araştırmalarında "sanal nefret" terimini kullanmışlardır. Bu araştırma sonucunda sanal nefreti, ortak bir özelliği paylaşan belirli bir grup insana yönelik şiddet içeren saldırgan bir dilin kullanılması olarak tanımlamışlardır. Araştırmacılara göre sanal nefret, sosyal ağların ve internetin kullanımı ile yaygınlaşmaktadır. Nobata ve ark. [13] ise, bir kişiyi veya grubu ırk, etnik köken, din, cinsiyet, yaş, engellilik durumu veya cinsel yönelim/cinsiyet kimliğine dayalı olarak aşağılayan veya onlara saldıran bir dil olarak tanımlamaktadırlar. McNamee, Peterson ve Pena [14] tarafından yürütülen, çeşitli nefret gruplarına ait 21 web sitesinin incelendiği kuramsal temelli bir çalışmada ise, "eğit, katıl, çağrıda bulun ve suçlama" olmak üzere başlıca dört tema ortaya çıkmıştır. Schmidt ve Wiegand [15], nefret söylemini aşağılayıcı içerik türleri için bir genelleme olarak kullanmıştır. Papcunova ve ark. [16] yürüttükleri araştırmada iki odak grup kurmuşlar ve bu odak grupların 2 ay süren 8 görüşme gerçekleştirmesini sağlamışlardır. Birinde her yaş, cinsiyet ve meslek grubundan insanların diğerinde psikologların bulunduğu bu odak grupların tartışmaları sonucunda göç ve mülteciler bağlamında nefret söylemini tanımlamışlardır. Buna göre nefret söyleminin "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" olduğunu belirtmişlerdir.

Avrupa Birliği Nefret Söylemine İlişkin Strateji ve Eylem Planı kapsamında ise nefret söylemi: "Bir kişiye veya gruba kimlik unsurlarına (din, etnik köken, milliyet, ırk veya renk) dayalı aşağılayıcı, saldırgan veya ayrımcı bir dil içeren her türlü konuşma, yazı veya davranış" şeklinde tanımlanmaktadır [17]. Nefret söylemi, bir çeşit taciz dili olmasına karşın mutlaka küfür içermesi gerekmez, insanları kışkırtıcı bir yönü olabilir ve ayrıca ayrımcılığın farklı bir formu olarak ifade edilebilir [5]. Nefret söyleminin kuramsal tanımına yönelik bu çalışmalar; nefret söyleminin hedefinde milliyet, ırk, cinsiyet/cinsel yönelim gibi ortak özellikleri paylaşan herhangi bir grup olabileceğini göstermektedir. Bu gruba yöneltilen aşağılama, ayrımcılık veya ötekileştirme içeren söylemler nefret söylemi olarak nitelendirilebilir. Nefret söyleminde öne çıkan bu öğeler Şekil 2.2'de özetlenmiştir.

Nefret söyleminin yayılmasının önüne geçmek, toplumsal barış ve huzur için önemlidir. Bu söylemin kontrolsüz ve çok hızlı bir şekilde yayılmasına ortam sağlayan sosyal ağ şirketlerinin bu konuda belirli ilkeler belirledikleri görülmektedir. Twitter, bu konudaki politikasını şu şekilde ifade etmektedir: "Irk, etnik köken, ulusal köken, kast, cinsel yönelim, cinsiyet, cinsel kimlik, dini görüş, yaş, engellilik durumu veya ciddi hastalık temelinde diğer insanlara karşı şiddeti teşvik edemez veya doğrudan onlara saldıramaz veya onları tehdit edemezsiniz. Ayrıca, bu kategoriler temelinde başkalarına zarar vermeye teşvik edilmesine de izin vermeyiz."

Şekil 2.2: Nefret söyleminde öne çıkan kavramlar

Nefret söyleminin inanç ve mezhebe, ırka, cinsiyete, siyasi görüşe, yabancı ve göçmenlere, engelli ve çeşitli hastalıklara yönelik olmak üzere altı boyutta ele alındığı görülmektedir [18]. Bu çalışma kapsamında mülteci ve göçmenlere yönelik nefret söylemine odaklanılmıştır. İzleyen bölümde nefret söyleminin tespiti üzerine yapılan çalışmalar incelenecektir.

2.3 Nefret Söyleminin Tespitine Yönelik Yapılan Araştırmalar

Sosyal ağlar üzerinde nefret söyleminin tespiti ve nefret dolu içeriklerin önüne geçilmesi, araştırmacı ve uygulayıcılar için hala birçok zorluk içeren bir konu olarak karşımıza çıkmaktadır. Bu zorlukların aşılması ve etkili bir modelin ortaya konması için birçok araştırma yürütüldüğü görülmektedir. Bu bölümde, "nefret söyleminin tespiti" ve "hate speech detection" anahtar kelimeleri kullanarak Google Scholar ve Web of Science veri tabanları üzerinden yapılan tarama sonucu öne çıkan çalışmaların bir kısmı sunulmuştur. Bu çalışmalardan ise özellikle Twitter sosyal medya platformu üzerinden elde edilen veri setleri ile yapılmış veya mültecileri konu edinen çalışmalara ağırlık verilmiştir.

Sosyal medya üzerinden yapılan paylaşımlarda, dilin esnek yapısı nefret söyleminin tespitinde önemli bir engel olarak karşımıza çıkabilir. Nitekim Chakraborty ve Masud [4] bu konuya dikkat çekmişler, yaptıkları araştırmalarında nefret söyleminin tespitinde öne çıkan sorunları dilin bağlam ve öznellik özelliklerinden kaynaklanan sıkıntılar, internetteki iletişimin çok yönlü doğası, standart ve geniş ölçekli bir nefret söylemi veri setinin eksikliği ve nefret söyleminin çok dilli doğası olarak sıralamışlardır. Araştırmacılar sosyal medyada nefret söyleminin yayılmasını sınırlamak için metin madenciliği alanında bir dizi çözüm önermişlerdir. Araştırmacıların çözüm aradıkları ilk sorun Twitter 'da nefret söyleminin yayılmasını tahminlemek olmuştur. Nefret söyleminin yayılımında yalnızca topolojinin yeterli olmadığını görmüşlerdir. Tweetlere verilen yanıtların nefret söylemi içerip içermediğine yönelik araştırmalarında ise kaynak tweet ile verilen yanıtlar arasında nefret doluluğu açısından bir ilişki olmadığını ortaya koymuşlardır. Ayrıca açıkça

saldırgan bir metnin nefret dolu olma olasılığını diğer sınıflara göre %25 daha yüksek ve saldırgan olmadığını bildiğimiz bir tweetin, nefret dolu olma şansını %50 az olarak tespit etmişlerdir.

Kuş [19] tarafından gerçekleştirilen bir çalışmada ise, dijital nefret söyleminin Facebook üzerinden tespiti üzerinde durulmuştur. Bu kapsamda BBC World Haber Servisinin Facebook'ta paylaştığı haberler incelenmiş, mültecilerle ilgili haberlerin yorumları metin madenciliği yöntemlerine tabi tutulmuştur. Sonuç olarak yorumların büyük çoğunluğunun mültecilere yönelik olumsuz duygular içerdiği ve nefret söyleminin artış gösterdiği ortaya konmuştur. Bu çalışmada metnin analizi için Rapidminer Studio kullanıldığı görülmüştür.

Simon, Baha ve Garba [5] yürüttükleri sistematik tarama çalışmasında, sosyal ağlardaki nefret söylemi ve diğer anti-sosyal davranışların tespiti için önerilen modellerin çoğunun bir metin sınıflandırma görevi olarak ele alındığını ve bu modellerin çoğunun denetimli makine öğrenimi algoritmalarına dayandığını tespit etmişlerdir. Bir makine öğrenimi algoritması olan DVM algoritmasının, metin sınıflandırması için yüksek performans düzeyi ve doğruluğu sunması nedeniyle nefret barındıran içeriklerin tespitinde özellikle tercih edildiğini görmüşlerdir. TF-IDF kelime temsil yönteminin ise nefret söylemi tespitinde ve diğer çevrimiçi anti-sosyal davranışların sınıflandırılmasında etkili olduğunu tespit etmişlerdir. Benzer şekilde derin öğrenme algoritmalarının da son zamanlarda birçok araştırmacı tarafından metin sınıflandırma problemleri için kullanıldığını ve özellikle hibrit derin öğrenme modellerinin başarımlarının yüksek olduğunu görmüşlerdir. Araştırmacılar inceledikleri makalelerin hiçbirinde daha yüksek n değerleri denenmemiş olmasına karşın, n değeri ne kadar yüksek olursa model eğitiminin o kadar zaman alacağını ve bir o kadar fazla veri gerektireceğini belirtmişlerdir.

Malmasi ve Zampieri [20], sosyal medya üzerinden nefret söyleminin tespiti için 14,509 tweet içeren bir veri seti kullanmışlardır. Kullandıkları veri seti içerisinde 2,399 "nefret söylemi", 4,836 "saldırgan dil" ve 7,274 "OK" etiketli veri bulunmaktadır. Çalışmalarında DVM modelinden yararlanmışlar ve Ana Dil Tanımlaması için çok etkili bir sınıflandırıcı olduğu kanıtlanmış olan LIBLINEAR paketini kullanmışlardır. Çalışmalarında %78 doğruluk oranına ulaşmışlardır.

Mathew ve ark. [21] tarafından yürütülen araştırmada ise nefret dolu içerikler paylaşan kullanıcıların paylaşımlarının yayılma hızına bakılmıştır. DeGroot'un modelinin kullanıldığı bu çalışmada araştırmacılar nefret dolu içerikler paylaşan kullanıcıların benzer paylaşımlar yapan diğer kullanıcılarla yoğun bir bağ ile bağlı olduklarını ve paylaşımlarının daha geniş ve uzak bir alana daha hızlı bir şekilde ulaşabildiğini tespit etmişlerdir. Ayrıca nefret içeren gönderilerin büyük bir kısmı görsel içerikli olduğu için ileriki çalışmalar için görsel ve videolar arasında ayrım yapabilen bir sınıflandırma üzerinde çalışılmasını tavsiye etmişlerdir. Başka bir çalışmada ise [22], nefret dolu YouTube videolarına yapılan yorumlar ile oluşturulan bir veri seti üzerinde Gauss Naive Bayes (GNB), Rastgele Orman, Lojistik Regresyon (LR), DVM, XGBoost (XGB), CatBoost (CB), Karar Ağaçları ve MLP, LSTM (Long Short-Term Memory Networks, Uzun Kısa Dönemli Bellek) gibi modeller uygulanmıştır. Bu çalışmada CatBoost (CB) %78 doğrulukla en iyi performansı gösterirken, bunu %74 doğrulukla XGBoost takip etmiştir. Yine benzer bir çalışmada ise [23], nefret söylemine yönelik kapsamlı bir veri seti oluşturulmuş; 20,148 İngilizce tweet içeren veri setinde, CNN-GRU, BiRNN, BiRNN-Attention ve BERT modellerine ilişkin sonuçlar doğruluk, macro F1-Ölçütü, ve AUROC ölçütü ile değerlendirilmiştir. En iyi performansın 0.698 doğruluk, 0.687 F1- Ölçütü ve 0.851 AUROC olarak elde edildiği görülmüştür.

Mullah ve Zainon [24], sosyal medyada nefret söylemi tespiti için makine öğrenimi algoritmalarını ve tekniklerini gözden geçirmişlerdir. Nefret söylemi sınıflandırmasını beş temel bileşen olan veri toplama ve araştırma, özellik çıkarma, boyut azaltma, sınıflandırıcı seçimi ve eğitimi, model değerlendirmesi açısından incelemişlerdir. Araştırmaları kapsamında, klasik makine öğrenimi kullanan nefret söylemi tespitinde topluluk ve derin öğrenme tekniklerinden daha fazla araştırma olduğunu ortaya çıkarmışlardır. Ayrıca kullanılabilecek veri setleri ile ilgili zorlukları dile getirmişlerdir. Araştırmacılar Nijerya'da bazı özel karakterlerin, medeni durum ya da sağlık durumu gibi değişkenlerin nefret söylemi özelliği taşıyabildiği örneğini vererek nefret söyleminin kültürel farklılıklar nedeniyle her ülke bazında incelenmesi gerektiğini vurgulamışlardır.

Burnap ve Williams [25] tarafından yürütülen çalışmada, İngilizce Twitter üzerinden toplanan veriler: "Bu metin ırk, etnik köken veya din açısından saldırgan mı yoksa düşmanca mı?" sorusu temelinde incelenmiş ve veri setindeki tweetler "evet, hayır, kararsız" etiketleri kullanılarak insan eliyle kodlanmıştır. Bu aşamada görevlerin

birden fazla kişiye dağıtılmasını sağlayan "CrowdFlower" hizmeti kullanılarak tweetler 158 kişi tarafından kodlanmıştır. Aynı etiketle etiketlenme oranı %75'ten az olan ve "kararsız" olarak etiketlenen tweetler kaldırılmıştır. Araştırmacılar, on kat çapraz doğrulama yaklaşımı, BoW ve n-gram öznitelik çıkarımlarını olasılıksal, kurala dayalı ve uzamsal tabanlı sınıflandırıcılar kullanarak sınıflandırmışlar, %98 başarım oranına ulaşmışlardır.

Waseem ve Hovy [26], nefret söyleminin sınıflandırmasına odaklandıkları çalışmalarında yine Twitter üzerinden İngilizce bir veri seti elde etmişlerdir. Karakter ve kelime n-gramlarını kullandıkları çalışmalarında %64 başarım elde etmişlerdir.

Davidson ve ark. [27], sosyal medyada otomatik nefret söylemi tespiti için temel zorluğun, nefret söyleminin diğer saldırgan dil örneklerinden ayrılması olduğunu ve sözcüksel algılama yöntemlerinin, belirli terimleri içeren tüm mesajları nefret söylemi olarak sınıflandırmaları nedeniyle başarımlarının düşük olduğunu belirtmişlerdir. Araştırmalarında nefret söylemi anahtar sözcükleri içeren tweetleri toplamak için kitle kaynaklı bir nefret söylemi sözlüğü kullanmışlar ve İngilizce tweetlerden oluşan veri setindeki verileri "nefret söylemi, yalnızca saldırgan dil içerenler ve hiçbirini içermeyenler" şeklinde etiketlemişlerdir. Nefret söylemi ve saldırgan dil sınıflandırması yapmak üzere TF-IDF ve Naive Bayes (NB), Lojistic Regresyon (LR), Rastgele Orman (RF), Karar Ağaçları ve doğrusal Destek Vektör Makinelerini kullandıkları çalışmalarında %90 başarım elde etmişlerdir. Irkçı ve homofobik tweetlerin nefret söylemi olarak sınıflandırılma olasılığının daha yüksek olduğunu, ancak cinsiyetçi tweetlerin genellikle saldırgan olarak sınıflandırıldığını, açıkça nefret anahtar sözcükleri içermeyen tweetlerin sınıflandırılmasının daha zor olduğunu bulmuşlardır.

İngilizce tweetler üzerine yapılan başka bir çalışma ise De Smedt ve ark.'na [28] aittir. Nefret söyleminin sınıflandırmasına yönelik olan çalışmada, Ekim 2014- Aralık 2016 tarihleri arasında toplanan 45.000 Twitter mesajı doğrulsal DVM ve Karar Ağaçları kullanılarak sınıflandırılmış ve %80 başarım elde edilmiştir.

Gaydhani ve ark. [29], üç farklı veri setinin kombinasyonundan elde ettikleri ve "Nefret dolu, saldırgan, temiz" şeklinde etiketlenmiş İngilizce tweetler üzerinde nefret dolu ve saldırgan dil sınıflandırmasını yapmışlardır. Bağlantılar, dolgu kelimeler,

etiketler gibi öğelerden arındırıldıktan sonra n-gram ve TF-IDF ile LR, NB ve Doğrulsal DVM kullanarak %95.6 başarım elde etmişlerdir.

Park ve Fung [30], LR, SVM, FastText, CharCNN, WordCNN ve araştırma kapsamında oluşturdukları HybridCNN algoritmalarını karşılaştırdıkları çalışmalarında hibrit modellerin daha iyi sonuçlar ürettiğini belirtmişler ve araştırmalarında %82.7 başarıma ulaşmışlardır. Pelle ve ark. [31] ise, İngilizce ve Portekizce Twitter ve diğer sosyal medya platformlarını saldırgan yorumların tespiti kelime ve karakter n-gramları ve Word2Vec ile LR modelleri F1-Ölçütü ile değerlendirmişler ve %90-97 başarım elde etmişlerdir.

Faris ve ark. [32], Twitter üzerinde Arapça dilinde siber nefretin tespitini Hibrit CNN ve LSTM modelini kullanarak %71.6 başarım elde ederken; Doris ve ark. [33] ise, yine Twitter üzerinde İngilizce nefret söylemi ve saldırgan dil tespiti için GloVe ve LSTM ile Sinir Ağları modelini kullanarak nefret söyleminin tespitinde 90.82% ve saldırgan dilin tespitinde 89.10% başarım elde etmişlerdir. Altin ve ark. [34], saldırgan dil tespitinde sinir ağları tabanlı bi-LSTM (Bidirectional Long Short-Term Memory Networks, Çift Yönlü Uzun Kısa Dönemli Bellek) modelini kullanarak %82.9 doğruluk değerine erişmişlerdir. Çalışmalarında ilk olarak, tweetleri noktalama işaretlerini kaldırılarak ama anlamı tanımlamaya katkıda bulunabileceği için emojileri ve başlık etiketlerini (hashtag) tutularak simgeleştirmişlerdir. İkinci aşamada, gömme katmanı, simgeleştirilmiş tweetteki her öğeyi (kelimeler, emojiler ve etiketler gibi) düşük boyutlu bir vektöre dönüştürmektedir. biLSTM'nin standart LSTM ile aynı işlemleri yaptığını, ancak metni soldan sağa ve sağdan sola paralel olarak işlediğini belirtmişlerdir. Jain, Kumar ve Garg'ın [35], İngilizce ve Hinçe Twitter iğneleyici tweetlerin tespiti için GloVe kelime temsil yöntemi ile bi-LSTM ve CNN modelleri ile 92.71% ve 89.05% başarım elde ettikleri görülmektedir.

Nefret söyleminin, dile özgü belirli özellikler ve Twitter gibi sosyal medya platformlarında kısa metinlerin paylaşılması nedeniyle tespitinin zorluğuna dikkat çekilen başka bir çalışmada, metinsel belirteçlerin frekanslarına ve psikolojik özelliklere dayalı olarak yeni özellik setleri geliştirilmiştir. Ardından geniş bir veri kümesi üzerinden çeşitli makine öğrenimi yöntemlerini değerlendirildiği çalışmada, Rastgele Orman ve BERT yöntemlerinin nefret söylemi içeriğini tespit etmede en etkili yöntemler olduğu ortaya konmuştur [36]. Başka bir çalışma ise Kürtçe dilinde

nefret söyleminin tespitine yönelik yürütülmüştür. Araştırmacılar Facebook sosyal medya platformu üzerinden topladıkları 6882 yorum üzerinde DVM modelini kullanmışlar, F1-Ölçütü ile modellerini değerlendirmişler ve 0.68 başarım elde etmişlerdir [37]. Twitter üzerinde saldırgan dilin tespit edilmesine yönelik yürütülen bir çalışmada, LMTweets isimli kodlayıcı 20001 adet tweet ile eğitilerek öznitelik çıkarımı yapılmıştır. Bu öznitelikler kullanılarak metin, saldırgan / saldırgan olmayan olarak sınıflandırılmış ve makine öğrenmesi ve Transformatör-Tabanlı evrişimli sinir ağı modelleri kullanılmıştır. En iyi performans LMTweets + CNN modeli ile elde edilmiştir [38].

Çalışmalar genel anlamda incelendiğinde, nefret söyleminin tespitinde yaşanan başlıca zorluklar:

- Dilin esnek yapısı
- Türkçe dilinde veri seti eksikliği
- Diğer saldırgan dil örneklerinden ayrılmasının zorluğu
- Bağlam temelli paylaşımlar
- Twitter gibi sosyal medya platformlarında çok kısa metinlerin paylaşılması şeklinde özetlenebilir.

Nefret söyleminin metin madenciliği yöntemleriyle tespitinde veri toplama ve araştırma, özellik çıkarma, boyut azaltma, sınıflandırıcı seçimi, eğitimi ve model değerlendirmesi şeklinde beş adımın uygulandığı görülmektedir. Yaygın olarak DVM modeli ve TF-IDF kelime temsil yönteminin kullanıldığı, hibrit modellerin başarımlarının daha yüksek olduğu ve son dönemde derin öğrenme modellerinin kullanıldığı çalışmaların sayısının arttığı gözlenmiştir. Twitter'ın veri seti oluşturmak için sıklıkla kullanıldığı ve İngilizce diline yönelik çalışmaların ağırlık kazandığı tespit edilmiştir. Ayrıca kültürel farklılıklar ve dilin özelliği nedeniyle nefret söylemi tespitinin ülke bazında çalışılması gereken bir konu olduğu görülmüştür.

Bu bölümde nefret söylemine yönelik yapılan çalışmalar özetlenmiştir. İzleyen bölümde metin madenciliğinde kullanılan yöntemler sunulacaktır.

Bölüm 3

Metin Madenciliği Yöntemleri

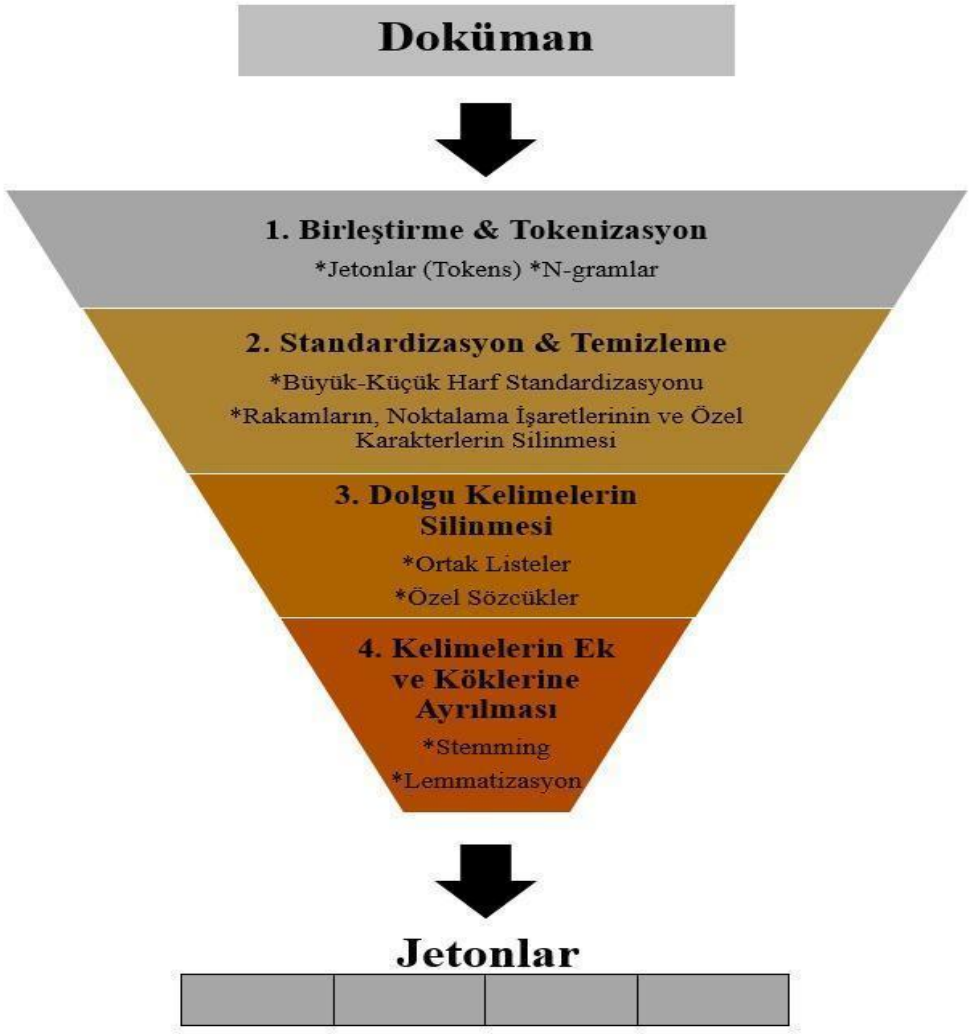
Veri madenciliğinin bir alanı olan metin madenciliği, yazılı kaynaklardan bilgi çıkarımını ifade etmektedir. Veri madenciliğinden farklı olarak metin madenciliğinde yapılandırılmış veri setlerinden değil, insanların okuması için yazılmış metinlerden, yani yapılandırılmamış ya da yarı yapılandırılmış veri setlerinden anlam çıkarılmaya çalışılmaktadır. Dolayısıyla verinin işlenebilmesi için gerçekleştirilmesi gereken birçok aşama vardır. Gerçekleştirilmesi gereken bu küçük adımlar, doğal dil işleme (NLP) alanı tarafından çalışılmaktadır [39]. NLP herhangi bir toplum tarafından kullanılan doğal dilin bilgisayar sistemleri ile mekanik olarak nasıl işlenebileceğini araştıran bir Bilgisayar Bilimi (CS) alanıdır ve doğal dil yapılarına ilişkin matematiksel modellerin tasarımını ve bunların uygulanmasını araştırır [40]. NLP uygulamaları metin işleme, özetleme, konuşma tanıma, uzman sistemleri, diller arası bilgi aktarımı gibi birçok alanı kapsar [41].

Büyük miktarlarda verinin metin olarak saklandığı ve metin tabanlı araçların kullanım sıklığı göz önüne alındığında metin işlemede başarılı modellerin ortaya konmasının önemi daha iyi anlaşılmaktadır. Bu noktada karşılaşılan önemli zorluklardan biri, sözcüklerin farklı konular söz konusu olduğunda farklı anlamlara gelebilmesi nedeniyle dilin bağlamı da dâhil olmak üzere yazılı metinlerin bilgisayar tarafından anlaşılamamasıdır. Özellikle Türkçe gibi sondan eklemeli dillerde eklere göre anlamın değişmesi önemli bir zorluk olarak karşımıza çıkmaktadır. Ayrıca kelimelerin cümle içerisindeki konumlarına göre de anlam farklılıkları oluşabilmektedir [42]. Tüm bu zorlukların üstesinden gelmek ve metinden bilgi çıkarımını sağlamak için geliştirilen kelime tabanlı, istatistiksel veya dilbilim çalışmalarını temel alan birçok yöntem bulunmaktadır. İstatistiksel veya makine öğrenimi yaklaşımı, metinlerin matematiksel temsilini kullanır. Doğal dil işleme tekniklerini kullanan dilbilimsel yöntemler, anlamın ve farklı ilişkilerin yer aldığı dil modellerini kullanan metinleri temsil eder. Metin madenciliği, genellikle büyük miktardaki metinlerde, bilgi bulmak için her iki yaklaşımı da kullanır [43].

Bu bölümde metin madenciliğinde kullanılan ön işleme ve sınıflandırma yöntemlerine ilişkin bilgi verilecektir.

3.1 Metin Ön İşleme Yöntemleri

Metin ön işleme (preprocessing), metin madenciliğinde verinin toplanması ve belge içeriğine göre özniteliklerin çıkarılmasını takip eden ilk adımdır. Özellikle makine öğrenmesi modellerinin uygulandığı durumlarda, daha etkili ve verimli bir sonuç alabilmek için ön işleme daha da önem kazanmaktadır. Farklı kaynaklardan toplanan metin şeklindeki veriler, öncelikle ön işlemden geçirilir, sonrasında metin madenciliği teknikleri kullanılarak metnin analizi gerçekleştirilir.



Şekil 3.1: Metin ön işleme adımları [44]

3.1.1 Metnin Temizlenmesi

Metnin temizlenmesi, modelin iyi bir performans göstermesi açısından çok önemlidir. Twitter gibi sosyal medya ortamlarında, yazım hatalarına, sözcüklerin yanlış kullanımına vb. sıklıkla rastlanmaktadır. Ayrıca sosyal medyaya özgü etiketler, farklı sayfalara bağlantılar gibi öğeler modelin performansını düşürebilir. Bunların temizlenerek metnin hafifletilmesi gerekir. Sayılar ve noktalama işaretleri de eğer kullanılmayacaksa bu aşamada temizlenmelidir [45]. Metin ön işlemede önemli adımlardan biri de "stop words" olarak adlandırılan gereksiz kelimelerden metnin temizlenmesidir. Bu kelimeler, doğal dilde kelimeleri veya cümleleri birbirine bağlayan veya anlamı destekleyen "ve, ile, ama, ancak, belki, biri" gibi kelimelerdir. İstatistiksel açıdan bu kelimeler daha az önemli olarak değerlendirilmektedir ve performansın arttırılması için anahtar kelimeler olarak görülmez [46]. Metni sadeleştirmek ve uygulamadan tasarruf etmek için bu kelimeler çıkartılmalıdır. Metin içerisindeki tüm harflerin küçük harfe çevrilmesi işlemi ise aynı kelimelerin büyük veya küçük harfle yazılmış versiyonlarının model tarafından farklı kelimelermiş gibi algılanmasının önüne geçmek için gerçekleştirilmesi gereken bir adımdır. Sonraki adımda, kelimelerin köklerine indirgenmesi işlemi gerçekleştirilir. Türkçe gibi sondan eklemeli dillerde kelime sonuna gelen ekler ile kelimenin anlamı tamamen değişebilmektedir. Dolayısıyla bu adımın dikkatle gerçekleştirilmesi önemlidir. Bu aşamada "stemming" ve "lemmatization" yöntemleri kullanılmaktadır. "Stemming" kelimenin sonundaki ekleri silerek en az iki harften oluşan köke ulaşmaya çalışırken [47], "lemmatization" problemi morfolojik bir yaklaşım olarak ele almakta ve kelimenin türüne bağlı olarak farklı anlamlara gelebileceğini göz önünde bulundurarak ekleri silmektedir [48]. Örneğin Türkçe bir metinde stemming işlemi ile bir hayvan ismi olan "koyun" kelimesi "koy" fiiline indirgenebilirken, lemmatizing ile kelimenin cümle içerisindeki anlamı göz önünde bulundurulduğundan bu hataya düşülmez .

3.1.2 Kelime Temsil Yöntemleri

Metin madenciliğinde en önemli noktalardan biri, eğitim verisinin vektörel olarak temsil edilmesidir. Eğitim verisinde geçen her sözcük bir özellik vektörü ile temsil edilerek; yani veride bir kelime geçiyorsa karşılık gelen niteliği 1'e, değilse 0'a ayarlanarak metin içerdiği kelime grubuyla temsil edilir [49]. Kelime temsilinde

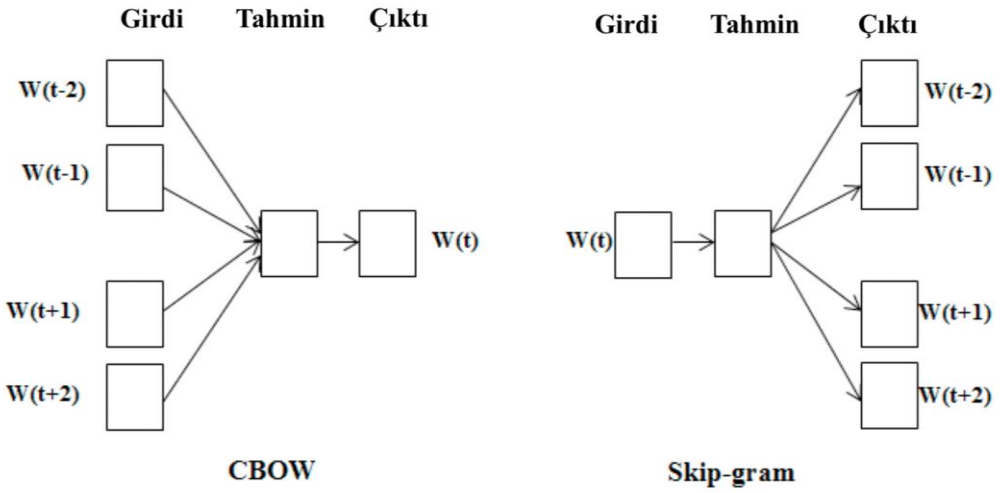
kullanılan yöntemler bağlamsal olmayan ve bağlamsal kelime temsil yöntemleri olarak sınıflandırılabilir.

Kelime Çantası Yöntemi (Bag of Words - BoW), yönteminde veri setinde geçen kelimeler için bir hazne oluşturulmaktadır. Bu şekilde veri setinde geçen kelimelerin frekansları hesaplanmaktadır. Ana fikir, her bir anahtar noktayı, genellikle kümeleme yoluyla türetilen görsel kelimelerden birine nicelleştirmektir [50]. Bu kelime haznesinin hafifletilmesi için noktalama işaretlerinin, büyük-küçük harf farklılıklarının, yazım yanlışlarının ortadan kaldırılması önemlidir. Metin içerisinde, bazı kelimeler birlikte bir anlam ifade edebilirler. Örneğin "Kara Kuvvetleri Komutanlığı" kelime öbeğinde, kelimeler tek tek ele alındığında veya bir araya geldiklerinde farklı anlamlar ifade etmektedir. Bu kelime öbekleri, frekansları hesaplanarak ortaya konmaktadır. Kelime frekansları kelimeler tek tek ele alınarak hesaplanırsa unigram, iki kelimelik gruplar olarak ele alındığında bigram, üç kelimelik gruplar oluşuyorsa trigram olarak ifade edilir [51]. BoW modeli kelimelerin sıralamasını dikkate almazken n-gram modeller, izleyen kelimenin gelme olasılığını hesaplamaktadır. Böylece analiz esnasında bilgi kaybını en aza indirgemektedirler.

Terim Sıklığı – Ters Belge Sıklığı **(TF-IDF)**, bir kelimenin bir metin belgesi için ne kadar önemli olduğunu gösteren sayısal bir istatistiktir. Metin madenciliğinde bir ağırlıklandırma faktörü olarak kullanılır. TF-IDF değeri bir kelimenin belgede görünme sayısıyla orantılı olarak artarken, kelimenin kullanım sıklığı ile dengelenmektedir. Gereksiz kelimelerin ayıklanmasında başarılı sonuçlar veren bir yöntem olan TF-IDF, metin özetleme ve sınıflandırma alanlarında sıklıkla kullanılmakta ve belge ve kelime sayısının çok olduğu büyük veri setlerinde daha iyi sonuçlar vermektedir [49]. TF-IDF yönteminde iki öğe bulunur: TF - j belgesindeki i teriminin terim sıklığı ve IDF - i teriminin ters belge sıklığı [52]. TF-IDF, değeri Denk. 1'de gösterildiği gibi hesaplanmaktadır (3.1). Eşitlikteki N değeri eğitim setindeki tüm dokümanların sayısını belirtmektedir.

$$a_{ij} = tf_{ij} idf_i = tf_{ij} \times log_2\left(\frac{N}{df_i}\right) \quad (3.1)$$

Word2Vec, kelime temsili için sunulan ön eğitimli bir algoritmadır. Kelimeleri vektör uzayında ifade etmeye çalışan tahmin temelli bir modeldir ve CBOW (Continous Bag of Words) ve Skip-Gram alt yöntemleriyle çalışmaktadır [53]. Bu iki yöntemden en uygun olanı seçilerek kelime demetlerinin vektörel temsili gerçekleştirilir [54]. Girdi, çıktı ve gizli katmandan oluşan ve bir yapay sinir ağı olan Word2Vec için önemli hiperparametrelerden biri olan window_size hedef kelimenin her iki tarafında kaç tane kelime olabileceğini belirtmektedir [55]. Modelde CBOW mimarisi bağlama dayalı olarak mevcut kelimeleri tahmin ederken, Skip-Gram mimarisi ise hâlihazırda verilen kelimenin etrafındaki kelimeleri tahmin etmektedir [56]. Başka bir deyişle CBOW modeli bir kelimenin bağlamındaki diğer kelimelerden yola çıkarak o kelimenin olasılıklı vektörünü tahminlerken, Skip-Gram modeli bunun tersi bir yaklaşım ile bir kelimenin vektöründen yola çıkar ve o kelimenin bağlamındaki diğer kelimeleri tahmin etmeye çalışır. CBOW ve Skip-Gram mimarileri Şekil 3.2'de görülebilir.



Şekil 3.2: Word2Vec CBOW ve Skip-Gram model mimarisi [56]

Bir $w_1, w_2, ..., w_t$ eğitim kelimeleri dizisi verildiğinde, CBOW modelinden ortalama log olasılığını maksimize etmesi istenir (Denk. 3.2).

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p\left(w_t \mid w_{t+j}\right) \tag{3.2}$$

Fakat Skip-Gram modelinden ortalama log olasılığını maksimize etmesi istenir (Denk. 3.3).

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p\left(w_{t+j}\,|w_t\right) \tag{3.3}$$

Burada c eğitim içeriğinin boyutudur [57]. Sonuç olarak CBOW modelinde bir kelimenin bağlamından, başka bir deyişle yakınındaki kelimelerden hareketle o kelime tahmin edilmeye çalışılırken; Skip-gram modelinde bir kelimeden hareketle o kelimenin bağlamındaki diğer kelimeler tahminlenir.

Başka bir ön eğitimli model olan Kelime Temsili için Global Vektörler (Global Vectors for Word Representation, **GloVe**) ise, metinsel veriden toplanan su ve buhar benzeri ikili sözcüklerin birlikte oluşum istatistikleri üzerinde gerçekleştirilir. Ortaya çıkan kelime vektörleri, Word2Vec paketinde olduğu gibi kelime analojisi görevlerinde çok iyi performans göstermektedir [58]. Kelime düzeyinde çalışan Word2Vec'ten farklı olarak karakter düzeyinde çalışan **FastText**, 157 farklı dilde, Common Crawl ve Wikipedia üzerinde, 300 boyutlu, karakter n-gram uzunluğu 5, pencere boyutu 5 ve 10 negatif olan konum ağırlıkları ile CBOW kullanılarak eğitilmiş kelime vektörleri sunmaktadır [59].

Kelime temsilinde farklı bir yaklaşım ise bağlamsal kelime temsilidir. Bu alanda, 2018 yılında Peters ve ark. [60] tarafından yayınlanan bir çalışma ile hem kelime kullanımının sözdizimi ve semantik gibi karmaşık özelliklerini hem de bu özelliklerin farklı bağlamlarda nasıl değiştiğini modelleyen yeni bir bağlamsallaştırılmış kelime temsili türü sunulmuştur. **ELMo** (Embeddings from Language Models) adını verdikleri bu model ile doğal dil işleme alanındaki soru cevaplama, metinsel gereklilik ve duygu analizi gibi problemlerin çözümünde büyük bir adım atılmıştır. ELMo kelime temsilleri, biLM'deki (bidirectional language model - çift yönlü dil modeli) ara katman temsillerinin göreve özel bir kombinasyonu olarak açıklanabilir. Her bir belirteç (token) için t_k, bir L-katmanı çift yönlü dil modeli bir dizi 2L +1 temsilini hesaplar (Denk. 3.4).

$$\begin{aligned} R_k &= \{\mathbf{x}_k^{LM}, \overrightarrow{\mathbf{h}}_{k,j}^{LM}, \overleftarrow{\mathbf{h}}_{k,j}^{LM} \mid j=1,\ldots,L\} \\ &= \{\mathbf{h}_{k,j}^{LM} \mid j=0,\ldots,L\}, \end{aligned} \tag{3.4}$$

Burada, $h_{k,0}^{LM}$ belirteç katmanıdır ve her bir biLSTM katmanı için,

$$h_{k,j}^{LM} = [\overrightarrow{\mathbf{h}}_{k,j}^{LM}; \overleftarrow{\mathbf{h}}_{k,j}^{LM}] \quad (3.5)$$

eşitliği bulunur. Bir aşağı akış modelinde ELMo, *R*'deki tüm katmanları tek bir vektöre daraltır:

$$ELMo_k = \mathrm{E}(R_k; \Theta_e) \quad (3.6)$$

En basit durumda, ELMo sadece üst katmanı seçer. Modelde tüm biLM katmanlarının göreve özel ağırlıklandırması hesaplanmaktadır:

$$\mathbf{ELMo}_k^{task} = E(R_k; \Theta^{task}) = \gamma^{task} \sum_{j=0}^{L} s_j^{task} \mathbf{h}_{k,j}^{LM} \quad (3.7)$$

Bu denklemde s^{task} softmax-normalleştirilmiş ağırlıklarını temsil eder ve γ^{task} skaler parametresi, görev modelinin tüm ELMo vektörünü ölçeklendirmesini sağlamaktadır. ELMo modeli ile biLM katmanlarının bağlam içindeki kelimeler hakkında farklı sözdizimsel ve anlamsal bilgileri verimli bir şekilde kodlaması sağlanmış ve genel görev performansı iyileştirilmiştir.

Bağlamsal kelime temsili sunan başka bir model olan BERT ise, 2019 yılında tanıtılmıştır [61]. Google arama sonuçları ve Google Çeviri gibi alanlarda da kullanılan BERT, bir girdi dizisindeki her kelimenin tam bağlamını anlama yeteneğine sahiptir. Fakat toplu cümle temsillerinde kodlanan örtük bilgi, yine de bağlamdan bağımsız bir modelinkiyle karşılaştırılabilir. Ayrıca, BERT'in cümle seviyesindeki özellikleri tek kelimelik gömmelerde bile kodlayabildiği ve cümle temsilleriyle ulaşılandan daha üstün sonuçlar elde edebildiği görülmektedir [62].

OpenAI ekibi tarafından yürütülen çalışmada ise, insanların genellikle bir dil görevini yalnızca birkaç örnekten yola çıkarak veya basit talimatlarla gerçekleştirebildiğine ve bunun, mevcut NLP sistemlerinin hala yapmakta güçlük çektiği bir durum olduğuna dikkat çekilmiştir. Çalışmalarında, önceki dil modellerinden 10 kat daha fazla, yani 175 milyar parametreye sahip bir otoregresif dil modeli olan **GPT-3**'ü eğitmişlerdir. GPT-3 çeviri, soru yanıtlama, sözcükleri çözme görevleri, yeni bir sözcüğü bir dizide kullanma veya akıl yürütme gerektiren çeşitli görevler üzerinde güçlü performans elde etmiştir. İnsan değerlendiricilerin, GPT-3'ün oluşturduğu haber makalelerini, insanlar tarafından yazılan makalelerden ayırt etmekte zorlandıkları ortaya koyulmuştur [63]. Sonuç olarak, son dönem kelime temsil yöntemlerinde bağlamsal temelli

yaklaşımların öne çıktığı, genel olarak daha iyi performans sergileyen bu modellerin, doğal dil işleme alanındaki gelişmeleri bir basamak ileriye taşıdığı söylenebilir. BERT ve OpenAI GPT'ye ilerleyen bölümlerde tekrar değinilecektir.

3.3 Metin Sınıflandırma Yöntemleri

İnternet teknolojisindeki gelişmeler ile birlikte, elektronik dokümanların sayısında hızlı bir artış olmuş ve günümüzde de bu artış daha da hızlanarak devam etmektedir. Devasa miktarlardaki metin içerisindeki verinin düzenlenmesi, sınıflandırılması ya da bu veriden bilgi çıkarımı için birçok yöntem işe koşulmaktadır. Bu bölümde bahsedilen yöntemler metin madenciliği bağlamında ele alınarak; makine öğrenmesi modelleri, derin öğrenme modelleri ve transformer tabanlı dil modelleri başlıkları altında incelenecektir.

3.3.1 Makine Öğrenmesi Modelleri

Makine öğrenmesi, bilgisayarların doğrudan programlanmadan "öğrenmesini" sağlamayı amaçlayan bir bilgisayar bilimi dalıdır [64]. Metin sınıflandırma için denetimli, denetimsiz veya yarı denetimli olmak üzere birçok farklı makine öğrenmesi modeli kullanılmaktadır. Bu bölümde metin madenciliğinde kullanılan bazı denetimli sınıflandırma yöntemleri üzerinde durulacaktır.

3.3.1.1. Lojistik Regresyon

Regresyon bir denetimli öğrenme yaklaşımıdır. Doğrusal Regresyon sürekli değişkenin tahminlemesini yaparken, Lojistik Regresyon kategorik değişkenin tahminlenmesinde kullanılabilir. Genel olarak lojistik regresyon veri kümesindeki iki kategoriden biri için hangi sınıfa ait olacağının olasılığını hesaplar [65].

LR modeli [66], girdi değeri x_i, ağırlık katsayısı β_i ve bias değeri olarak ifade edildiğinde Denk. 3.8'deki gibi ifade edilmektedir. z değeri, Sigmoid fonksiyonundan geçirilerek çıktı değeri olan y_p değerine ulaşılır (Denk. 3.9).

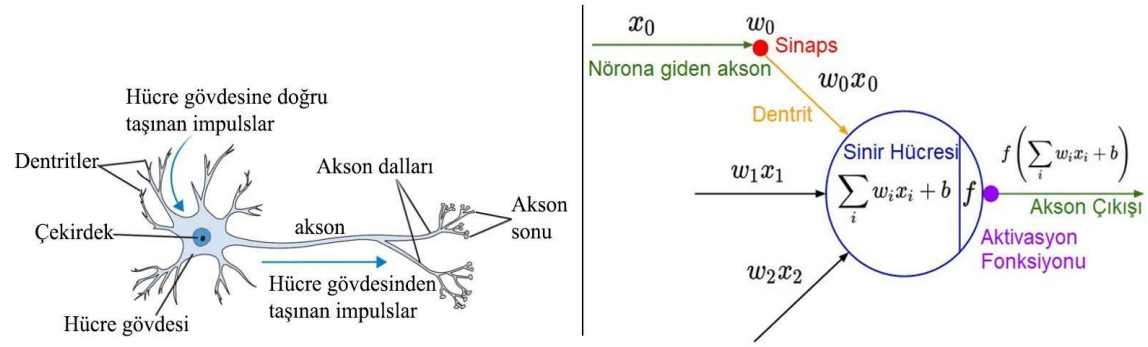
$$z = \sum_{i=0}^{n} \beta_i . x_i + b \tag{3.8}$$

$$y_p = \frac{1}{1+e^z} \tag{3.9}$$

Bu yöntem uygulama kolaylığı, verimlilik gibi avantajlara sahiptir. Lojistik regresyon, bağımlı değişken kategorik bir yapıda olduğunda kullanılır. Eğitim ortamına ilişkin problemlerde başarılı/başarısız, sağlıkla ilgili problemlerde ise hastalığın varlığı/yokluğu gibi tahminlerde kullanılabilir [67]. Literatür incelendiğinde, lojistik regresyonun metin sınıflandırma için sıklıkla kullanıldığı görülmektedir. Bu araştırmalardan biri Indra, Wikarsa ve Turang [68] tarafından yürütülmüş, araştırmacılar Twitter'dan topladıkları verileri sağlık, müzik, spor ve teknoloji başlıkları altında lojistik regresyon kullanarak sınıflandırmışlardır. Araştırmada eğitim verisi olarak kullanılan her konu için 1800 adet etiketli tweet bulunmaktadır. Ön işleme aşamasında, URL'lerin kaldırılması, noktalama işaretleri ve durdurma sözcükleri, simgeleştirme ve köklerine ayırma dahil olmak üzere çeşitli işlemler yapılmıştır. Daha sonra tweetler Bag of Words kullanarak otomatik olarak özellik vektörüne dönüştürülmüştür. Eğitimli sınıflandırıcı daha sonra her konu için 450 tweet ile 1800 tweet kullanılarak değerlendirilmiş ve sınıflandırma doğruluğunun %92 olduğu görülmüştür.

3.3.1.2. Yapay Sinir Ağları

Yapay sinir ağlarının (YSA), anlaşılabilmesi için nörobiyolojik mimarinin özünün anlaşılması gerekir. Biyolojik bir beyinde her nöron binlerce girdi almaktadır. Bir nöron, bir giriş/çıkış cihazı olarak hayal edilebilir. Ayrıntılar değişiklik gösterse de, nöronlar temel olarak darbe kodlu analog bilgileri iletir. Bu darbe kodlu sistemin girdi/çıktı ilişkisi basit bir sigmoiddir. Sigmoid özelliği, nöral hesaplama özellikleri için çok önemlidir [69]. Girdi, ara katman ve çıktı katmanından oluşan yapay sinir ağlarında, düğüm adı verilen bu yapay nöronlar temel yapıyı oluşturmaktadır. Nöronun basitleştirilmiş bir matematiksel modelinde, sinapsların etkileri, ilişkili giriş sinyallerinin etkisini modüle eden bağlantı ağırlıkları ile temsil edilir ve nöronların sergilediği doğrusal olmayan özellik, bir transfer fonksiyonu ile temsil edilir. Nöron impulsu, transfer fonksiyonu tarafından dönüştürülen giriş sinyallerinin ağırlıklı toplamı olarak hesaplanır. Yapay bir nöronun öğrenme yeteneği, ağırlıkların seçilen öğrenme algoritmasına göre ayarlanmasıyla elde edilir [70].



Şekil 3.3: Biyolojik bir nöron ile yapay bir nöron arasındaki benzerlik [71]

3.3.1.3. K-En Yakın Komşu (KNN)

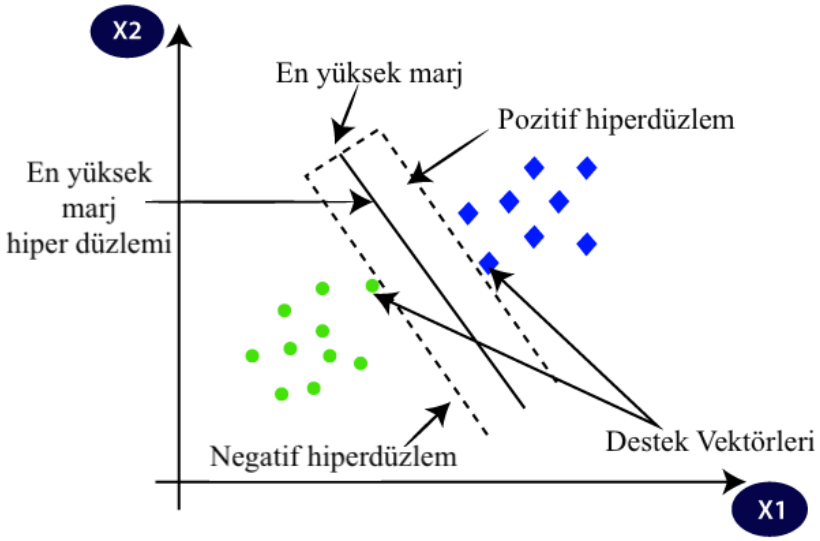
Temel olarak öğeler arasındaki benzerliğe göre çalışan KNN, mesafe fonksiyonunu kullanmaktadır. Kendisine en yakın yani en çok benzeyen K tane örneğe dayanarak bir sonuca varan bir öğrenme algoritmasıdır. Burada K sayısı komşu sayısını ifade etmektedir ve olası bir eşitliğin önüne geçilebilmesi için özellikle sınıflandırma problemlerinde, tek sayı olarak belirlenir. Kendisine en yakın örneği bulabilmesi için çeşitli matematiksel formüllerle uzaklık hesabı yapılır. KNN için en yaygın olarak kullanılan mesafe fonksiyonlarından biri Minskowski uzaklık fonksiyonu Denk. 3.10'da verilmiştir [72]. Burada x ve z gözlem değerleri arasındaki mesafe p sayıda değişken göz önüne alınarak hesaplanmaktadır.

$$\mathrm{dist}(\mathbf{x}, \mathbf{z}) = \left(\sum_{r=1}^{d} |x_r - z_r|^p \right)^{1/p} \qquad (3.10)$$

Daha doğru sonuçlar elde edilebilmesi için özniteliklerin tamamı belli bir aralığa normalize edilmektedir. Böylece her özniteliğin sonuca eşit derecede etki etmesi sağlanır. Ağırlıklandırılmış yaklaşımda ise daha yakın komşular sonucu daha ağırlıklı olarak etkilemektedir. Yalın bir model olması nedeniyle çok yaygın olarak kullanıldığı görülmektedir fakat KNN ile sınıflandırma süresi çok uzundur ve K'nın optimal değerini bulmak zordur. Genel olarak, seçilecek en iyi k alternatifi verilere bağlıdır [73].

3.3.1.4. Destek Vektör Makineleri

Destek Vektör Makineleri (Support Vector Machine, DVM) daha çok küçük ve orta ölçekli veri setlerinde, genellikle sınıflandırma problemlerinde kullanılır. DVM, gerçek ve sahte kredi kartı faaliyetlerini belirlemek, el yazısından rakamları inceleyerek tanımak gibi problemlerin çözümünde etkili sonuçlar ortaya koymaktadır [74]. SVM'nin öncelikli hedefi veri kümeleri arasında hiper düzlemler bulmaktır ve optimum hiperdüzlemleri tespit etmek için maksimum kenar boşlukları bulunur [75].



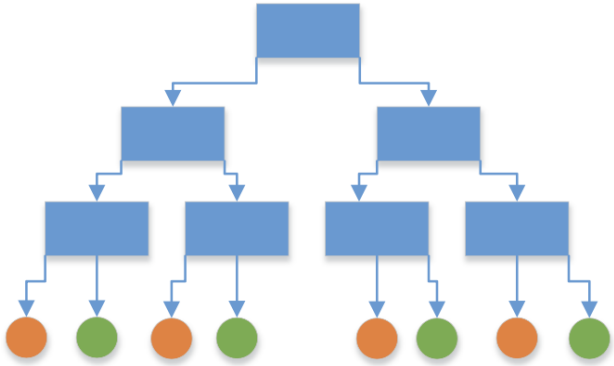
Şekil 3.4: Destek Vektör Makineleri [75]

DVM, metin sınıflandırması da dahil olmak üzere birçok uygulama için en başarılı sınıflandırma yöntemlerinden biri olarak kabul edilmiştir. Eğitimin öğrenme yeteneği ve hesaplama karmaşıklığı, özellik uzayının boyutundan bağımsız olabilse de, hesaplama karmaşıklığının azaltılması, metin sınıflandırmasının pratik uygulamalarında çok sayıda terimi verimli bir şekilde ele almak için önemli olduğu görülmektedir [76].

3.3.1.5. Karar Ağaçları

Karar ağaçları modelinde sınıflandırma yapılırken veri, tıpkı bir ağacın yapısında olduğu gibi dallanarak değerlendirilmektedir (Şekil 3.5). Ağaç yapısı oluşturulurken hangi özniteliğin ağırlığının en fazla olması isteniyorsa o öznitelik köke yerleştirilir. Köke yerleştirilecek öznitelik, bazı metrikler yardımıyla hesaplanmaktadır. Kullanımı oldukça kolay olan bu model, özellikle varılan sonuca ne şekilde ve nasıl varıldığının

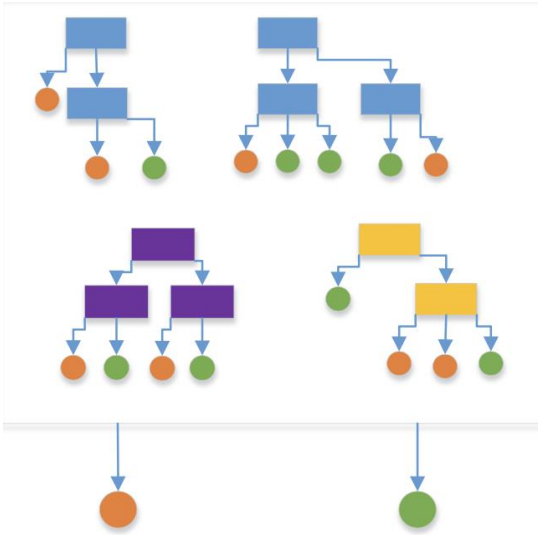
açık olması gereken durumlarda kullanılmaktadır. Güçlü ve aynı zamanda esnek olan bu modelin uygulanması için çok az varsayım gerekmesine rağmen yüksek kaliteli sonuçlar elde edilmektedir [77].



Şekil 3.5: Karar ağaçları modeli [75]

3.3.1.6. Rastgele Orman

Rastgele ormanlar, her bir ağacın bağımsız olarak örneklenen ve ormandaki tüm ağaçlar için aynı dağılıma sahip rasgele bir vektörün değerlerine bağlı olduğu ağaç tahmincilerinin bir kombinasyonudur (Şekil 3.6). Bir topluluk algoritması olan bu modelde, genelleme hatası ormandaki ağaç sayısı arttıkça sınırlanır. Bir ağaç sınıflandırıcı ormanının genelleme hatası, ormandaki ağaçların gücüne ve aralarındaki korelasyona bağlıdır. Dâhili tahminler hatayı, gücü ve korelasyonu izler ve bunlar, bölmede kullanılan öznitelik sayısının artmasının yanıtını göstermek için kullanılır. Değişkenin önemini ölçmek için dâhili tahminler de kullanılır [78].



Şekil 3.6: Rastgele orman modeli [75]

3.3.2 Derin Öğrenme Modelleri

Metin sınıflandırmada oldukça iyi sonuçlar veren derin öğrenme modelleri, yapay sinir ağları kullanılarak oluşturulmuştur [79]. Metin sınıflandırmalarında Derin Nöral Ağlar, Yinelenen Nöral Ağlar ve Evrişimsel Nöral Ağlar (Convolutional Neural Networks, CNN) gibi modellerin kullanılabildiği görülmektedir. Goldberg [80], doğal dil işleme için derin öğrenme konusunda, genel olarak sinir ağlarının, özellikle önceden eğitilmiş kelime gömmeleriyle kullanıldığında, klasik doğrusal sınıflandırıcılardan daha iyi performans gösterdiğini belirtmektedir.

CNN resim sınıflandırma için kullanılan başarılı bir model olmasının yanı sıra metin sınıflandırma görevleri için de kullanılmaktadır. Johnson ve Zhang [81] yürüttükleri araştırmada, CNN'nin, geleneksel n-gram çantası yaklaşımı veya kelime vektörü CNN'den farklı olarak, küçük metin bölgelerinin doğrudan gömülmesi yoluyla metin kategorizasyonu için kelime sırasının etkin kullanımı için alternatif bir mekanizma sağladığını göstermişlerdir. CNN kullanarak duygu sınıflandırması ve konu sınıflandırmasında oldukça iyi bir performans elde etmişlerdir. Lai ve ark. [82] ise, yürüttükleri çalışmada metin sınıflandırması için evrişimli sinir ağı kullanmışlardır. Modelde, geleneksel pencere tabanlı nöral ağlara kıyasla çok daha az gürültüye yol açan kelime temsillerini öğrenirken mümkün olduğunca bağlamsal bilgileri yakalamak için tekrarlayan bir yapı uygulamışlardır. Metinlerdeki temel bileşenleri yakalamak için metin sınıflandırmasında hangi kelimelerin kilit rol oynadığını otomatik olarak karara varan bir maksimum havuzlama katmanı da kullanmışlardır. Elde ettikleri deneysel sonuçlar, modelin oldukça iyi performans elde ettiğini göstermektedir.

Kısa metinlerin sınıflandırılması, uzun belgelere kıyasla daha zordur. Çünkü paragraflardan veya belgelerden farklı olarak kısa metinler, yeterli bağlamsal bilgiye sahip olmadıkları için daha belirsizdir. Chen ve ark. [83], bu zorluğun üstesinden gelmek için derin nöral ağları kullanarak bir model geliştirmişlerdir. Kısa metinlerin anlamsal temsilini geliştirmek için YAGO ve Freebase gibi açık bilgi tabanlarının yardımıyla anlamsal sunumu zenginleştirmek yoluyla kavramsal bilgiyi derin sinir ağlarına dâhil etmişlerdir. STCKPA (Short Text Classification with Knowledge Powered Attention) adını verdikleri modelleri geleneksel sinir ağı modellerinden daha iyi performans göstermektedir. Yapılan çalışmaların gösterdiği gibi derin öğrenme modelleri, metin sınıflandırma problemleri için etkili çözümler sunabilmektedir. Fakat

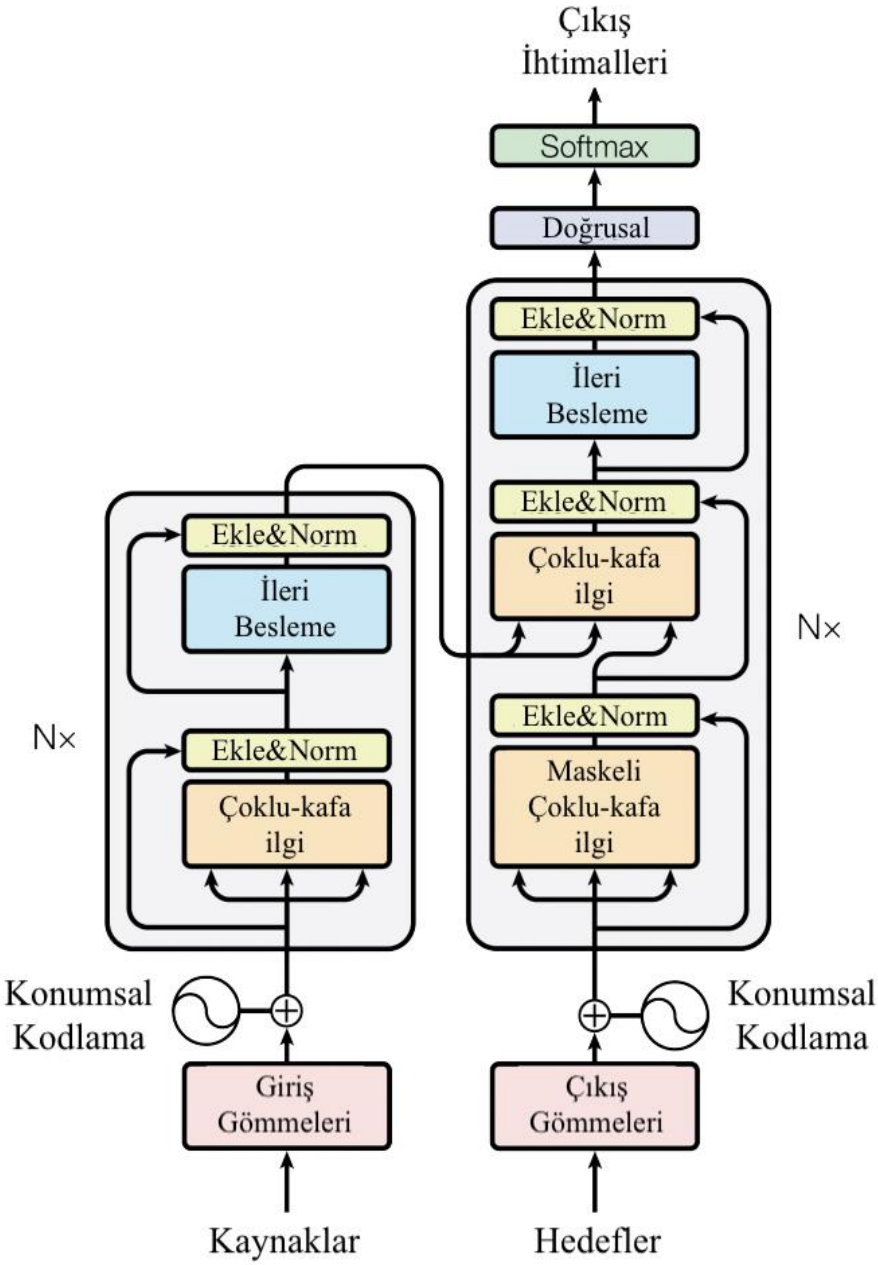
metin madenciliğinde transformer tabanlı dil modelleri ile farklı bir boyuta geçildiği görülmektedir.

3.3.3 Transformer Tabanlı Dil Modelleri

Transformer mimarisi [84], yayınlandığı dönemde dil işleme görevlerinde en iyi performansı sunan LSTM ve RNN (Recurrent Neural Networks - Yinelenen Nöral Ağlar) modellerindeki uzun mesafeli bağımlılıkları çözmek için bir dikkat mekanizması öneren bir model olarak karşımıza çıkmaktadır. Model bir kodlayıcı (encoder) ve çözücü (decoder) yapısına sahiptir. Kodlayıcı yapısı 6 eş katmandan oluşan kodlayıcı yığınını, çözücü ise yine aynı sayıdaki kod-çözücü yığınını içerir. Bu yığındaki her kodlayıcı bir ileri beslemeli sinir ağı (feedforward neural network), katman normalizasyonu (layer normalization) ve çok-kafalı ilgi mekanizması içerir ve girdi metnin her bir kelimesi için bir gömme vektörü oluşturur (Şekil 3.7).

Bu gömme vektörleri bir sonraki katmana iletilir ve her bir katman mevcut gömülü kelime vektörünü kullanarak yeni bir gömülü kelime vektörü oluşturur. Kodlayıcı ve çözücü katmanlar çoklu-kafa ilgi mekanizması yoluyla birbirleri ile etkileşim halindedir. Modeldeki çoklu-kafa ilgi mekanizması, bir sorguyu ve bir dizi anahtar-değer çiftini; sorgunun, anahtarların, değerlerin ve çıktının tümünün vektör olduğu bir çıktıya eşlemek olarak tanımlanabilir. Çıktı, değerlerin ağırlıklı toplamı olarak hesaplanır ve burada her bir değere atanan ağırlık, karşılık gelen anahtarla sorgunun bir uyumluluk işlevi tarafından hesaplanır. Softmax katmanı ise, bir dağılım oluşturmak için benzerlik ölçülerini normalleştirir ve dikkat ağırlıklarının doğru bir şekilde hesaplanmasını sağlar.

Öz-ilgi katmanı, modelin bir kelimenin ilgili olduğu diğer kelimeleri anlamasını sağlamaktadır. Örneğin "Köpek karşıya geçmedi çünkü o çok yorgundu." cümlesinde "o" zamiri ile "köpek" kelimesinin ilişkilendirmesi öz-ilgi sayesinde gerçekleşmektedir [85]. Sonuç olarak, bu mekanizma ile her kelime vektörü diğer tüm kelime vektörleri ile birlikte ele alınır ve kelime vektörlerinin birbirlerine olan bağımlılıkları hesaplanır. Katman normalizasyonu, çıktıları ölçeklendirmektedir. Daha sonra ileri beslemeli sinir ağı, çoklu-kafa ilgi mekanizması ve katman normalizasyonu çıktılarını işler. Modeldeki tüm alt katmanlar ve gömülü katmanlar, d_{model}= 512 boyutunda çıktılar üretmektedirler.



Şekil 3.7: Transformer model mimarisi [84]

Öz-ilgi katmanları, Denk. 3.11'de gösterildiği gibi üç doğrusal dönüşüm uygulayarak belirteçleri anahtar (K), sorgu (Q) ve değer (V) matrislerine dönüştürmektedir. X, verilen girdileri, W ise eğitilebilir ağırlık matrislerini göstermektedir. K ve Q'nun iç çarpımı dikkat puanını oluşturur, V orijinal belirteci temsil eder. Bu dikkat ağırlıkları, softmax fonksiyonu kullanılarak hesaplanmaktadır. Dikkat puanının hesaplanması Denk. 3.12'de gösterilmektedir [75].

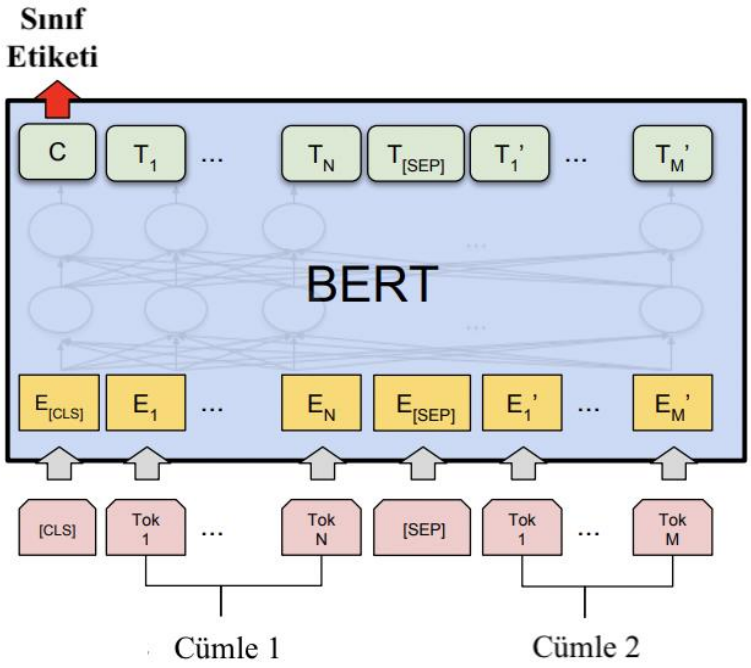
$$V = W^V X,\ K = W^k X,\ Q = W^q X \tag{3.11}$$

$$Attention(Q, K, V) = softmax(\frac{QK^T}{d_k})V \quad (3.12)$$

Özet olarak, transformer tabanlı modeller, bağlamı anlamak için bir cümle içerisinde yer alan, kelimeler benzeri sıralı verilerdeki ilişkileri izlemektedir. Kelimeler arasındaki ilişki, cümlenin anlamını vermektedir. Birbirini etkileyen veya birbirine bağlı olan öğelerin arasındaki ilişkileri tespit etmek için kullanılan öz-ilgi adı verilen matematiksel yöntemler sayesinde bu ilişkilerin cebirsel bir haritasının çıkarılması ve böylece anlamın model tarafından öğrenilmesi sağlanmaktadır.

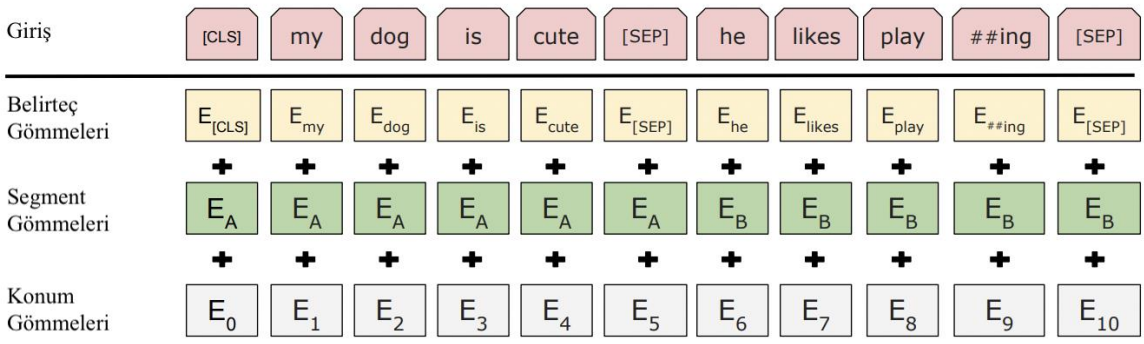
Transformers Tabanlı Çift Yönlü Kodlayıcı Temsilleri (Bidirectional Encoder Representation from Transformers, BERT) modeli 2018 yılında Google araştırmacıları tarafından sunulmuştur ve ilk aşamada $BERT_{BASE}$ ve $BERT_{LARGE}$ olmak üzere iki model boyutu üzerinde eğitilmiştir. BERT dil modeli, çift yönlü transformatör mimarisi içerir. Model denetimsiz öğrenme ile ön eğitime tabi tutulurken çift yönlü bir sinir ağı kullanılmaktadır. Bu modeli göreve özel olarak eğitmek için ise, fine-tuning olarak isimlendirilen ince ayar aşaması uygulanmaktadır. Model ön-eğitim boyunca etiketlenmemiş veriler ile farklı ön-eğitim görevleri üzerinde eğitilmekte, etiketli verilerle ince ayar yapılmaktadır. Bu sayede, sınırlı veri setleri ile bile yüksek başarımlı sonuçlar elde edilebilmekte ve zamandan tasarruf sağlanmaktadır [61]. $BERT_{BASE}$ modeli 768 gizli büyüklükte 12 katman ve 12 öz-dikkat başlığı, $BERT_{LARGE}$ modeli, 1024 gizli boyutta 24 katman ve 16 öz-dikkat başlığı içermektedir [86].

BERT modeli [61], giriş boyutu 512 olan sabit uzunluklu girdiler üzerinde çalışır. Orjinal cümle 30.000 belirteç dağarcığı bulunan WordPiece kullanılarak ek ve köklerine ayrılır. Her dizinin ilk simgesi özel bir sınıflandırma simgesi olan [CLS]'dir. İlk olarak, cümleler özel bir belirteç olan [SEP] ile ayrılır (Şekil 3.8). İkinci olarak, her simgeye A veya B cümlelerinden hangisine ait olduklarına dair öğrenilmiş bir gömme eklenir. Girdi yerleştirmeleri, belirteç yerleştirmelerinin, segmentasyon yerleştirmelerinin ve konum yerleştirmelerinin toplamıdır. (Şekil 3.9).



Şekil 3.8: Sınıflandırma görevinde cümle çiftleri [61]

BERT modeli geliştirilerek elde edilen RoBERTa [87], daha büyük miktarda veri ile eğitilmiştir ve daha iyi bir performans sunmaktadır. Ayrıca eğitim verileri rastgele sıralanarak modelin farklı koşullara daha iyi uyum sağlaması ve metinlerin cümle sınırlamasının kaldırılması ile daha uzun metinlerin anlaşılması sağlanmıştır. ALBERT modeli [88] ile ise GPU/TPU bellek sınırlamaları göz önüne alınarak, bellek tüketimini azaltmak ve BERT'nin eğitim hızını artırmak için iki parametre azaltma tekniği sunulmuştur. DistilBERT [89] ise BERT modelinin daha hızlı ve küçük bir modeli olarak karşımıza çıkmaktadır.



Şekil 3.9: BERT giriş gösterimi [61]

Alandaki en son çalışmalardan biri olan OpenAI Generative Pretrained Transformer 3 (GPT-3) 2020 yılında duyurulmuştur [63]. GPT-3 ince ayar gerektirmeden dil işleme

görevlerinde çok yüksek performans göstermiş ve kullanıcı sorgularına yüksek doğrulukta yanıtlar verebilmiştir. BERT modeli ile aralarındaki en önemli fark, BERT Transformer'ın çift yönlü öz-ilgi kullanırken GPT Transformer'ın her belirtecin yalnızca solundaki bağlama katılabildiği kısıtlı öz-ilgi kullanmasıdır. Ayrıca;

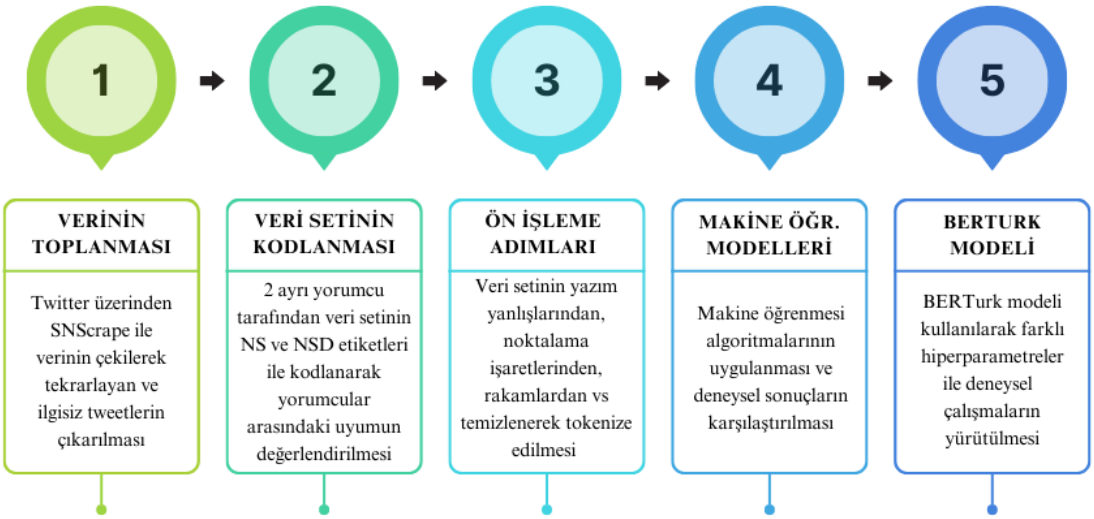
- GPT BooksCorpus (800 milyon kelime) üzerinde eğitilmiştir; BERT ise, BooksCorpus (800 milyon kelime) ve Wikipedia (2,5 milyar kelime) üzerinde eğitilmiştir.
- GPT, yalnızca ince ayar zamanında tanıtılan bir cümle ayırıcı ([SEP]) ve sınıflandırıcı belirteci ([CLS]) kullanır; BERT ise, ön eğitim sırasında [SEP], [CLS] ve cümle A/B yerleştirmelerini öğrenir.
- GPT, tüm ince ayar deneyleri için aynı öğrenme oranı olan 5e-5'i kullanmaktayken BERT ile geliştirme setinde en iyi performansı gösteren, göreve özel ince ayarlı bir öğrenme oranı seçilebilmektedir [61].

Bu bölümde doğal dil işleme kavramına değinilmiş ve özellikle metin sınıflandırma problemleri için kullanılan modeller üzerinde durulmuştur. İzleyen bölümde çalışmanın gerçekleştirilmesi için kullanılan materyal ve yöntem açıklanmaktadır.

Bölüm 4

Materyal ve Yöntem

Bu bölümde araştırmada kullanılan veri setinin oluşturulması için izlenen yöntem, veri setinin ön işleme adımları, kullanılan performans değerlendirme metrikleri, makine öğrenmesi ve derin öğrenme modellerinden sırasıyla bahsedilecektir. Bu araştırmada öncelikle veri toplanarak ilgisiz veya tekrarlayan tweetlerden temizleniş; daha sonra etiketleme işlemine geçilmiştir. Etiketleme işlemi tamamlandıktan sonra ön işleme adımları tamamlanmıştır. Daha sonra makine öğrenmesi ve BERTurk ile deneysel çalışmalar yürütülmüştür. Deneysel süreçte Python programlama dili ve bir masaüstü grafik kullanıcı arayüzü olan Anaconda Navigator kullanılmıştır. Eğitim ve testler MacBook Air (1,6 GHz Intel Core i5 işlemci, 4 GB 1600 MHz DDR3 bellek) ile gerçekleştirilmiştir. Veri seti eğitim ve test olarak ayrılırken makine öğrenmesi modelleri için 10 kat çapraz doğrulama kullanılmış, BERTurk için ise Numpy kütüphanesi içerisindeki random fonksiyonu kullanılarak %80 eğitim ve %20 test şeklinde bölünmüştür. Araştırmanın aşamaları Şekil 4.1'de sunulmuştur.



Şekil 4.1: Araştırmanın aşamaları

4.1 Veri Seti

Bu tez çalışmasında kullanılan veri seti araştırmacı tarafından oluşturulmuş ve Github platformu üzerinden açık erişime sunulmuştur[1].

4.1.1. Veri Setinin Oluşturulması

Bu çalışma kapsamında Twitter üzerindeki veriler kullanılmıştır. Öncelikle hangi anahtar kelimeler ile veri çekileceğine yönelik Twitter üzerinden genel bir inceleme yapılarak "mülteciler", "mülteci", "göçmenler", "ulkemdemulteciistemiyorum" gibi etiketlerle yapılan paylaşımların yoğunluk kazandığı tespit edilmiştir. Veri setinin elde edilmesi için SNScrape aracı kullanılmıştır. Bu modül birçok sosyal medya platformundan veri çekmek için kullanılabilmektedir. Python 3.8 sürümü ve sonrası için çalışmaktadır [90]. Belirlenen anahtar kelimeleri ile yayınlandığı tarih dâhil edilerek tweetler çekilmiştir. Çekilen 12.200 adet tweet tekrarlayan ve alakalı olmayan tweetlerden temizlenmiş, kalan 10659 tweet alanında uzman 2 kodlayıcı tarafından ayrı ayrı "Nefret söylemi değil" (0) ve "Nefret Söylemi" (1) etiketleriyle etiketlenmiştir. Verileri etiketleyen uzmanların önyargılardan arınması ve nefret söylemi tespitini yaparken kavram yanılgılarına düşmemeleri için etiketleme işlemi başlamadan önce iki farklı bilgilendirme toplantısı yapılmıştır. Bu toplantılarda nefret söylemi üzerine yapılan araştırmalar incelenmiş, literatüre dayanan bir tanım ortaya konmaya çalışılmıştır. Nefret söyleminin tespitinde dikkat çeken öğelerin yanı sıra, neyin nefret söylemi sayılamayacağı üzerinde de durulmuştur. Etiketlemeler ayrı ayrı yapılarak yorumcuların birbirinden etkilenmemesi sağlanmıştır. Elde edilen veri setine ilişkin bir kesit Tablo 4.1'de sunulmuştur.

Etiketleme işlemi bittikten sonra 10659 tweet için ayrı ayrı kodlanan veri setleri karşılaştırıldığında %91,73 oranında benzer kodlama yapıldığı görülmüştür. İki kodlayıcı arasındaki uyumun ölçülmesi için gerçekleştirilen Cohen's Kappa testi sonucunda Cohen's K değeri 0,835 ($p<0,01$) olarak hesaplanmıştır. Cohen'e [91] göre iki kodlamacı arasında "Güçlü düzeyde uyum" bulunduğu görülmüştür. Kodlayıcılar tarafından farklı etiketlenen 881 tweet silindikten sonra 4831 adet "Nefret söylemi

[1] Github linki: https://github.com/fegin80/Turkish-hate-speech-dataset.git

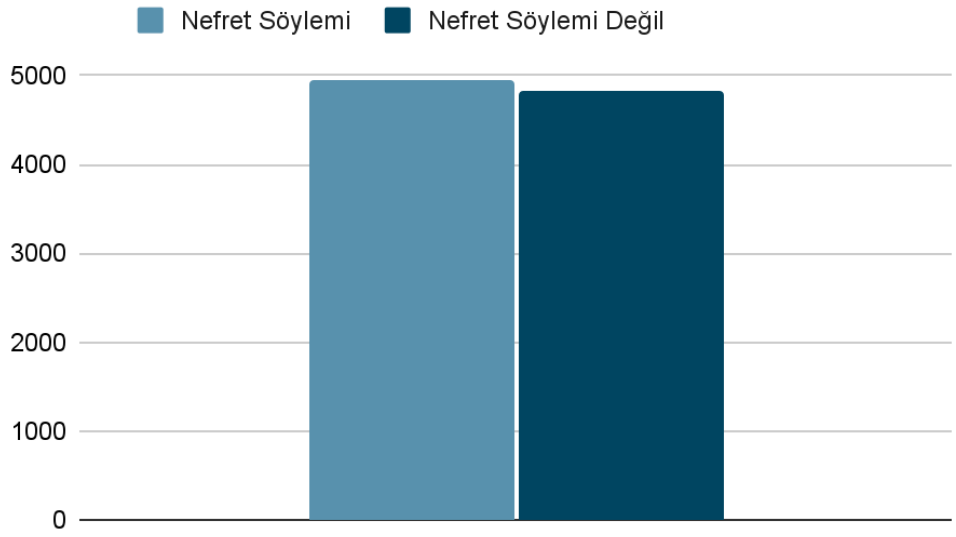
değil” ve 4947 adet “Nefret söylemi” etiketli toplam 9778 tweet içeren bir veri seti elde edilmiştir.

Tablo 4.1: Etiketli veri setinden bir kesit
(Nefret söylemi:1, Nefret Söylemi Değil:0)

Tweet	Sınıf Etiketi
Sığınmacılar, mülteciler adına ne denirse densin Türkiye için her zaman tehlikedir!	1
İstanbul'da bayramda mülteciler için ücretsiz ulaşım imkanı tanınmayacak.	0
Avrupaya göç eden Türkler ile mülteciler aynı sebeple ülkelerini terk etmedi	0
Bu mülteciler ve sığınmacılar gittikleri her yeri karıştırıyor!	1

4.1.2. Veri Setine İlişkin İstatistikler

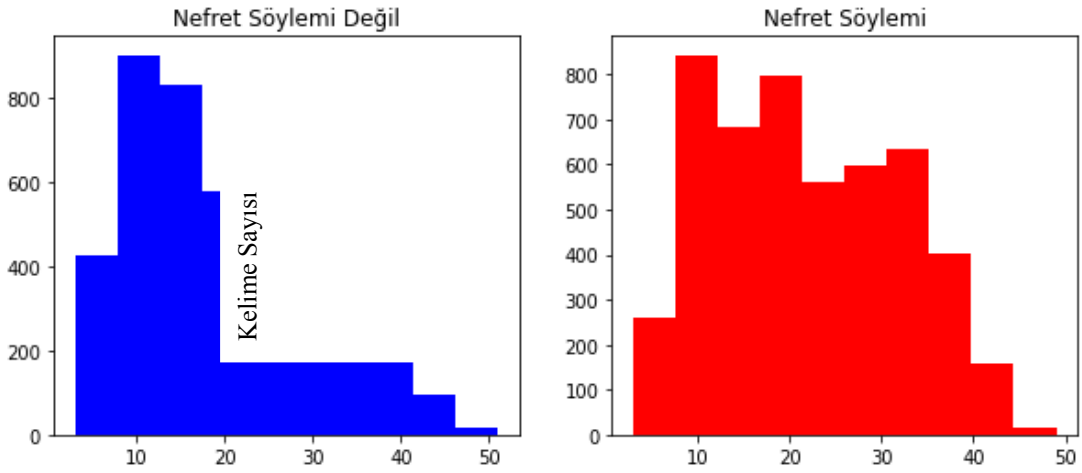
Veri setinin, 4831 adet “Nefret söylemi değil” ve 4947 adet “Nefret söylemi” etiketli toplam 9778 tweet’ten oluştuğu ve çok ayrık bir veri seti olmadığı görülmektedir (Şekil 4.2).



Şekil 4.2: Veri setinin dağılımı (Tweet sayısı)

Tweetler için kullanılan kelime sayıları incelendiğinde yoğunlukla hem NSD için hem NS için 10-20 kelime aralığında olduğu görülmektedir (Şekil 4.3). NS içeren

tweetlerin ortalamada daha fazla karakter, kelime ve tekil kelime içerdiği görülmektedir.



Şekil 4.3: Tweet başına kelime sayıları

Tablo 4.2: Tweetlerin ortalama karakter ve kelime sayıları
(NS: Nefret Söylemi, NSD: Nefret Söylemi Değil)

	Etiket	
	NS	NSD
Karakter Sayısı	179.9	173.09
Kelime Sayısı	21.891	21.316
Benzersiz Kelime Sayısı	21.065	20.537

Veri setini daha iyi anlayabilmek için karakter, kelime ve benzersiz kelime sayıları (Tablo 4.2) ile unigrams, bigrams ve trigrams sıklıkları çıkarılmıştır (Tablo 4.3, Tablo 4.4, Tablo 4.5, Şekil 4.3, Şekil 4.4, Şekil 4.5, Şekil 4.6). Veri setinde NS veya NSD olarak etiketli verilerin karakter ve kelime sayılarının oldukça yakın olduğu görülmüştür. Ayrıca benzersiz kelime sayıları da nefret söylemi içeren ve içermeyen tweetlerde birbirine yakın çıkmıştır.

Tablo 4.3: Veri setine ait unigram sıklıkları

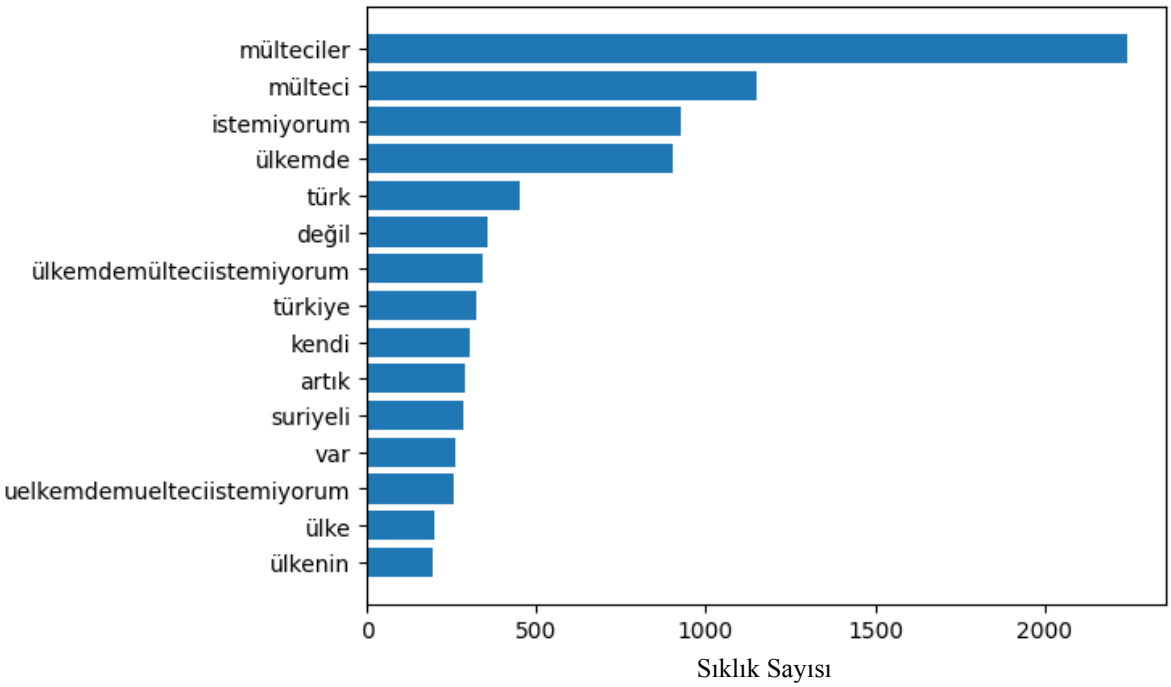
Sıra No	Frekans	Unigram
1	7639	mülteciler
2	2194	mülteci
3	1311	istemiyorum
4	1258	ülkemde
5	951	türkiye
6	909	değil
7	840	türk
8	809	var
9	751	suriyeli
10	557	ülkemdemülteciistemiyorum

Tablo 4.4: Veri setine ait bigram sıklıkları

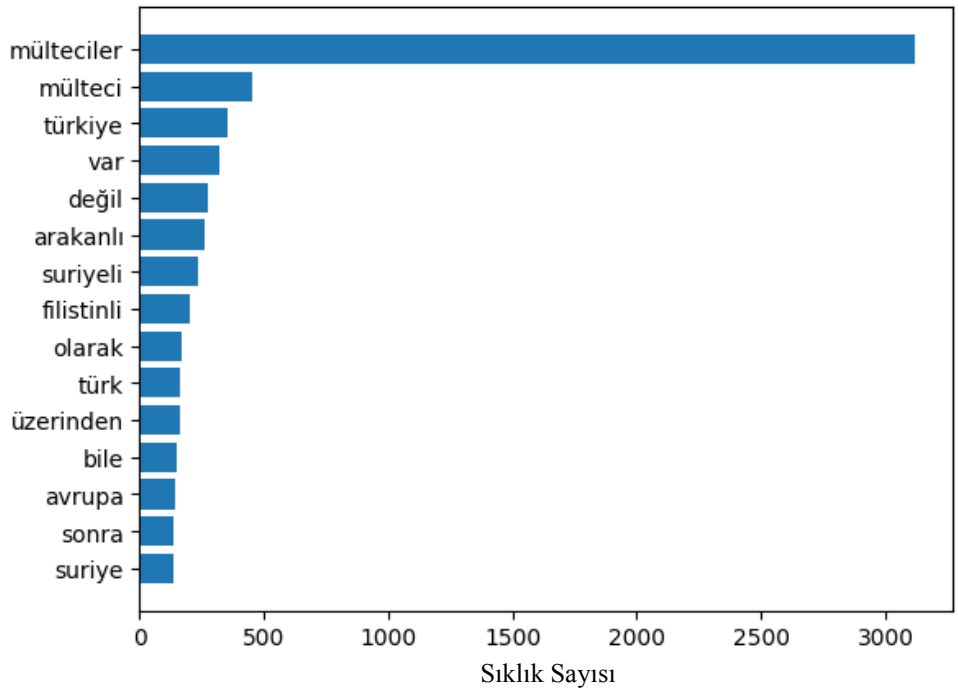
Sıra No	Frekans	Unigram
1	1083	mülteci istemiyorum
2	1073	ülkemde mülteci
3	259	arakanlı mülteciler
4	203	suriyeli mülteciler
5	198	filistinli mülteciler
6	177	mülteciler üzerinden
7	165	mülteciler konusunda
8	163	mülteciler yüzünden
9	146	mülteciler tarafından
10	130	türkiye ye

Tablo 4.5: Veri setine ait trigram sıklıkları

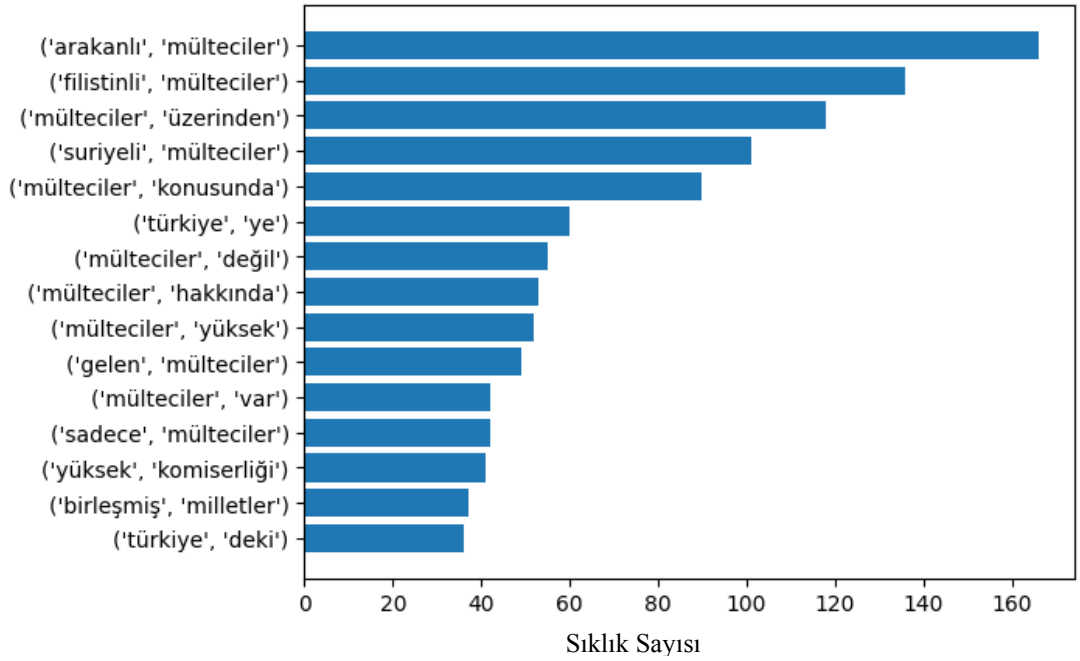
Sıra No	Frekans	Unigram
1	1059	ülkemde mülteci istemiyorum
2	59	mülteciler yüksek komiserliği
3	47	istemiyorum ülkemde mülteci
4	45	milletler mülteciler yüksek
5	44	birleşmiş milletler mülteciler
6	29	artık ülkemde mülteci
7	27	mülteciler tarafından öldürülen
8	27	mülteci istemiyorum bebek
9	27	bm mülteciler yüksek
10	25	mülteciler gelmeden önce



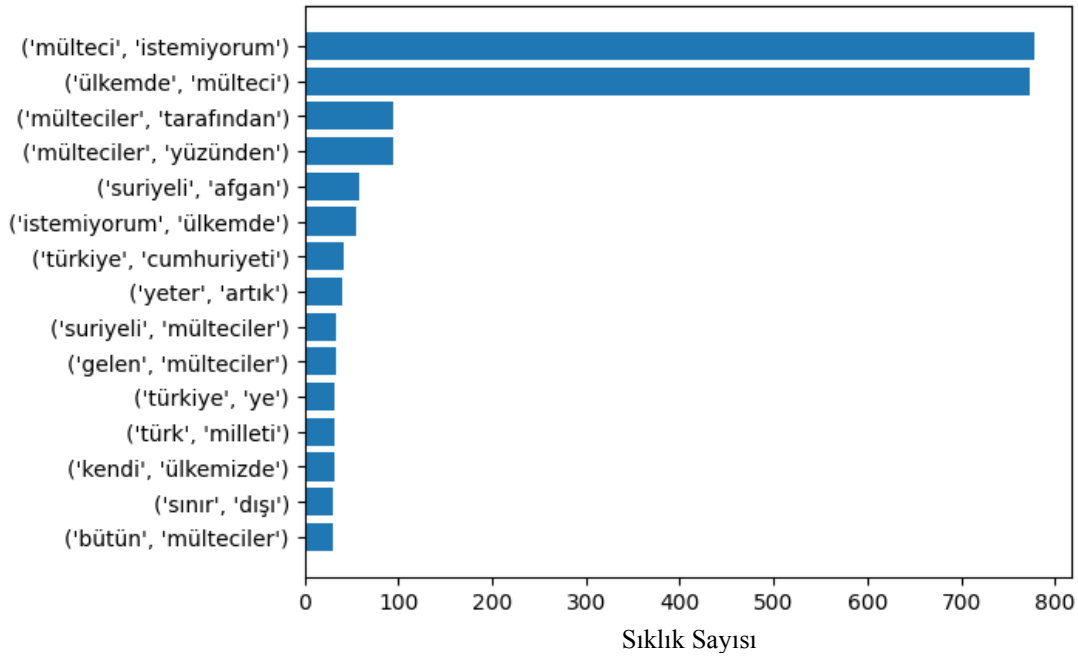
Şekil 4.4: NS olarak etiketlenen tweetler için en yüksek frekansa sahip unigramlar



Şekil 4.5: NSD olarak etiketlenen tweetler için en yüksek frekansa sahip unigramlar



Şekil 4.6: NS olarak etiketlenen tweetler için en yüksek frekansa sahip bigramlar



Şekil 4.7: NSD olarak etiketlenen tweetler için en yüksek frekansa sahip bigramlar

4.2 Veri Ön İşleme Yöntemleri

Verinin işlenebilmesi için bazı ön işleme adımlarına tabi tutulması gerekmektedir. Bu araştırmada, öncelikle yazım yanlışları düzeltilmiş, sonra büyük küçük harf değişiklikleri nedeniyle aynı kelimelerin farklı kelimeler olarak algılanmalarının önüne geçmek için tweet içerikleri küçük harfe çevrilmiştir. İzleyen adımlarda ise tweetler;

- özel karakterlerden,
- rakamlardan,
- noktalama işaretlerinden,
- fazla boşluklardan,
- linklerden,
- kullanıcı isimlerinden,
- Türkçe dolgu kelimelerden temizlenmiştir.

Dolgu kelimelerden temizlemek için NLTK kütüphanesi kullanılmış ve kütüphane içerisinde sunulan mevcut dolgu kelimeler silinmiştir. Temizlenen tweetler ek ve köklerine ayrılarak ön işleme adımları tamamlanmıştır.

4.3 Makine Öğrenmesi Modelleri ile Elde Edilen Sonuçlar

Bu araştırmada Word2Vec ve TF-IDF kelime gömme yöntemleri ile Lojistik Regresyon (LR), Karar Ağaçları, Destek Vektör Makineleri (DVM), Rastgele Orman ve Yapay Sinir Ağı (YSA) modelleri kullanılmıştır. Modellerin performansları doğruluk, F1-Ölçütü (F1-Score), kesinlik (precision), duyarlılık (recall) ve doğruluk (accuracy) ölçütleri ile değerlendirilmiştir. Sınıflandırma modellerinin değerlendirilmesinde kullanılan bu ölçütler karmaşıklık matrisi üzerinden açıklanmaktadır. Karmaşıklık matrisi, gerçek sınıf etiketleri ile tahmin edilen sınıf etiketlerinin karşılaştırılması sonucu elde edilir. Sınıf sayısı N ile gösterildiğinde, karmaşıklık matrisi NxN boyutunda bir matris olarak hesaplanmaktadır. Bu matriste satırlar gerçek sınıfları temsil ederken, sütunlar tahmin edilen sınıfları gösterir. Matrisin sol üst köşesi gerçek pozitifleri (True Positive - TP) temsil ederken, sağ üst köşesi yanlış pozitifleri (False Positive - FP), sol alt köşesi yanlış negatifleri (False Negative - FN) ve sağ alt köşesi gerçek negatifleri (True Negative - TN) temsil eder (Tablo 4.6).

Tablo 4.6: Karmaşıklık matrisi

Tahmin edilen sınıf Gerçek Sınıf	N	P
N	TN	NP
P	FN	TP

Sınıflandırma modelinin performansını ölçmek için kullanılan metrikler ise karmaşıklık matrisi temel alınarak şu şekilde hesaplanır [92]:

Doğruluk, nefret söyleminin doğru tahminlerinin toplam tahmine oranını vermektedir.

$$\text{Doğruluk} = (TP + TN) / (TP + TN + FP + FN) \quad (4.1)$$

Kesinlik, nefret söylemi olarak doğru tahminlenen etiketlerin nefret söylemi olarak tahminlenen tweet sayısına oranıdır.

$$\text{Kesinlik} = TP / (TP + FP) \quad (4.2)$$

Duyarlılık nefret söylemi olarak tahmin edilen etiketlerin nefret söylemi olan etiketlere oranını gösterir.

$$\text{Duyarlılık} = TP / (TP + FN) \quad (4.3)$$

F1-Ölçütü, kesinlik ve duyarlılık ölçütlerinin uyumlu ortalamasıdır (harmonic mean).

$$F1 = 2 * (\text{kesinlik} * \text{duyarlılık}) / (\text{kesinlik} + \text{duyarlılık}) \quad (4.4)$$

Deneysel süreçte, kullanılan makine öğrenmesi modelleri için seçilen parametreler ise Tablo 4.7'te gösterilmektedir.

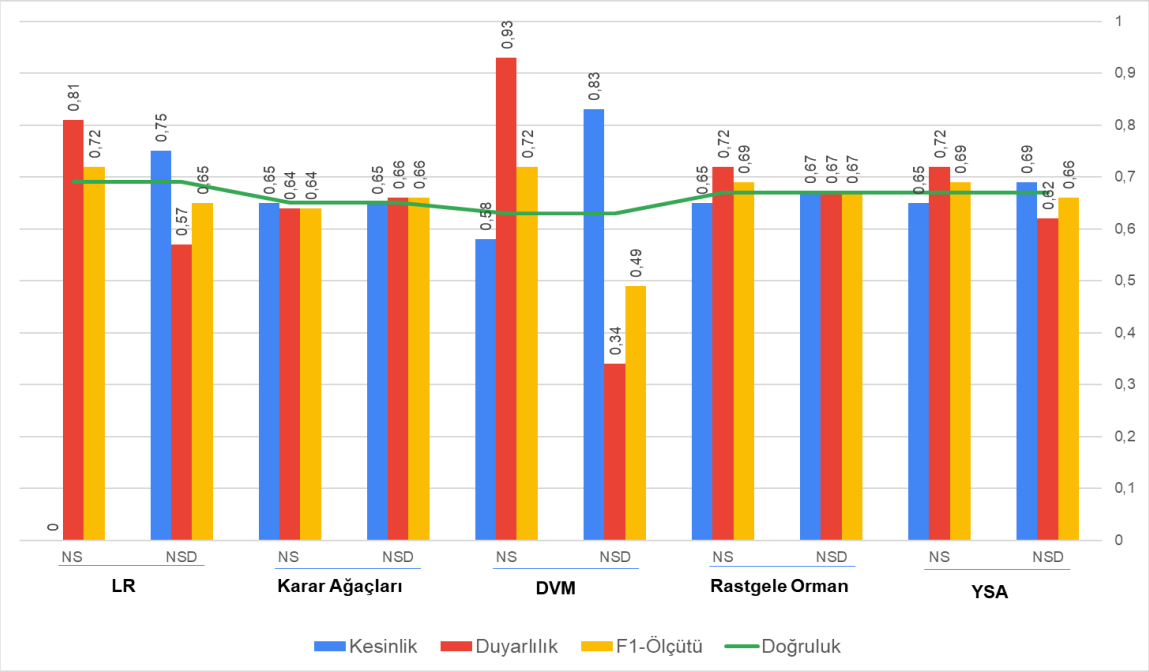
Tablo 4.7: Makine öğrenmesi modellerinde kullanılan parametreler

Model	Parametre	Parametre Değeri
Lojistik Regresyon	solver	liblinear
	c	10
	penalty	12
Karar Ağaçları	criterion	entropy
	random_state	0
DVM	decion_function_shape	ovo
	probability	true
Rastgele Orman	n_estimators	100
	criterion	gini
	max_depth	none
YSA	hidden_layer_size	10,10,10,10
	max_iter	10000

Makine öğrenmesi modelleri ile gerçekleştirilen uygulamanın sonuçları incelendiğinde, genel olarak TF-IDF kelime gömme yöntemi ile elde edilen sonuçların Word2Vec ile elde edilen sonuçlardan daha iyi olduğu ve en iyi performansın 0.81 doğruluk değeri ile TF-IDF kelime temsil ve LR, DVM, Rastgele Orman modelleri ile elde edildiği görülmektedir (Tablo 4.8 ve Tablo 4.9).

Tablo 4.8: Word2Vec Kelime gömme yöntemi ile elde edilen sonuçlar

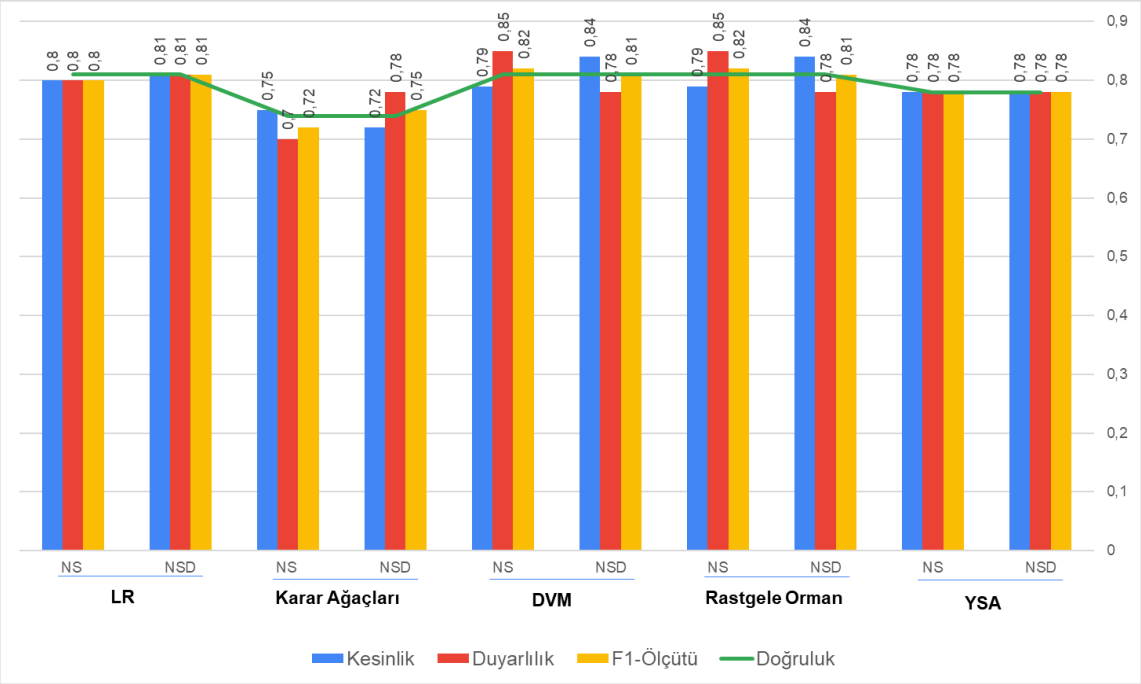
	Etiket	Kesinlik	Duyarlılık	F1-Ölçütü	Doğruluk
LR	0	0.65	0.81	0.72	**0.69**
	1	0.75	0.57	0.65	
Karar Ağaçları	0	0.65	0.64	0.64	0.65
	1	0.65	0.66	0.66	
DVM	0	0.58	0.93	0.72	0.63
	1	0.83	0.34	0.49	
Rastgele Orman	0	0.65	0.72	0.69	0.67
	1	0.67	0.67	0.67	
YSA	0	0.65	0.72	0.69	0.67
	1	0.69	0.62	0.66	



Şekil 4.8: Word2VEc ile elde edilen sonuçlar

Tablo 4.9: TF-IDF Kelime gömme yöntemi ile elde edilen sonuçlar

	Etiket	Kesinlik	Duyarlılık	F1-Ölçütü	Doğruluk
LR	0	0.80	0.80	0.80	**0.81**
	1	0.81	0.81	0.81	
Karar Ağaçları	0	0.75	0.70	0.72	0.74
	1	0.72	0.78	0.75	
DVM	0	0.79	0.85	0.82	**0.81**
	1	0.84	0.78	0.81	
Rastgele Orman	0	0.79	0.85	0.82	**0.81**
	1	0.84	0.78	0.81	
YSA	0	0.78	0.78	0.78	0.78
	1	0.78	0.78	0.78	



Şekil 4.9: TF-IDF ile elde edilen sonuçlar

4.4 BERTurk ile Elde Edilen Sonuçlar

Bu tez kapsamında ele alınan metin sınıflandırma problemi için transformers tabanlı dil modellerinden biri olan BERT altyapısını kullanan BERTurk modeli de kullanılmıştır. BERTurk modeli Türkçe Vikipedia, makine öğrenmesi ve yapay zeka çalışmaları için çok dilli kaynaklar sunmayı amaçlayan açık kaynak bir proje olan OSCAR ve yine açık kaynak bir başka veri sağlama projesi olan OPUS külliyatı üzerinden elde edilen 35 GB boyutunda ve 44.04.976.662 token'dan oluşan bir derlem ile eğitilmiştir [93].

BERTurk uygulanmadan önce veri seti numpy kütüphanesi ve kütüphane içerisinde bulunan random fonksiyonu kullanılarak rastgele bir şekilde %80 eğitim ve %20 test verisi şeklinde bölünmüştür. Özel karakterler ve performansa olumlu katkısı olmadığı düşünülen dolgu kelimeler silinmiştir. BERT'e özgü tokenizasyon kullanılmıştır. BERT'in en yüksek dizi uzunluğu 512'dir. Tweet'ler kısa metinlerden oluştuğu için bu araştırmada dizi uzunluğu 128 olarak seçilmiştir. Optimizer olarak Hugging Face'in PyTorch'a karşılık gelen kütüphanesi olan AdamW ($\beta_1 = 0.9$ ve $\beta_2 = 0.999$, $\varepsilon = 1e\text{-}8$) kullanılmıştır[2]. Eğitim 3 devrede (epoch) tamamlanmıştır.

BERTurk ile elde edilen sonuçların iyileştirilmesi için farklı öğrenme oranları ile deneysel uygulamalar yapılmıştır. En yüksek doğruluk değeri öğrenme oranı (lr) parametresi "3e-5" olarak ayarlandığında elde edilmiştir. Sonuçlar Tablo 4.10 ve 4.11'da sunulmuştur.

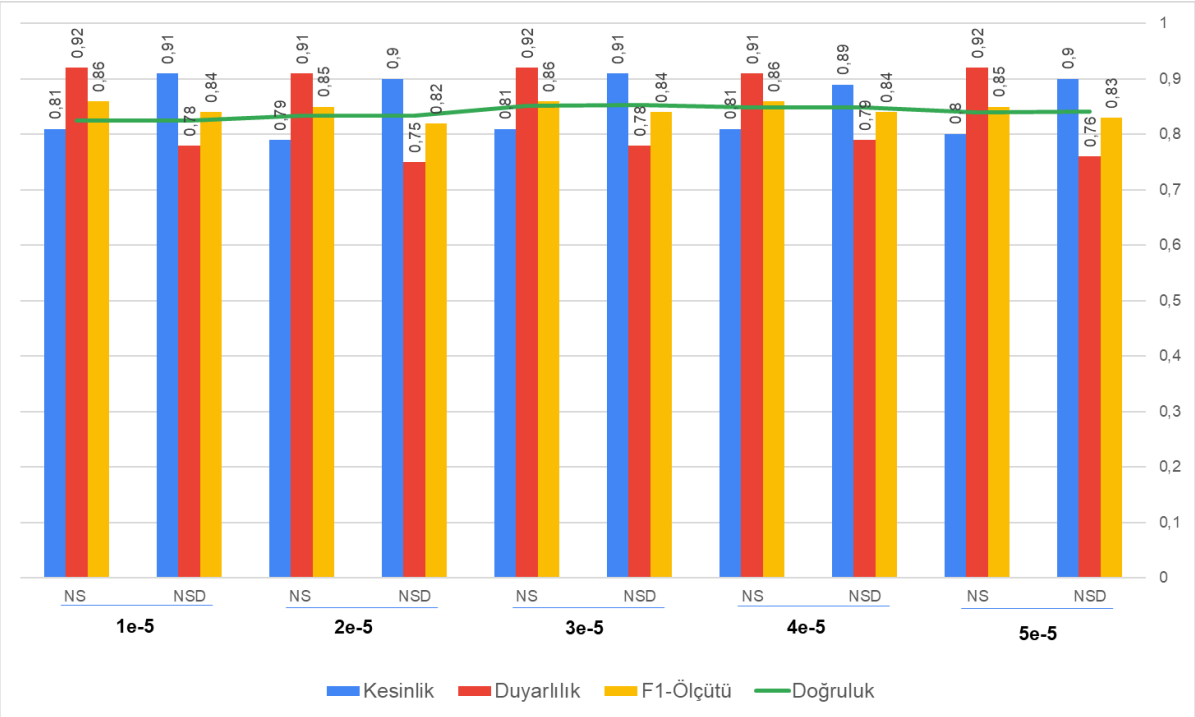
[2]huggingface.co/docs/transformers/main_classes/optimizer_schedules#transformers.AdamW.eps
https://arxiv.org/pdf/1711.05101.pdf

Tablo 4.10: BERTurk ile elde edilen sonuçlar

Öğrenme Oranı	Devre					
	1		2		3	
	Ort. Kayıp	Doğruluk	Ort. Kayıp	Doğruluk	Ort. Kayıp	Doğruluk
1e-5	0.54	0.78	0.39	0.81	0.33	0.81
2e-5	0.40	0.81	0.32	0.84	0.22	0.85
3e-5	0.49	0.82	0.29	0.85	0.17	0.86
4e-5	0.47	0.84	0.27	0.54	013	0.85
5e-5	0.47	0.83	0.26	0.83	0.12	0.85

Tablo 4.11: BERTurk ile elde edilen sonuçlar

Öğrenme Oranı	Etiket	Kesinlik	Duyarlılık	F1-Ölçütü	Doğruluk
1e-5	0	0.81	0.92	0.86	0.825
	1	0.91	0.78	0.84	
2e-5	0	0.79	0.91	0.85	0.834
	1	0.90	0.75	0.82	
3e-5	0	0.81	0.92	0.86	**0.852**
	1	0.91	0.78	0.84	
4e-5	0	0.81	0.91	0.86	0.848
	1	0.89	0.79	0.84	
5e-5	0	0.80	0.92	0.85	0.840
	1	0.90	0.76	0.83	



Şekil 4.10: BERTurk ile elde edilen sonuçlar

Bu bölümde Word2Vec ve TF-IDF kelime temsil yöntemleri ile Lojistik Regresyon, Karar Ağaçları, DVM, Rastgele Orman, YSA makine öğrenmesi ve BERTurk modeli ile yürütülen deneysel sonuçlar sunulmuştur. İzleyen bölümde bulgular tartışılacaktır.

Bölüm 5

Bulgular ve Tartışma

Bu bölümde bu araştırmada yürütülen deneysel çalışmalar sonucu elde edilen bulgular sunulmuş ve literatürdeki benzer çalışmalarla karşılaştırılarak tartışılmıştır. Bu araştırma kapsamında, Twitter'dan elde edilen veriler kullanılmıştır. Twitter'ın kullanıcılarına tweet atarken belirli bir karakter sınırlaması uygulaması nedeniyle kısa metinler üzerinde çalışılmıştır. Paragraf veya dökümanların aksine kısa metinler her zaman doğal dilin sözdizimine uymamakta, genellikle oldukça belirsiz olmakta ve sıklıkla yazım hataları içermektedirler [94]. Bu çalışmada karşılaşılan önemli zorluklardan biri, metinlerin kısa olması nedeniyle bağlamın anlaşılmasının zorlaşması yanı sıra Twitter üzerinden çekilen verilerde kelimelerin yazımında yapılan ve Türkçe karakterlerin kullanılmamasından kaynaklanan yanlışların çokluğu olmuştur. Uygulanan modellerin başarımının arttırılabilmesi için bu yazım yanlışları düzeltilmiştir.

Araştırmada öncelikle Word2Vec kelime temsil yöntemi ile Karar Ağaçları, Lojistik Regresyon, DVM, YSA ve Rastgele Orman modelleri uygulanmıştır. Word2Vec, Google tarafından sunulan ve BoW ile SkipGram modellerini içeren bir kelime temsil yöntemidir. Word2Vec'in büyük miktarlardaki verilerde boyut azaltmada [95] veya metin sınıflandırma problemlerinde [96] kullanıldığı görülmektedir. Word2Vec'in özellik çıkarımı için kullanıldığı ve derin öğrenme yöntemlerinin işe koşulduğu bir çalışmada sinir ağlarında başlangıç ağırlığını belirlemek için kullanılmış ve sonuçları iyileştirdiği görülmüştür [97]. Bu araştırmada Word2Vec kelime temsil yöntemi olarak kullanıldığında makine öğrenmesi modellerinde en yüksek doğruluk değeri Lojistik Regresyon ile elde edilmiş ve 0.69'a ulaşılmıştır. Rastgele Orman ve YSA modeliyle 0.67 doğruluk oranı elde edilmiştir. Karar Ağaçları ve DVM, Word2Vec kelime temsil yöntemiyle birlikte uygulanan bu modeller arasında en az başarım elde edilen modeller olmuş ve sırasıyla 0.65 ve 0.63 doğruluk oranlarına ulaşılabilmiştir. Nefret söyleminin tespiti için F1-Ölçütü değerleri Rastgele Orman modeli için 0.67,

YSA ve Karar Ağaçları modelinde 0.66 olmuştur. Lojistik Regresyon ile uygulanan modelde F1-Ölçütü 0.65 olarak bulunurken, en düşük F1-Ölçütü ise 0.49 ile DVM modeliyle elde edilmiştir.

Araştırmanın ikinci bölümünde TF-IDF kelime temsil yöntemi kullanılmış ve yine aynı modellerle deneysel çalışmalar sürdürülmüştür. TF-IDF kelime temsilinin, metin sınıflandırması üzerinde yapılan çalışmalarda, derin öğrenme algoritmalarıyla [98] ve özellik çıkarımı için [99] kullanıldığı görülmektedir.

TF-IDF modelinin en önemli kısıtlılığı bağlamdan bağımsız sınıflandırma yapması ve sözlüksel düzeyde kalmasıdır [100]. Bu araştırmada TF-IDF kelime temsil yöntemi kullanıldığında, Lojistik Regresyon, DVM ve Rastgele Orman modelleri ile 0.81 doğruluk oranına ve 0.81 F1-Ölçütüne ulaşılmıştır. YSA ile elde edilen doğruluk oranı ve F1-Ölçütü ise 0.78 olmuştur. Son olarak Karar Ağaçları ile 0.74 doğruluk oranı ve 0.75 F1-Ölçütü değerlerine ulaşılmıştır. Genel olarak en iyi sonuçların TF-IDF kelime temsil yöntemi kullanıldığında elde edildiği görülmüştür. TF-IDF ile elde edilen performans, tüm modellerde Word2Vec uygulandığında elde edilen performanstan daha yüksektir. Literatür incelendiğinde benzer şekilde metin sınıflandırma görevinde Word2Vec kelime gömme yönteminin TF-IDF'in gerisinde kaldığı görülmektedir [101]. Word2Vec ve TF-IDF ile elde edilen sonuçlar arasındaki fark gözetildiğinde kelime temsil yöntemi seçiminin, makine öğrenmesi modellerinin performansını önemli ölçüde etkilediği söylenebilir. Nitekim Forman ve Kirshenbaum [102] da belge temsili seçiminin sınıflandırıcının kalitesi üzerinde derin bir etkisi olduğunu göstermişlerdir. TF-IDF ve Word2Vec yöntemlerinin her ikisine modellerinde yer veren çalışmalara da rastlanmaktadır. Muhammed ve Omer [103] tarafından gerçekleştirilen bir araştırmada öğrenci başarısını değerlendirmek için sorulan sorular Bloom Taksonomisine göre otomatik olarak sınıflandırılmıştır. Araştırmada TF-IDF önemli kelimeleri ağırlıklandırmak için, Word2Vec ise sınıflandırma sürecini hızlandırma amacıyla kullanılmıştır. Çalışma sonuçlarına göre en yüksek F1-Ölçütü 0.89 olarak elde edilmiştir.

Bu araştırmada her iki kelime gömme yöntemiyle de nefret söyleminin tespitinde lojistik regresyon modelinin ön plana çıktığı görülmektedir. Word2Vec kelime gömme yönteminde en yüksek başarım lojistik regresyonla elde edilmiş, TF-IDF kelime gömme yönteminde ise lojistik regresyonla birlikte Rastgele Orman ve DVM ile YSA

ve Karar Ağaçlarına göre daha yüksek doğruluk oranlarına ulaşılmıştır. Elde edilen bu sonuçların literatürdeki çalışmalarla paralel olduğu, Naïve Bayes, Rastgele Orman, Karar Ağaçları, DVM ve Lojistik Regresyonun metin sınıflandırma başarısının karşılaştırıldığı bir çalışmada, Lojistik Regresyonun diğer modellere göre daha iyi doğruluk oranı yakaladığı görülmektedir [104]. Başka bir çalışmada ise K-en Yakın Komşu ve Rastgele Orman ile Lojistik Regresyon modellerinin performansları BBC Haberlerinin sınıflandırılmasına yönelik bir çalışma ile karşılaştırılmış ve Lojistik regresyon ile daha iyi performans elde edilmiştir [105].

Metin sınıflandırmada, transformers tabanlı modellerin yüksek performans gösterdiği ve BERT modelinin sıklıkla kullanıldığı görülmektedir. Sel ve Hanbay [106] tarafından Türkçe dilinde kısa metinlerden cinsiyet tespitinin yapılmasına yönelik yürütülen çalışmada en yüksek başarım %80.1 doğruluk değeri BERT ile elde edilmiştir. Doğal felaketlere yönelik atılan tweetlerin gerçek veya gerçek dışı olarak etiketlendiği 7613 tweet içeren bir veri seti ile yapılan çalışmada BERT ile %98 doğruluk oranına ulaşıldığı görülmektedir [107]. Suriyelilere yönelik ayrımcılık içeren tweetlerin sınıflandırıldığı başka bir çalışmada ise BERT modeli ile %85 doğruluk oranı elde edilmiştir. Literatür incelendiğinde BERT modelinin doğal dil işleme görevlerinde güçlü bir çözüm sunduğu, literatüre benzer şekilde bu araştırmada da en iyi performansın BERT temelli BERTurk ile elde edildiği görülmektedir.

Bu araştırmada mültecilere yönelik nefret söyleminin tespiti için BERT tabanlı bir model olan BERTurk kullanılmıştır. Farklı öğrenme oranları ile yapılan deneysel çalışmalar ile “1e-5” doğruluk oranı kullanıldığında ilk devrede 0.78 doğruluk, ikinci devrede 0.81 ve 3. devrede yine 0.81 doğruluk değerine ulaşılmıştır. BERT’e ait doğruluk oranı ise %82.5 olmuştur. “2e-5” öğrenme oranı uygulandığında ise, ilk devrede 0.81, ikinci devrede 0.84 ve 3. devrede 0.85 doğruluk oranı elde edilmiştir. Bu öğrenme oranı uygulandığında BERT’e ait doğruluk oranının %83.4 olduğu görülmüştür. “3e-5” en yüksek değerlerin elde edildiği parametre değeri olmuştur. İlk devrede 0.82 doğruluk değerine erişilirken 2. ve 3. devrelerde artarak 0.85 ve 0.86 ile bu araştırmadaki en yüksek doğruluk değerlerine ulaşmıştır. BERT’in doğruluk değeri ise %85.2 olmuştur. Öğrenme oranı “4e-5” ve “5e-5” olarak ayarlandığında da yakın sonuçlar elde edildiği görülmüştür. “4e-5” parametresi ile birinci devrede 0.83, ikinci devrede 0.84 ve üçüncü devrede 0.85 doğruluk değerlerine ulaşılmıştır. BERT’in doğruluk değeri ise %84.8 olmuştur. “5e-5” parametresi uygulandığında ise birinci ve

ikinci devrede 0.83, üçüncü evrede ise 0.85 doğruluk değerine ulaşılmış ve BERT'in doğruluk değeri %84.0 olarak ölçülmüştür. Literatür incelendiğinde, sosyal medyadaki nefret söyleminin tespitine yönelik çalışmalarda BERT modelinin sıklıkla kullanıldığı görülmektedir. Nefret söyleminin tespiti için oluşturulmuş farklı veri setleri üzerinde BERT modeli ile deneysel çalışmalar yürüten Mozafari ve ark. [108], etiketlemeleri yapan yorumculardan kaynaklı ön yargının önüne geçmek için yeniden ağırlıklandırılmış yeni bir eğitim seti ile ince ayar yapmış ve bu şekilde bir önyargı azaltma mekanizması önermişlerdir. Nefret söyleminin tespitinde DVM ve BERT modellerinin kıyaslandığı bir başka araştırmada ise dengeli olmayan veri setlerinde bile BERT modeliyle daha yüksek başarım elde edildiği vurgulanmıştır [109].

Bu bölümde araştırma sonucunda elde edilen bulgular literatürdeki benzer çalışmalar ile kıyaslanarak tartışılmıştır. İzleyen bölümde ise sonuçlar sunulacaktır.

Bölüm 6

Sonuçlar

Nefret söylemi, belirli bir gruba ya da kişiye yönelik olarak, sahip olduğu kimlik unsurları nedeniyle aşağılama, ötekileştirme, ayrımcılık veya saldırgan dil içeren söylemler olarak tanımlanmaktadır. Nefret söyleminin hedefinde kadınlar, eşcinseller, yabancılar, belirli bir dinin mensubu kişiler ya da mülteciler gibi gruplar olabilmektedir. Türkiye'de mültecilere karşı nefret söyleminin, Suriye iç savaşını takiben karşı karşıya kalınan yoğun göç nedeniyle özellikle sosyal medya ortamında yükseldiği görülmektedir. Türkiye'de sosyal medyanın 16 milyonun üzerinde kullanıcısı bulunmaktadır ve özellikle kötü niyetli kişiler tarafından sosyal medya ortamının kontrolünün zor olması fırsat bilinerek nefret söyleminin yaygınlaşabildiği görülmektedir. Sosyal medyada paylaşım hızı göz önüne alındığında, nefret söyleminin otomatik olarak tespit edilmesi, nefret söylemini takip edebilecek şiddet eylemlerinin de önüne geçilmesi açısından önemlidir. Mültecilere yönelik nefret söyleminin tespit edilmesi için yapılan çalışmalar incelendiğinde Türkçe dilinde kapsamlı bir veri setinin olmadığı görülmüştür. Bu tez çalışması kapsamında ilk olarak nefret söyleminin tespitine yönelik bir veri setinin ortaya konması amaçlanmıştır. Twitter üzerinden SNScrape aracı kullanılarak çekilen veriler, ön elemeye tabi tutulmuş, 2 uzman tarafından ayrı ayrı etiketlenmiş ve yorumcular arasındaki uyum Cohen's Cappa metriği kullanılarak hesaplanmıştır. Sonuç olarak 9774 adet tweet içeren ve NS ve NSD etiketleriyle etiketlenmiş dengeli bir veri seti elde edilmiştir. Veri setinde 4947 adet NS ve 4831 adet NSD etiketli veri ile çeşitli deneysel çalışmalar yürütülmüştür.

Çalışmada, en büyük zorluk veri setindeki yanlış yazımların varlığı olmuştur. Kısa metinlerle yapılan paylaşımlar bağlamın anlaşmasını zorlaştırdığından uygulanan modellerin performanslarını düşürdüğü tahmin edilmektedir. Öncelikle Word2Vec kelime temsil yöntemi ile LR, Karar Ağaçları, DVM, Rastgele Orman ve YSA modelleri denenmiştir. En yüksek performans LR ile 0.69 doğruluk oranıyla elde edilmiştir.

İkinci aşamada, TF-IDF kelime temsil yöntemi denenmiştir. LR, DVM ve Rastgele Orman modelleri 0.81 doğruluk değeri verirken Karar Ağaçları ve YSA 0.78 ve 0.74 oranları ile diğer modellerin gerisinde kaldığı, TF-IDF ile genel olarak Word2Vec kelime temsil yöntemine göre daha iyi sonuçlar elde edildiği görülmüştür.

Araştırmanın üçüncü aşamasında, Transformatör tabanlı bir model olan BERT temel alınarak eğitilen Türkçe diline özgü BERTurk modeli kullanılmıştır. Model farklı hiperparametreler kullanılarak denenmiştir. Literatürde BERT modeli ile elde edilen sonuçların makine öğrenmesi ve diğer derin öğrenme modellerinden elde edilen sonuçlardan daha iyi olduğu görülmektedir. Bu çalışmada da BERTurk ile elde edilen sonuçlar, 3e-5 öğrenme oranı kullanılarak 0.85 doğruluk oranına ulaşmış ve makine öğrenmesi modellerinden daha iyi bir performans elde edilmiştir.

Bu çalışmanın, literatürdeki mültecilere yönelik nefret söylemi için kullanılabilecek Türkçe veri seti eksikliğini gidereceği ve uygulanan modeller ile nefret söyleminin tespitinde yürütülecek çalışmalara yön vereceği düşünülmektedir.

İleriki çalışmalarda, literatüre sunulan bu veri seti üzerinde BERT modeline ait farklı hiperparemetreler ile deneysel çalışmalar yürütülerek sonuçlar incelenebilir.

Kaynaklar

1. Babacan M, Haşlak İ, Hira İ. Sosyal medya ve Arap baharı. Akademik İncelemeler Dergisi 2011; 6(2): 63-92.
2. Uluç G, Yarcı A. Sosyal medya kültürü. Dumlupınar Üniversitesi Sosyal Bilimler Dergisi 2017; 52: 88-102.
3. Yaşa H, Öksüz O. Nefret söyleminin inşasında sosyal medyanın rolü: Ekşi Sözlük örneği. Erciyes İletişim Dergisi 2020; 7(2): 1383-1408.
4. Chakraborty T, Masud S. Nipping in the bud: detection, diffusion and mitigation of hate speech on social media. ACM SIGWEB Newsletter 2022; 3: 1-9. doi.org:10.1145/3522598.3522601
5. Simon H, Baha BY, Garba EJ. Trends in machine learning on automatic detection of hate speech on social media platforms: A Systematic review. FUW Trends in Science & Technology Journal 2022; 7(1): 001-016.
6. Türkiye İstatistik Kurumu. Uluslararası Göç İstatistikleri [İnternet] Ankara; 2022 [erişim tarihi 05.12.2022]. https://data.tuik.gov.tr/Bulten/Index?p=Uluslararasi-Goc-Istatistikleri-2019-33709
7. T.C. İçişleri Bakanlığı Göç İdaresi Başkanlığı. Geçici Koruma [İnternet] Ankara; 2022 [erişim tarihi 05.12.2022]. https://www.goc.gov.tr/gecici-koruma5638
8. Caleb T, Hayes C, Hayes R. Social Media: Defining, Developing, and Divining. Atlantic Journal of Communication 2015; 23:1, 46-65. doi.org: 10.1080/15456870.2015.972282
9. Statistica. Social Media &User Generated Content. [İnternet] Hamburg; 2023 [erişim tarihi: 12.01.2023]. https://www.statista.com/statistics/315405/snapchat-user-region-distribution/
10. Statistica. Number of Social Nertwork Users in selected Countries. [İnternet] Hamburg; 2023 [erişim tarihi 12.01.2023] https://www.statista.com/statistics/278341/number-of-social-network-users-in-selected-countries/

11. Mathew B, Dutt R, Goyal P, Mukherjee, A. Spread of hate speech in online social media. In Proceedings of the 10th ACM conference on web science; 2019 June; 173-182. https://arxiv.org/pdf/1812.01693.pdf

12. Castaño-Pulgarín SA, Suárez-Betancur N, Vega LMT, López HMH. Internet, social media and online hate speech. Systematic review. Aggression and Violent Behavior 2021; 58: 101608.

13. Nobata C, Tetreault J, Thomas A, Mehdad Y, Chang Y. Abusive language detection in online user content. Proceedings of the 25th International Conference on World Wide Web; International World Wide Web Conferences Steering Committee; 2016 April; 145– 153. http://yichang-cs.com/yahoo/WWW16_Abusivedetection.pdf

14. McNamee G L, Peterson BL, Pena J. A call to educate, participate, invoke. and indict: Understanding the communication of online hate groups. Communication Monographs 2010; 77(2):257–280. doi.org:10.1080/03637751003758227

15. Schmidt A, Wiegand M. A Survey on Hate Speech Detection using Natural Language Processing. Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media; 2017 April; Valencia, Spain. 1–10. https://aclanthology.org/W17-1101

16. Papcunová J, Martončik M, Fedáková D, Kentoš M, Bozogáňová M, Srba I et all. Hate speech operationalization: a preliminary examination of hate speech indicators and their structure. Complex & Intelligent Systems 2021; 1-16: https://doi.org/10.1007/s40747-021-00561-0

17. United Nations. What is Hate Speech [İnternet]. 2023 [erişim tarihi: 12.01.2023]. https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech

18. Yaşa H, Öksüz O. Nefret söyleminin inşasında sosyal medyanın rolü: Ekşi Sözlük örneği. Erciyes İletişim Dergisi 2020; 7(2):1383-1408.

19. Kuş O. Dijital Nefret Söylemini Anlamak: Suriyeli Mülteci Krizi Örnek Olayı Bağlamında Bbc World Servıce Facebook Sayfasına Gelen Yorumların Metin Madenciliği Tekniği İle Analizi. İstanbul Üniversitesi İletişim Fakültesi Dergisi 2016; (51): 97-121.

20. Malmasi S, Zampieri M. Detecting hate speech in social media. Proceedings of Recent Advances in Natural Language Processing (RANLP); 2017 Dec 26; Bulgaria, Varna. 467-472. https://arxiv.org/abs/1712.06427

21. Mathew B, Dutt R, Goyal P, Mukherjee A. Spread of hate speech in online social media.WebSci '19: Proceedings of the 10th ACM Conference on Web Science; 2019 June; 173-182. https://arxiv.org/pdf/1812.01693.pdf

22. Mathew B, Kumar N, Goyal P, Mukherjee A. Analyzing the hate and counter speech accounts on twitter. arXiv preprint arXiv:1812.02712; 2018. https://arxiv.org/pdf/1812.02712.pdf

23. Mathew B, Saha P, Yimam SM, Biemann C, Goyal P, Mukherjee A. Hatexplain: A benchmark dataset for explainable hate speech detection. In Proceedings of the AAAI Conference on Artificial Intelligence; 2021 May; 35(17);14867-14875.

24. Mullah NS, Zainon WMNW. Advances in machine learning algorithms for hate speech detection in social media: A review. IEEE Access; 2021; 9: 88364-88376. doi.org: 10.1109/ACCESS.2021.3089515

25. Burnap P, Williams ML. Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making. Policy & Internet 2015; 7(2): 223-242.

26. Waseem Z, Hovy D. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In: Proceedings of the NAACL Student Research; 2016 June; San Diego, California. 88-93. https://aclanthology.org/N16-2013.pdf

27. Davidson T, Warmsley D, Macy M, Weber I. Automated Hate Speech Detection and the Problem of Offensive Language. Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM); 2017; 512-515.

28. De Smedt T, De Pauw G, Van Ostaeyen P. Automatic Detection of Online Jihadist Hate Speech. Computational Linguistics & Psycholinguistics Technical Report Series 2018; 7:1-31. https://arxiv.org/abs/1803.04596

29. Gaydhani A, Doma V, Kendre S, Bhagwat L. Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach. IEEE International Advance Computing Conference; 2018 Sep 23; https://doi.org/10.48550/arXiv.1809.08651

30. Park JH, Fung P. One-step and two-step classification for abusive language detection on twitter. In: 1st Workshop on Abusive Language Online to be held at the annual meeting of the Association of Computational Linguistics (ACL); 2017 Agust 4; Vancouver, Canada. https://doi.org/10.48550/arXiv.1706.01206

31. Pelle R, Alcântara C, Moreira VP. A classifier ensemble for offensive text detection. In: Proceedings of the 24th Brazilian Symposium on Multimedia and the Web; 2018 October. 237-243. https://doi.org/10.1145/3243082.3243111

32. Faris H, Aljarah I, Habib M, Castillo PA. Hate Speech Detection using Word Embedding and Deep Learning in the Arabic Language Context. In: ICPRAM 9th International Conference on Pattern Recognition Applications and Methods; 2020 February. 453-460. https://www.researchgate.net/profile/Hossam-Faris/publication/339920611_Hate_Speech_Detection_using_Word_Embedding_and_Deep_Learning_in_the_Arabic_Language_Context/links/5e7290ba92851c93e0ad4550/Hate-Speech-Detection-using-Word-Embedding-and-Deep-Learning-in-the-Arabic-Language-Context.pdf

33. Dorris W, Hu R, Vishwamitra N, Luo F, Costello M. Towards automatic detection and explanation of hate speech and offensive language. In: Proceedings of the sixth international workshop on security and privacy analytics; 2020 March 16; 23-29. https://doi.org/10.1145/3375708.3380312

34. Altin LSM, Serrano ÀB, Saggion H. Lastus/taln at semeval-2019 task 6: Identification and categorization of offensive language in social media with attention-based bi-lstm model. In Proceedings of the 13th International Workshop on Semantic Evaluation; 2019 June 16; Minneapolis, Minnesota, USA. 672-677. https://aclanthology.org/S19-2120

35. Jain D, Kumar A, Garg G. Sarcasm detection in mash-up language using soft-attention based bi-directional LSTM and feature-rich CNN. Applied Soft Computing 2020; 91, 106198. doi.org:10.1016/j.asoc.2020.106198

36. Alkomah F, Salati S, Ma X. A New Hate Speech Detection System based on Textual and Psychological Features. International Journal of Advanced Computer Science and Applications 2022; 13(8). doi.org:10.14569/IJACSA.2022.01308100

37. Saeed AM, Ismael AN, Rasul DL, Majeed RS, Rashid TA. Hate Speech Detection in Social Media for the Kurdish Language. In Proceedings of the ICR'22 International Conference on Innovations in Computing Research; 2022 August 11; 253-260. doi.org: 10.1007/978-3-031-14054-9_24

38. Özbay E. Transformatör-Tabanlı Evrişimli Sinir Ağı Modeli Kullanarak Twitter Verisinde Saldırganlık Tespiti. Konya Journal of Engineering Sciences; 2022:10(4): 986-1001.

39. What is Text Mining? [İnternet]. 2003 [erişim tarihi 12.01.2023]. https://www.jaist.ac.jp/~bao/MOT-Ishikawa/FurtherReadingNo1.pdf

40. Bitter C, Elizondo DA, Yang Y. Natural language processing: a prolog perspective. The Artificial Intelligence Review 2010; 33(1-2): 151.

41. Jusoh S, Alfawareh HM. Natural language interface for online sales; Proceedings of the International Conference on Intelligent and Advanced System (ICIAS2007); 2007 November 25-28; Kuala Lumpur, Malaysia; 224–228. https://ieeexplore.ieee.org/document/4658379

42. Şeker, SE. Doğal Dil İşleme (Natural Language Processing). YBS Ansiklopedi 2015; 2(4):14-31.

43. Žižka J, Dařena F, Svoboda A. Text mining with machine learning: principles and techniques. First, Crc Press; 2019

44. Anandarajan M, Hill C, Nolan T. Text preprocessing. Practical text analytics: Maximizing the value of text data 2019; 45-59

45. Onan A. Twitter mesajları üzerinde makine öğrenmesi yöntemlerine dayalı duygu analizi. Yönetim Bilişim Sistemleri Dergisi 2017; 3(2): 1-14

46. Porter MF. An Algorithm for Suffix Stripping Program 1980; 14 (3):130-137

47. Lovins J. B. Development of a stemming algorithm. Mech. Transl. Comput. Linguistics 1968; 11(1-2): 22-31.

48. Balakrishnan, V., & Lloyd-Yemoh, E. (2014). Stemming and lemmatization: A comparison of retrieval performances.

49. Fürnkranz J. A study using n-gram features for text categorization. Austrian Research Institute for Artifical Intelligence 1998; 3:1-10.

50. Zhang Y, Jin R, Zhou ZH. Understanding bag-of-words model: a statistical framework. International journal of machine learning and cybernetics 2010; 1: 43-52.

51. Güran A, Akyokuş S, Bayazıt NG, Gürbüz MZ. Turkish text categorization using n-gram words. Proceedings of the International Symposium on Innovations in Intelligent Systems and Applications (INISTA 2009); 2009, June; 369-373.https://www.researchgate.net/profile/Emre-Ozkop/publication/262001726_Proceedings_of_INISTA_2009_International_Symposium_on_INnovations_in_Intelligent_SysTems_and_Applications/links/0deec536355bb83080000000/Proceedings-of-INISTA-2009-International-Symposium-on-INnovations-in-Intelligent-SysTems-and-Applications.pdf#page=381

52. Uguz H. A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm; Knowledge-Based Systems 2011; 24:1024–1032.

53. Rong, X. word2vec parameter learning explained. 2014 arXiv preprint arXiv:1411.2738. https://doi.org/10.48550/arXiv.1411.2738

54. Google Code. word2vec [İnternet]. 2013 [erişim tarihi 13.01.2023]. https://code.google.com/archive/p/word2vec/]

55. Medium. word2vec nedir? [İnternet]. 2018 [erişim tarihi 13.01.2023] https://medium.com/@muhammedbuyukkinaci/word2vec-nedir-t%C3%BCrk%C3%A7e-f0cfab20d3ae

56. Jatnika D, Bijaksana M. A, Suryani A. A. Word2vec model analysis for semantic similarities in english words. Procedia Computer Science, 2019;157: 160-167.

57. Xiong Z, Shen Q, Xiong Y, Wang Y, Li W. New generation model of word vector representation based on cbow or skip-gram. Computers, Materials & Continua; 2019;. 60(1): 259–273.

58. Stanford. GloVe: Global Vectors for Word Representation [İnternet]. 2014 [erişim tarihi 13.01.2023]. https://nlp.stanford.edu/projects/glove/

59. Facebook Inc. Word vectors for 157 languages [İnternet] 2022 [erişim tarihi: 13.01.2023] https://fasttext.cc/docs/en/crawl-vectors.html

60. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, et all. Deep contextualized word representations. arXiv 2018;12. arXiv preprint arXiv:1802.05365.

61. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

62. Miaschi A, Dell'orletta F. Contextual and non-contextual word embeddings: an in-depth linguistic investigation. Proceedings of the 5th Workshop on Representation Learning for NLP; 2020 July. 10-119. https://aclanthology.org/2020.repl4nlp-1.15

63. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal, Amodei D, et all. Language models are few-shot learners. Advances in neural information processing systems, 2020; 33: 1877-1901.

64. Samuel AL. Some studies in machine learning using the game of checkers. IBM Journal of research and development 1959; 3(3): 210-229.

65. Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. Journal of biomedical informatics, 2002;35(5-6): 352-359.

66. Karayiğit H. Sosyal Medyada Türkçe Nefret Söylemlerinin ve Covid-19 Yorumlarının Makine Öğrenmesi, Derin Öğrenme ve BERT Teknikleri İle Analizi (doktora tezi) Mersin: Mersin Üniversitesi; 2022.

67. Dayton CM. Logistic regression analysis. Stat; 1992:474-574.

68. Indra ST, Wikarsa L, Turang R. Using logistic regression method to classify tweets into the selected topics. IEEE: 2016 international conference on advanced computer science and information systems (icacsis); 2016 October; 385-390.

69. Hopfield J. J.Artificial neural networks. IEEE Circuits and Devices Magazine 1988; 4(5): 3-10.

70. Abraham A. Artificial neural networks. Handbook of measuring system design, 2005.

71. Medium. A Beginner Intro to Neural Networks [İnternet]. [erişim tarihi 04.04.2023]. https://purnasaigudikandula.medium.com/a-beginner-intro-to-neural-networks-543267bda3c8

72. Cornell University. Lecture 2: k-nearest neighbors [İnternet]. [erişim tarihi 04.04.2023]. https://www.cs.cornell.edu/courses/cs4780/2018fa/lectures/lecturenote02_kNN.html#:~:text=The%20k%2DNN%20algorithm&text=Denote%20the%20set%20of%20the,furthest%20point%20in%20Sx

73. Khamar K. Short text classification using kNN based on distance function. International Journal of Advanced Research in Computer and Communication Engineering 2013; 2(4): 1916-1919.

74. Noble WS. What is a support vector machine?. Nature biotechnology 2006; 24(12): 1565-1567.

75. Özberk, A. Offensive Language Detection in Turkish Twitter Data with BERT Models (yüksek lisans tezi). Ankara: Hacettepe Üniversitesi; 2022. https://tez.yok.gov.tr/

76. Kim H. et al. Dimension reduction in text classification with support vector machines. Journal of machine learning research 2005; 6(1).

77. De Ville B. Decision trees. Wiley Interdisciplinary Reviews: Computational Statistics 2013; 5(6): 448-455.

78. Breiman L. Random forests. Machine learning 2001; 45: 5-32.

79. Brownlee J. Deep learning for natural language processing: develop deep learning models for your natural language problems. Machine Learning Mastery; 2017.

80. A Primer on Neural Network Models for Natural Language Processing, 2015.

81. Johnson R, Zhang T. Effective use of word order for text categorization with convolutional neural networks 2015; arXiv preprint arXiv:1412.1058, 2014.

82. Lai S, Xu L, Liu K, Zhao J. Recurrent Convolutional Neural Networks for Text Classification. Proceedings of the AAAI Conference on Artificial Intelligence; 2015; 29(1). https://doi.org/10.1609/aaai.v29i1.9513

83. Chen J, Hu Y, Liu J, Xiao Y, Jiang H. Deep short text classification with knowledge powered attention. Proceedings of the AAAI conference on artificial intelligence; 2019 July; 33(1); 6252-6259.

84. Vaswani A. et al. Attention is all you need. Advances in neural information processing systems 2017; 30:5998-6008.

85. Github. The Illustrated Transformer [İnternet]. [erişim tarihi 14.03.2023]. http://jalammar.github.io/illustrated-transformer/

86. Huang Z, Xu P, Liang D, Mishra A, Xiang B. TRANS-BLSTM: Transformer with bidirectional LSTM for language understanding. 2020 arXiv preprint arXiv:2003.07000.

87. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D.ve diğ. Roberta: A robustly optimized bert pretraining approach. 2019; arXiv:1907.11692.

88. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. Albert: A lite bert for self-supervised learning of language representations. 2019; arXiv:1909.11942.

89. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

90. Github. Snscrape [İnternet]. 2020 [erişim tarihi 14.02.2023]. https://github.com/JustAnotherArchivist/snscrape

91. Cohen JA. Coefficient of agreement for nominal scales. Educational and Psychological Measurement 1960; 20(1): 37-46.

92. Çelikten A, Bulut H. Turkish medical text classification using bert. 29th Signal Processing and Communications Application; 2021 June;1-4

93. Github. stefan-it/turkish bert [İnternet]. 2021 [erişim tarihi 14.02.2023] https://github.com/stefan-it/turkish-bert

94. Wang, J. et al. Combining Knowledge with Deep Convolutional Neural Networks for Short Text Classification. IJCAI; 2017. 3172077-3172295.

95. MA L, Zhang Y. Using Word2Vec to process big text data. 2015 IEEE International Conference on Big Data (Big Data). IEEE; 2015. 2895-2897.

96. Lilleberg J, Z Y, Zhang Y. Support vector machines and word2vec for text classification with semantic features. In: 2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC); 2015 June; 136-140.

97. Kurnia RI, Girsang, AS. Classification of user comment using word2vec and deep learning. Int J Emerg Technol Adv Eng 2021; 11(5): 1-8.

98. Zhou H. Research of Text Classification Based on TF-IDF and CNN-LSTM. Journal of Physics: Conference Series. IOP Publishing; 2022; 012021.

99. Setiawan Y, Gunawan D, Efendi R. Feature Extraction TF-IDF to Perform Cyberbullying Text Classification: A Literature Review and Future Research

Direction. In: 2022 International Conference on Information Technology Systems and Innovation (ICITSI). IEEE; 2022; 283-288.

100. Qaiser S, Ali R. Text mining: use of TF-IDF to examine the relevance of words to documents. International Journal of Computer Applications 2018; 181(1): 25-29.

101. Cahyani DE, Patasik I. Performance comparison of tf-idf and word2vec models for emotion text classification. Bulletin of Electrical Engineering and Informatics 2021; 10(5): 2780-2788.

102. Forman G. Kirshenbaum E. Extremely fast text feature extraction for classification and indexing. Proceedings of the 17th ACM conference on Information and knowledge management; 2008; 1221-1230.

103. Mohammed M, Omar N. Question classification based on Bloom's taxonomy cognitive domain using modified TF-IDF and word2vec. PloS one 2020; 15(3): e0230442.

104. Pranckevičius T, Marcinkevičius V. Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. Baltic Journal of Modern Computing 2017; 5(2): 221.

105. Shah K, Patel H, Sanghvi D. et al. A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification. Augment Hum Res 5; 2020:12 https://doi.org/10.1007/s41133-020-00032-0

106. Sel İ, Hanbay D. Ön Eğitimli Dil Modelleri Kullanarak Türkçe Tweetlerden Cinsiyet Tespiti. Fırat Üniversitesi Mühendislik Bilimleri Dergisi 2021; 33(2): 675-684.

107. Sevli O, Kemaloğlu N. Olağandışı olaylar hakkındaki tweet'lerin gerçek ve gerçek dışı olarak google BERT modeli ile sınıflandırılması. Veri Bilimi 2021; 4(1): 31-37.

108. Mozafari M, Farahbakhsh R, Crespi N. BERT modeline dayalı olarak sosyal medyada nefret söylemi algılama ve ırksal önyargıyı azaltma. PLoS BİR, 2020; 15(8): e0237861. doi.org:10.1371/journal.pone.0237861

109. Kumar R, Ojha AK. KMI-Panlingua at HASOC 2019: SVM vs BERT for Hate Speech and Offensive Content Detection. FIRE (Working Notes). 2019; 25: 285-292.

Ekler

Ek A Tezden Üretilmiş Yayınlar

Makaleler

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

Özgeçmiş

Adı Soyadı: Figen Eğin

Eğitim:

2006 Lisans - Ege Üniversitesi, Bilgisayar ve Öğretim Tek. Eğitimi Bölümü

2016 Yüksek Lisans - Ege Üniversitesi, Bilgisayar ve Öğretim Tek. Eğitimi Bölümü

İş Deneyimi:

2006 – Aliağa Lisesi Bilişim Tek. Öğretmeni

2007 – Aliağa İlçe Milli Eğitim Müdürlüğü İlçe Koordinatörü

2008 – Muş İMKB Anadolu Lisesi Bilişim Tek. Öğretmeni

2010 – Çorlu Ticaret Borsası Anadolu Lisesi Bilişim Tek. Öğretmeni

2014 – Turgutlu Gazi Ortaokulu Bilişim Tek. Öğretmeni

2016 – Turgutlu Bilsem Bilişim Tek. Öğretmeni

Yayınlar:

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

2. Bayburt B., Eğin F. Teknoloji ve Sanayideki Gelişmelerin Yansıması Olarak Eğitim 4.0. Bilgi Ekonomisi ve Yönetimi Dergisi, 2021; 16.2: 137-154.

3. Eğin F., Arıkan YD. Bilişim teknolojileri öğretmenlerinin kodlama öğretimine ilişkin görüşleri: Manisa örneği. Ege Eğitim Dergisi, 2020; 21(2): 57-75.

4. Uslu N., Mumcu F., Eğin F. Görsel programlama etkinliklerinin ortaokul öğrencilerinin bilgi-işlemsel düşünme becerilerine etkisi. Ege Eğitim Teknolojileri Dergisi; 2018, 2(1): 19-31. https://dergipark.org.tr/en/download/article-file/511326