VERİ BİLİMİ DERGİSİ
www.dergipark.gov.tr/veri

Türkçe Hakaret ve Nefret Söylemi Otomatik Tespit Modeli

Mehmet Salih KURT[1]*, Eylem YÜCEL DEMİREL[2]

[1]*Hakkari Üniversitesi, Enformatik Bölümü, Hakkari, TÜRKİYE*
[2]*İstanbul Üniversitesi-Cerrahpaşa, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, İstanbul, TÜRKİYE*

Makale Bilgisi

Başvuru:
22/05/2023
Kabul:
22/06/2023

Özet

İnsanların çevrimiçi dünyada, özellikle sosyal medya platformlarında iletişim kurmasıyla birlikte, kullanıcılar tarafından oluşturulan içeriklerin internet üzerindeki miktarı artmıştır. Bu platformların anonim yapısı nedeniyle, kullanıcılar hakaret ve nefret içeren düşünceleri paylaşabilmektedir. Bu istenmeyen içerikler, hem bireyler hem de toplumlar üzerinde olumsuz etkilere neden olabilir. Bu nedenle, hakaret ve nefret içeren içeriklerin tespit edilmesi ve filtrelenmesi önemlidir. Bu tür içeriklerin manuel olarak tespit edilmesi zordur, bu yüzden otomatik yöntemlere ihtiyaç duyulmaktadır. Son yıllarda, çevrimiçi hakaret ve nefret söylemlerinin tespitiyle ilgili akademik araştırmalarda artış görülmektedir. BERT gibi transfer öğrenme modelleriyle İngilizce hakaret ve nefret söylemlerinin otomatik tespiti konusunda umut verici sonuçlar elde edilmiştir. Ancak, Türkçe gibi sınırlı kaynaklara sahip dillerde hakaret ve nefret söyleminin otomatik tespiti üzerine yapılan araştırma sayısı oldukça azdır.

Bu çalışmada, Türkçe dili için hakaret ve nefret söylemi otomatik tespit sistemi geliştirme çabalarının sonuçları paylaşılmıştır. İlk olarak, Türkçe veri seti oluşturmak için otomatik etiketleme yöntemi önerilmiş ve bu yöntemle Türkçe hakaret ve nefret söylemi veri seti oluşturulmuştur. Doğal dil işleme alanında en iyi sonuçlar veren BERT modelinin farklı varyantları ve çeşitli Türkçe hakaret ve nefret söylemi veri setleri kullanılarak deneyler gerçekleştirilmiştir. Yapılan deneyler sonucunda, en iyi performansa sahip olan XLM-RoBERTa modeli için hiperparametre optimizasyonu yapılmış ve en kapsamlı veri setleri kullanılarak nihai Türkçe hakaret ve nefret söylemi otomatik tespit sistemi oluşturulmuştur. Oluşturulan Türkçe hakaret ve nefret söylemi otomatik tespit modeli, diğer çalışmalarla aynı test veri setini kullanarak karşılaştırılmıştır.

Anahtar Kelimeler: *Hakaret Söylemi, Nefret Söylemi, BERT, Doğal Dil İşleme*

Offensive Language and Hate Speech Detection in Turkish

Abstract

As a result of people communicating online, especially on social media platforms, the amount of user-generated content on the internet has increased. Due to the anonymous nature of these platforms, users can share content containing offensive language and hate speech. Such undesirable content can have negative effects on both individuals and societies. Therefore, it is important to detect and filter content that contains offensive language and hate speech. Detecting such content manually is challenging, which is why there is a need for automated methods. In recent years, there has been an increase in academic research on the detection of online offensive language and hate speech. Promising results have been achieved in the

* Iletişim e-posta: mehmetsalihkurt@hakkari.edu.tr

automatic detection of offensive language and hate speech in English using transfer learning models such as BERT. However, the number of studies on automatic detection of offensive language and hate speech in languages with limited resources such as Turkish is quite limited.

This study presents the results of efforts to develop an automatic detection system for offensive language and hate speech in the Turkish language. Firstly, an automatic labeling method was proposed to create a Turkish dataset, and using this method, a Turkish dataset for hate speech and offensive language was created. Experiments were conducted using various variants of the BERT model, which is considered state-of-the-art in natural language processing, along with various Turkish datasets related to offensive language and hate speech. Through these experiments, the XLM-RoBERTa model, which achieved the best performance, underwent hyperparameter optimization. Subsequently, using the most comprehensive datasets available, the final Turkish automatic detection system for offensive language and hate speech was developed. The developed Turkish automatic detection model for offensive language and hate speech was compared with other studies using the same test dataset.

Keywords: *Offensive Language, Hate Speech, BERT, Natural Language Processing*

1 Giriş

Çevrimiçi platformlarda kullanıcılar tarafından oluşturulan içeriklerin artması, ifade özgürlüğünü genişletme fırsatı sunarken, aynı zamanda bilgi kirliliği, şiddet ve aşağılayıcı söylemlerin yaygın olduğu bir ortamı da beraberinde getirmiştir. Sosyal medya siteleri, kullanıcıların anonimliklerini koruyarak nefret söylemi ve şiddet içeren içeriklerin yayılmasını kolaylaştırırken, bazı durumlarda gerçek dünyadaki şiddeti teşvik edebilmektedir [1]. Bu nedenle, sosyal medya platformları, şiddet, hakaret ve nefret söylemi gibi istenmeyen içerikleri barındırmaya müsait bir ortam oluşturabilmektedir [2, 3].

Sosyal medya platformlarında kullanıcıların oluşturduğu hakaret ve nefret söylemlerine karşı önlem alınmaması, demokratik bir toplumun gelişimi için önemli olan insan haklarının, hukukun üstünlüğünün ve ifade özgürlüğünün sekteye uğramasına neden olabilir. Bu durum, geniş çaplı çatışmalara ve şiddete yol açabilir [4]. Örneğin, Facebook'un Sri Lanka'da Müslüman karşıtı şiddeti kışkırtmakla suçlanması ve Myanmar'daki Rohingya nüfusuna yönelik olası soykırıma katkıda bulunacak şekilde nefret söylemi yaymakla suçlanması gibi olaylar, kullanıcıların sosyal medya platformlarında oluşturdukları şiddet, hakaret ve nefret söylemlerinin toplumsal olaylara sirayet ettiğini göstermektedir [5, 6]. Benzer şekilde, Twitter da nefret söylemiyle mücadelede yeterince etkili olmadığı gerekçesiyle eleştirilmiştir [7]. Bu nedenle, sosyal medya platformlarının bu tür içerikleri önlemek için daha etkili önlemler alması gerekmektedir.

Sosyal medya platformları, istenmeyen içeriklerin azaltılması için belirli kurallar oluşturmuştur ve kullanıcıların bu kurallara uyması gerekmektedir [8]. Kurallara uymayan gönderiler silinebilir veya kullanıcıların hesapları askıya alınabilir. Platformlar, sağlıklı, güvenilir ve kullanıcı dostu bir ortam sağlamak için gönderileri manuel olarak inceleyen moderatörlerle çalışmaktadır [9]. Ancak, bu moderasyon stratejisi, moderatörlerin hızı, argo ve jargonu anlama yetenekleri ve çok dilli içeriğe hakimiyeti gibi faktörlere bağlıdır. Ayrıca, veri akışının büyük hacmi nedeniyle her gönderiyi manuel olarak incelemek ve zararlı içeriği filtrelemek neredeyse imkansızdır. Bu nedenle, hakaret ve nefret söylemini tespit etmek için otomatik tekniklere ihtiyaç duyulmaktadır. Otomatik teknikler, bu tür içeriklerin tespit edilmesinde önemli ve kaçınılmaz bir rol oynamaktadır. Son yıllarda hakaret ve nefret söylemi ile ilgili araştırma sayısının artmasının başlıca sebepleri şunlardır:

Toplumsal Farkındalık: Toplumda artan farkındalık, hakaret ve nefret söylemiyle mücadeleye olan ilgiyi artırmıştır. Özellikle sosyal medya platformlarında yaygınlaşan bu tür içerikler, toplumda ciddi endişelere ve tepkilere yol açmıştır.

Yasal ve Politik Baskılar: Hükümetler ve kurumlar, hakaret ve nefret söyleminin yayılmasını kontrol altına almak için yasal düzenlemeler ve politikalar geliştirmekte ve bunları uygulamaya koymaktadır. Bu baskılar, araştırmacıları bu alanda çalışmaya teşvik etmektedir.

İletişim Teknolojilerindeki Gelişmeler: İnternetin ve sosyal medyanın yaygınlaşmasıyla birlikte insanlar arasındaki iletişim artmıştır. Bu da hakaret ve nefret söylemi içeren içeriklerin

çoğalmasına yol açmıştır. Bu durum, araştırmacıları bu tür içeriklerin tespiti ve önlenmesi için çalışmalara yönlendirmiştir.

Teknolojik İlerlemeler: Doğal dil işleme, makine öğrenimi ve yapay zeka gibi teknolojik alanlardaki ilerlemeler, hakaret ve nefret söylemiyle mücadelede otomatik tespit yöntemlerinin geliştirilmesine olanak sağlamıştır. Bu teknolojik imkanlar, araştırmacıları bu alanda daha fazla çalışmaya teşvik etmiştir.

Toplumsal Etki: Hakaret ve nefret söyleminin olumsuz etkileri, bireyler, toplumlar ve şirketler üzerinde ciddi sonuçlara yol açmaktadır. Bu etkilerin farkına varan insanlar, bu sorunla mücadele için araştırmalara ve çözümlere yönelmektedir.

Şirketlerin ve Sosyal Medya Platformlarının İlgisi: Sosyal medya platformları, kullanıcılar arasındaki iletişimi güvenli ve olumlu bir ortamda sağlamak için hakaret ve nefret söylemiyle mücadeleye önem vermektedir. Bu nedenle, bu platformlar ve diğer şirketler, bu alanda yapılan araştırmaları desteklemekte ve bütçe ayırmaktadır.

Bu faktörler, hakaret ve nefret söylemiyle ilgili araştırmaların artmasına ve bu alanda daha fazla çalışmanın yapılmasına yol açmıştır.

Hakaret ve nefret söylemini otomatik olarak tespit etme çalışmaları çoğunlukla İngilizce dilinde yapılmaktadır. Ancak bu çalışma, dünya genelinde milyonlarca kişi tarafından konuşulan Türkçe diline odaklanmaktadır. Türkçe, geniş bir coğrafyada büyük bir kullanıcı kitlesi tarafından aktif olarak kullanılmasına rağmen, hakaret ve nefret söylemi tespiti konusunda kapsamlı bir araştırmaya rastlanmamıştır. Bu çalışmada Türkçe diline yönelik hakaret ve nefret söylemi veri seti oluşturulmuş ve bu veri seti, makine öğrenme modellerinde kullanılarak otomatik hakaret ve nefret söylemi tespit sistemi geliştirilmiştir. Bu çalışma sonucunda, Türkçe dilinde yapılan en kapsamlı çalışma olması hedeflenen Türkçe hakaret ve nefret söylemi tespit sistemi geliştirilmesi amaçlanmaktadır. Bu çalışma, Türkçe dilinde hakaret ve nefret söylemini otomatik olarak tespit edebilen bir sistem geliştirme amacıyla literatüre şu katkıları sağlamıştır:

- Hakaret ve nefret söylemiyle ilgili detaylı bir literatür taraması gerçekleştirilmiş ve görev tanımları, kullanılan terimler, veri elde etme ve etiketleme yöntemleri, veriden özellik çıkarma ve sınıflandırma yöntemleri gibi alanlarda yapılan çalışmalar incelenmiştir.
- İlgili çalışmaların çoğu İngilizce dilinde gerçekleştirilmiştir. Bu çalışmada ise İngilizce dilinde el ile etiketlenmiş hakaret ve nefret söylemi veri setleri, çok dilli makine öğrenme modelleriyle Türkçe model geliştirme deneyleri için kullanılmış ve sonuçları paylaşılmıştır.
- Çalışma kapsamında farklı kaynaklardan faydalanarak kapsamlı bir Türkçe hakaret ve nefret söylemi veri seti oluşturulmuş ve bu veri seti araştırmacıların erişimine sunulmuştur.
- Oluşturulan Türkçe hakaret ve nefret söylemi veri seti, metin sınıflandırma alanında en iyi modellerin eğitiminde kullanılmış ve performans karşılaştırmaları yapılmıştır. En iyi performans gösteren modelin hiperparametre optimizasyonu gerçekleştirilerek nihai bir model oluşturulmuştur.
- Oluşturulan Türkçe hakaret ve nefret söylemi otomatik tespit modeli, aynı test veri setini kullanan modellerle karşılaştırılmış ve sonuçları analiz edilmiştir. Çalışma sonucunda geliştirilen model, kullanıcılar için bir web arayüzüyle sunulmuştur.

Bu katkılar, Türkçe dilinde hakaret ve nefret söylemini tespit etmek için geliştirilen sistemde önemli bir ilerleme sağlamış ve alanın literatürüne değerli bir katkı sunmuştur.

Literatürde kullanıcılar tarafından oluşturulan istenilmeyen söylemlerin (hakaret, küfür, cinsiyetçi, ırkçı, alay etme, trolleme, siber zorbalık, incitici ve küçük düşürücü söylemler) tespit edilmesiyle ilgili birçok araştırma bulunmaktadır. Hakaret söylemi insanlara yönelik tehditler, ayrımcılıklar, küfürler ve kaba hakaretlerdir [10]. Nefret söylemi ise ırka, etnik kökene, dine, engelliliğe, cinsiyete, yaşa veya cinsel yönelime dayalı olarak bir gruba saldıran veya küçük düşüren dildir [11]. Bu çalışmada hakaret ve nefret söylemi terimleri tercih edilmiştir.

Hakaret ve nefret söylemi otomatik tespiti ile ilgili araştırmalar çoğunlukla İngilizce dilinde gerçekleştirilmiştir. Davidson ve diğerleri (2017), nefret söylemini diğer hakaret söylemlerinden ayırt etmeyi amaçlamıştır [12]. Yapılan çalışmada, nefret söylemi, nefret söylemi içermeyen hakaret söylemi ve hiçbiri olmak üzere üç kategoriye ayrılan bir veri seti oluşturulmuştur. Çalışmada, açık bir şekilde nefret sözcükleri içermeyen tweetlerin

sınıflandırılmasının diğerlerine göre daha zor olduğu belirtilmiştir.

Nobata ve diğerleri (2016), hakaret söylemlerini tespit etmek için etiketli bir veri seti oluşturmuştur [11]. Hakaret söylemi terimi, aşağılayıcı dili, nefret söylemini ve küfürlü dili kapsayan bir terim olarak ele alınmıştır. Bu çalışmada, derin öğrenme yaklaşımlarından daha iyi performans gösteren doğal dil işleme özelliklerine sahip denetimli bir sınıflandırma metodolojisi geliştirilmiştir. Ayrıca, farklı alanlardan elde edilen verilerle eğitilen modellerin farklı zaman dilimlerindeki veriler üzerindeki performanslarına yönelik analizler yapılmıştır.

Waseem ve Hovy (2016), ırkçı ve cinsiyetçi hakaretleri belirlemek için eleştirel ırk teorisine dayalı bir kriter listesi sunmuşlardır [13]. Bu çalışmada, 16.000'den fazla tweet bu kriterlere göre etiketlenmiş ve herkese açık hale getirilmiştir. Ayrıca, karakter n-gramları ve diğer dil dışı özelliklerin sınıflandırmaya etkisi analiz edilmiştir. Karakter n-gram tabanlı yaklaşımların sınıflandırma performansına önemli katkı sağladığı belirtilmiştir.

Zampieri ve diğerleri (2019), hakaret söylemlerinin tespiti ve karakterizasyonu için yeni bir üç seviyeli hiyerarşik açıklama şeması önermişlerdir [14]. Bu şema, hakaret söyleminin türünü ve hedefini etiketlemek için OLID (Hakaret Söylemi Tespit Veri Seti - Offensive Language Identification Dataset) adlı İngilizce tweet veri kümesine uygulanmıştır. OLID üzerinde farklı makine öğrenimi modelleri kullanılarak her düzey için deneyler gerçekleştirilmiş ve gelecekteki çalışmalar için önemli bir temel oluşturulmuştur.

Türkçe Hakaret ve nefret söylemi otomatik tespiti için gerçekleştirilen araştırma sayısı oldukça azdır. Çöltekin (2020) tarafından Türkçe dilinde oluşturulan hakaret söylemi veri seti tanıtılmıştır [15]. Bu çalışmada Twitter'dan elde edilen veri seti analiz edilerek elle etiketlenmiştir. Aynı veri seti, SemEval-2020'de Türkçe veri seti olarak da kullanılmıştır [16]. Bu araştırmada hedeflenen Türkçe hakaret ve nefret söylemi tespit modelinin oluşturulmasında, Çöltekin tarafından oluşturulan veri setinin faydalı olabileceği düşünülmektedir. Söz konusu veri seti, elle etiketlenmiş hakaret ve nefret söylemi örneklerini içermesi bakımından değerlidir. Ancak, örnek sayısının kısıtlı olması ve sadece Twitter verilerini içermesi nedeniyle, her platformda kullanılabilecek kapsamlı bir modelin oluşturulması için yeterli olmadığı sonucuna varılmıştır. Türkçe hakaret ve nefret söylemi alanında yapılan başka araştırmalar da bulunmaktadır. Şahi ve diğerleri (2018) tarafından Türkçe dilinde kadınlara yönelik nefret söylemini tespit etmek için bir sınıflandırma modeli önerilmiştir [17]. Mayda ve diğerleri (2021) farklı ırklara ait tweetleri kullanarak Türkçe tweetlerden nefret ve hakaret söylemi tespit modeli oluşturmuştur [18]. Her iki çalışmada kullanılan veri setleri, örnek sayılarının sınırlı olmasının yanı sıra hakaret ve nefret söylemi kapsamının dar olmasından dolayı, kapsamlı bir model oluşturma hedefi açısından yetersiz kalmaktadır.

Bu çalışmada, literatürde incelenen Türkçe veri setlerindeki örnek sayılarının kapsamlı bir model oluşturmak için yetersiz olduğu tespit edilmiştir. Bu nedenle, çalışma kapsamında kapsamlı bir Türkçe veri seti oluşturma kararı alınmıştır.

2 Materyal ve Metod

Bu bölümde, Türkçe hakaret ve nefret söylemi otomatik tespit modeli oluşturmak için kullanılacak olan veri setleri ve makine öğrenme modelleri hakkında bilgiler sunulmuştur. İlk kısımda, literatürde bulunan Türkçe veri setleri tanıtılmış ve Türkçe veri seti oluşturulmasıyla ilgili bilgiler paylaşılmıştır. İkinci kısımda ise, metin sınıflandırma alanında yaygın olarak kullanılan makine öğrenme modelleri hakkında bilgiler verilmiştir.

2.1 Mevcut Veri Setleri

Bu çalışmada, Türkçe hakaret ve nefret söylemi tespit sistemi için kullanılabilecek veri setleriyle ilgili detaylı bir araştırma yapılmıştır. Araştırma sonucunda, tespit edilen veri setleriyle ilgili bilgiler Tablo 1'de listelenmiştir.

Tablo 1. Hakaret ve nefret söylemi otomatik tespiti için kullanılabilir Türkçe veri setleri.

Ad	Boyut	Kaynak
OffensEval 2020 Eğitim Seti	31756	[15, 16]
OffensEval 2020 Test Seti	3528	[15, 16]
Kıyafetimekarisma Veri Seti	318	[17]
Irksal Nefret Söylemi Veri Seti	1000	[18]

Tablo 1'de gösterilen bilgilere dayanarak, Türkçe dilinde hakaret ve nefret söylemi konusunda kullanılabilecek kaynakların sınırlı olduğu görülmektedir. Şahi ve diğerleri (2018) tarafından yapılan çalışmada cinsiyetçi tweetleri belirlemek için 318 tweet örneği kullanılmış, Mayda ve

diğerleri (2021) ise ırkçı tweetleri bulmak için 1000 tweet örneği üzerinde çalışmışlardır. Bu veri setleri, özel bir alana odaklanmış olduklarından hedeflenen model için son derece yetersiz kalmaktadır. Tabloda listelenen Türkçe veri setleri içinde, yalnızca Çöltekin (2020) tarafından paylaşılan veri setinin hedeflenen modelin oluşturulmasında faydalı olabileceği düşünülmektedir. Bu veri seti, el ile etiketlenmiş hakaret ve nefret söylemine ait örnekler içermesi bakımından değerlidir, ancak örnek sayısı sınırlı ve sadece Twitter verilerini içermesinden dolayı her platformda kullanılabilecek kapsamlı bir model oluşturmak için yeterli değildir. Bu nedenle, bu çalışmanın sürecinde kapsamlı bir Türkçe veri seti oluşturulması kararı alınmıştır.

2.2 Türkçe Veri Seti Oluşturulması

Türkçe hakaret ve nefret söylemi ile ilgili veri seti oluşturmak için literatürde kullanılan veri setlerinin oluşturulma yöntemleri incelenmiştir. Ancak, bu çalışmada hedeflenen kapsamlı Türkçe hakaret ve nefret söylemi veri setini elle etiketlemenin insan kaynağı ve bütçe yetersizlikleri nedeniyle mümkün olmadığı sonucuna varılmıştır. Bunun üzerine, Türkçe hakaret ve nefret söylemi içeren ve içermeyen metinler için farklı kaynaklar ve yöntemler kullanılarak, makine öğrenme modelleri için kullanılabilecek etiketli bir veri seti oluşturulmuştur. Türkçe hakaret ve nefret söylemi veri setinin oluşturulması sürecinde pozitif, negatif ve etiketsiz metin örneklerinin nasıl elde edildiğiyle ilgili bilgiler aşağıda ayrı ayrı sunulmuştur.

Pozitif metin örneklerinin oluşturulması: Pozitif metin örnekleri için Twitter web sitesi kullanılmıştır. Türkçe hakaret ve nefret söylemiyle ilgili metinleri elde etmek için literatürdeki çalışmalarda da kullanılan bir yöntem olan anahtar kelimelerin seçilmesi ve bu anahtar kelimelerin geçtiği tweet'lerin elde edilmesi amaçlanmıştır. Anahtar kelimeler için github.com ve tscorpus.com gibi kaynaklardaki Türkçe projelerde kullanılan kara listelerdeki kelimeler birleştirilmiştir. Bu kara listede, hakaret, müstehcenlik, küfür, ırkçılık, cinsiyetçilik anlamlarını içeren ve cümle içinde kullanıldığında kesinlikle hakaret ve nefret söylemi anlamını taşıyan anahtar kelimeler seçilmiştir. Bu anahtar kelimeler, github.com web sitesinde proje sayfasında paylaşılmıştır [21]. Seçilen anahtar kelimelerin geçtiği Türkçe tweet'leri çekmek için Python programlama dilinde yazılmış olan Scrapy kütüphanesi kullanılmıştır. Elde edilen tweet'ler, hakaret ve nefret söylemi içerdiğinden emin olduğumuz için elle etiketlemeye gerek duyulmadan otomatik olarak pozitif örnek olarak etiketlenmiştir. Bu işlemler sonucunda hakaret ve nefret söylemini içeren toplamda 1.67 milyon tweet elde edilmiştir.

Negatif metin örneklerinin oluşturulması: Türkçe dilinde oluşturulmayı hedeflediğimiz hakaret ve nefret söylemi veri setinin negatif (hakaret ve nefret söylemi ile ilgili metinleri içermeyen) örneklerini oluşturmak için Wikipedia web sitesinde yer alan Türkçe içerikler kullanılmıştır. Negatif örneklerin oluşturulması, "Wikipedia sitesinden elde edilen Türkçe metinlerde hakaret ve nefret söylemi olmaması veya önemsiz düzeyde olması" varsayımına dayanmaktadır. Bu varsayımın doğruluğunu kontrol etmek amacıyla Wikipedia'dan elde edilen 10.000 cümle tek tek incelenmiş ve hakaret ve nefret söylemi içermediği tespit edilmiştir. Bu test sonucunda, varsayımımızı destekleyen Wikipedia web sayfasındaki metinlerden elde edilen cümleler otomatik olarak negatif örnek olarak etiketlenmiştir. Böylelikle, hakaret ve nefret söylemi içermeyen 2.51 milyon cümle Wikipedia sitesinden elde edilmiştir.

Etiketsiz metin örneklerinin oluşturulması: Türkçe hakaret ve nefret söylemi veri setinin pozitif metin örneklerini oluşturmak amacıyla Türkçe istenilmeyen kelimeleri içeren kara listeler kullanılmıştır. Bu kara listelerdeki kelimeler, Twitter platformundan çekilen tweet'lerde kullanılarak pozitif örnekler olarak otomatik olarak etiketlenmiştir. Etiketsiz metin örneklerinin oluşturulmasının amacı, makine öğrenme modellerinin tweet'lerde geçen bu anahtar kelimeleri ezberlemesini engellemek ve Twitter metinlerinde çeşitlilik sağlamaktır. Bu doğrultuda, Türkçede metinlerde sıkça kullanılan "ve", "veya", "ile", "için", "gibi", "ama", "kadar" gibi edat ve bağlaçlar anahtar kelimeler olarak seçilmiş ve bu anahtar kelimelerin geçtiği Türkçe tweet'ler çekilmiştir. Bu işlem için Python programlama dilinde yazılmış olan Scrapy kütüphanesi kullanılmıştır. Elde edilen tweet'lerin hakaret ve nefret söylemi içerip içermediği bilinmediği için ilk aşamada etiketlenme işlemi gerçekleştirilmemiştir. Bu süreç sonucunda toplamda 10.67 milyon tweet'ten oluşan etiketlenmemiş bir veri seti elde edilmiştir. Burada oluşturulan etiketsiz Türkçe Twitter veri setine ek olarak Tscorpus.com sitesinin Twitter veri seti (Sezer, 2020) ve Yıldız Teknik Üniversitesinin Kemik Twitter veri seti (Kemik,

2020) gibi diğer Twitter veri setlerinin de olduğu tespit edilmiştir [19, 20]. Bu Twitter veri setleri de Türkçe tweet'lerden oluşmaktadır ve hakaret ve nefret söylemi örnek çeşitliliğine katkıda bulunması amacıyla ilgili kaynaklardan temin edilmiştir.

Türkçe hakaret ve nefret söylemi için elde edilen pozitif ve negatif metin örnekleri birleştirilerek "Hakaret ve Nefret Söylemi Veri Seti" oluşturulmuştur. Bu veri seti daha sonra eğitim ve test verisi olarak ikiye ayrılmıştır. Bu ayrım yapılırken, pozitif metin örneklerinde bulunan anahtar kelimelerin eğitim ve test setleri arasında dengeli bir şekilde dağılımı sağlanmıştır. Tablo 2'de, bu araştırma kapsamında oluşturulan veri seti ile birlikte diğer kaynaklardan temin edilen veri setleri hakkında detaylı bilgiler paylaşılmıştır.

Tablo 2'de listelenen veri setlerinden OffensEval 2020 Türkçe veri seti, el ile etiketlenmiş olmasından dolayı, model eğitimi ve testi açısından son derece kritik bir rol oynamaktadır. Bu çalışmada, hem modellerin eğitiminde hem de performansının değerlendirilmesinde OffensEval 2020 Türkçe veri setleri etkin bir şekilde kullanılmıştır. Ancak, OffensEval 2020 Türkçe veri setindeki örnek sayısı yetersiz olduğundan, model eğitimlerinde otomatik olarak etiketlenmiş hakaret ve nefret söylemi veri setleri ve etiketsiz Twitter veri setleri de kullanılmıştır.

Tablo 2. Hakaret ve nefret söylemi otomatik tespiti için kullanılabilir Türkçe veri setleri.

Ad	Etiket Durumu	Boyut	Kaynak
OffensEval 2020 Eğitim Seti	Elle Etiketlenmiş	31756	[15, 16]
OffensEval 2020 Test Seti	Elle Etiketlenmiş	3528	[15, 16]
Hakaret ve Nefret Söylemi Eğitim Seti	Otomatik Etiketlenmiş	3.79M	
Hakaret ve Nefret Söylemi Test Seti	Otomatik Etiketlenmiş	385K	
Twitter Veri Seti	Etiketsiz	10.67M	
TSCorpus Twitter Seti	Etiketsiz	17 Milyon	[19]
Kemik Twitter Seti	Etiketsiz	3.5 Milyon	[20]

Türkçe hakaret ve nefret söylemi sistemi için yarı denetimli bir makine öğrenme yaklaşımı benimsenmiştir. Bunun için model eğitiminde önce elle etiketlenmiş OffensEval 2020 Türkçe veri seti ve otomatik olarak etiketlenmiş hakaret ve nefret söylemi veri seti kullanılmıştır. Daha sonra oluşturulan model, Tablo 2'de listelenen etiketsiz Twitter veri setlerindeki metinleri sınıflandırmak için kullanılmıştır. Sınıflandırılan bu metinler, etiketli eğitim setlerine dahil edilerek hakaret ve nefret söylemini içeren örneklerin sayısı ve çeşitliliği artırılmıştır. Sonuç olarak, Tablo 2'de listelenen tüm veri setleri, bu çalışma kapsamında oluşturulması hedeflenen Türkçe hakaret ve nefret söylemi tespit sistemi için kullanılmıştır.

Bu çalışmada oluşturulan hakaret ve nefret söylemi veri seti, farklı platformlara ait metinleri içermesi ve Türkçe dilinde oluşturulan en kapsamlı hakaret ve nefret söylemi veri seti olması açısından önemlidir. Bu çalışmada gerçekleştirilen tüm kodlamalar ve oluşturulan hakaret ve nefret söylemi veri seti, bu alanda çalışma yapmak isteyen araştırmacıların kullanımına sunulmuştur [21].

2.3 Kullanılan Makine Öğrenme Modelleri

Bu çalışmada Türkçe hakaret ve nefret söylem otomatik tespit modelini oluşturmak için Trannsformer modellerinden BERT modeli ve varyantları kullanılmıştır.

2.3.1 BERT (Bidirectional Encoder Representations from Transformers)

BERT, Devlin ve diğerleri tarafından 2018 yılında geliştirilmiş bir doğal dil işleme modelidir [22]. Bu model, önceden eğitilmiş dil modellerinin doğal dil işlemedeki yeni bir standardı olarak kabul edilmiş ve ELMo, ULMFiT ve OpenAI'nin GPT girişimleriyle birlikte doğal dil işleme alanında yeni bir çağın başlamasına öncülük etmiştir [23-25]. Bu model, hem önceden eğitimli bir dil modeli olarak hem de göreve özgü ince ayarlar uygulanarak kullanılmaktadır. BERT, doğal dil işleme alanında büyük bir etki yaratmış ve birçok görevde en iyi sonuçları elde etmiştir.

BERT, Transformer modelinin bir türevidir ve diğer dil modellerinden farklı olarak çift yönlü bir bağlamı dikkate alır. Bu, hem önceki hem de sonraki kelimelerin bağlamını kullanarak bir kelimenin

anlamını belirleyebilmesini sağlar. BERT'in önceden eğitim süreci, büyük bir metin veri kümesi üzerinde gerçekleştirilir ve maskelenmiş dil modellemesi ve cümle seviyesi sınıflandırma görevleriyle eğitilir. Ardından, göreve özgü ince ayar için kullanılır, bu da modelin belirli doğal dil işleme görevlerinde daha iyi performans göstermesini sağlar. BERT, doğal dil işleme alanında önemli bir ilerleme sağlamış ve çeşitli görevlerde yüksek başarı elde etmiştir. Bu model, metin anlama, duygu analizi, soru-cevap sistemleri ve dil çevirisi gibi birçok alanda kullanılmaktadır.

2.3.2 BERT'in Varyantları

Bu bölümde, bu çalışmada kullanılan deneylerde yer alan BERT modelinin farklı varyantları olan DistilBERT, RoBERTa ve XLM-RoBERTa hakkında bilgiler sunulmuştur.

DistilBERT: BERT modelinin daha hafif bir versiyonudur ve BERT'in özünü daha küçük bir yapıda yoğunlaştırır [26]. Modelin temel amacı, BERT'in boyutunu ve hesaplama maliyetini azaltarak daha hızlı ve daha hafif bir alternatif sunmaktır. DistilBERT, daha hafif yapısıyla kaynak ve hesaplama gücü kısıtlı ortamlarda kullanımı kolaylaştırırken BERT'in dil anlama yeteneğinin önemli bir kısmını korur. Bu nedenle, doğal dil işleme görevlerinde etkili bir seçenek olarak tercih edilmektedir.

RoBERTa: BERT modelinin performansını geliştirmek için yapılan bir varyasyondur [27]. Daha büyük bir eğitim veri kümesi ve ek eğitim teknikleriyle eğitilen RoBERTa, dil temsillerinde daha iyi bir performans sergiler. RoBERTa, BERT'in mimarisini takip eder, ancak eğitim sürecinde bazı değişiklikler yapar. Bu değişiklikler arasında "Sonraki Cümle Tahmini" görevinin kaldırılması, dinamik maskeleme, daha büyük toplu işlemler ve Byte-level BPE kullanımı bulunur. RoBERTa, doğal dil işleme görevlerinde daha iyi sonuçlar elde etmiş ve dildeki anlam ve yapısal örüntüleri daha iyi yakalama yeteneğine sahip olduğunu kanıtlamıştır.

XLM-RoBERTa: Çoklu dil desteği sağlamak amacıyla özel olarak tasarlanmış RoBERTa tabanlı bir modeldir [28]. Bu modelde geniş bir dil veri kümesi üzerinde eğitilerek geliştirilmiştir. Geniş bir dil veri seti üzerinde eğitilerek oluşturulan bu modelde farklı diller arasındaki dil anlama becerisi geliştirilmiştir. XLM-RoBERTa, çeşitli dillerdeki metinleri işleyebilme yeteneği sayesinde dil çevirisi, çok dilli metin sınıflandırması ve diğer çok dilli doğal dil işleme görevlerinde kullanılabilir. Ayrıca, dil sınırlarını aşarak farklı diller arasında transfer öğrenimini kolaylaştıran bir yaklaşım sunmaktadır. Bu model, dil özelliklerini daha iyi öğrenerek farklı dillerdeki semantik ve sözdizimsel desenleri anlama yeteneğini arttırmaktadır. Yüksek performans sergileyen XLM-RoBERTa, çeşitli dil kaynaklarını etkin bir şekilde kullanarak çok dilli metinlerdeki anlamı yakalama konusunda dikkate değer bir başarı elde etmiştir.

mBERT: mBERT, çok dilli doğal dil işleme için tasarlanmış bir dil modelidir [22]. Bu model, farklı diller arasındaki dil anlama yeteneğini geliştirmek amacıyla kullanılmaktadır. mBERT, çeşitli dillerdeki metinleri işleyebilmekte ve farklı dillerdeki kelime anlamlarını ve ilişkilerini daha iyi yakalamak için dil temsillerini öğrenebilmektedir. Dil çevirisi, çok dilli metin sınıflandırması gibi görevlerde etkin bir şekilde kullanılabilmekte ve diller arasında transfer öğrenimine olanak sağlamaktadır.

3 Bulgular ve Tartışma

3.1 Hakaret ve Nefret Söylemi Otomatik Tespit Modelinin Oluşturulması

Bu bölümde bu çalışmada oluşturulan Türkçe hakaret ve nefret söylemi veri seti ile OffensEval 2020 Türkçe veri seti kullanılarak Türkçe hakaret ve nefret söylemi otomatik tespit modelinin geliştirilmesi ile ilgili deneylerin sonuçları paylaşılmıştır. Türkçe hakaret ve nefret söylemi otomatik tespit sistemi oluşturma sürecinde, modellerin eğitimi büyük bir sorun olarak ortaya çıkmaktadır. Bu süreçte, dönüştürücü modellerin büyük veri setleriyle eğitilmesi maliyetli ve yüksek GPU işlem gücü gerektirmektedir. Bu sorunların üstesinden gelmek için, Google Colaboratory ve Kaggle gibi platformlardan temin edilen bilgisayarlar kullanılmıştır. Bu bilgisayarlar, NVidia K80 GPU özelliklerine sahip olduğu için modellerin eğitim sürelerini ve parametre optimizasyonunu daha etkili bir şekilde gerçekleştirmiştir. Bu bölümdeki tüm model eğitimleri ve testleri bu platformlarda yapılmıştır.

3.1.1 Model ve Veri Seti ile ilgili Deneyler

Bu bölümde, Türkçe hakaret ve nefret söylemi veri seti ile OffensEval 2020 Türkçe veri seti kullanılarak, yüksek doğruluk seviyesinde sınıflandırma yapabilen kapsamlı bir Türkçe hakaret ve nefret söylemi otomatik tespit modeli oluşturulması amaçlanmaktadır. Bu amaç doğrultusunda, doğal dil işleme alanında en iyi metin sınıflandırma modelleri, söz konusu veri

setleri üzerinde önce eğitilmiş ve ardından test edilerek performansları karşılaştırılmıştır.

Bu çalışmada, oluşturulacak modellerin otomatik etiketli veri setinde bulunan anahtar kelimelerin ezberlemesi ihtimali de göz önünde bulundurulmuştur. Bu durumdan kaçınmak için modellerin hem eğitiminde hem de testinde elle etiketlendiği için güvenilir bir veri seti olarak OffensEval 2020 Türkçe veri seti kullanılmıştır. Makine öğrenme modelleri olarak, metin sınıflandırma alanında başarılı sonuçlar elde edilen DistilBERT, BERT, RoBERTa ve çok dilli XLM-RoBERTa modelleri kullanılmıştır. Bu modellerin eğitiminde, OffensEval 2020 eğitim seti ile birlikte otomatik etiketlenmiş Türkçe hakaret ve nefret söylemi eğitim seti kullanılmıştır. Oluşturulan modellerin performansını değerlendirmek için OffensEval 2020 test seti ve Türkçe hakaret ve nefret söylemi test seti kullanılmıştır. Modellerin f1 skorları ve doğruluk değerleri Tablo 3'te listelenmiştir. Burada gerçekleştirilen model eğitimlerinden ve testlerden elde edilen sonuçlar aşağıdaki şekilde özetlenebilir:

- OffensEval 2020 eğitim seti, sadece 31756 adet örneği kapsamaktadır. Bu örnekler içerisinde 6131'i hakaret ve nefret söylemi içerirken, 25625'i ise hakaret ve nefret söylemi içermemektedir. Bu elle etiketlenmiş veri seti, model eğitiminde belirli bir seviyeye ulaşmayı sağlamaktadır. Veri setine eklenen otomatik etiketli yüzbinlerce örneğin, model performansını çok az bir şekilde artırdığı gözlenmiştir. Bu sonuçlar, otomatik etiketlenmiş veri setinin modelin performansına olumlu bir katkı sağladığını, ancak elle etiketlenmiş veri setinin model eğitiminde en etkin kaynak olduğunu göstermektedir.
- Elde edilen makine öğrenme modelleri, otomatik etiketlenmiş Türkçe hakaret ve nefret söylemi test setinde yüksek doğruluk değerleri ve düşük f1 skorları sergilemiştir. Bu durum, modellerin eğitim verisindeki örnekleri ezberlemesi sonucunda ortaya çıkan bir durumdur. Bu nedenle, performans değerlendirmesinde doğruluk yerine f1 skoru kullanılmıştır. Sonuçlar, modellerin Türkçe hakaret ve nefret söylemi eğitim setindeki anahtar kelimeleri ezberlediğini göstermektedir. Bu tür sorunlar nedeniyle, elle etiketlenmiş OffensEval 2020 test seti, otomatik etiketlenmiş Türkçe hakaret ve nefret söylemi test setinden daha güvenilir bir performans değerlendirme aracı olarak tercih edilmiştir.
- Deneylerde, farklı eğitim veri setleriyle eğitilen dönüştürücü modellerin performansı test veri setleri üzerinde değerlendirilmiştir. Başlangıç deneylerinde sadece OffensEval 2020 eğitim seti kullanılarak DistilBERT, BERT, RoBERTa ve XLM-RoBERTa modelleri için sırasıyla 0.74, 0.76, 0.79 ve 0.79 f1 skorları elde edilmiştir. Daha sonra, OffensEval 2020 eğitim setine ek olarak 200 bin adet otomatik etiketli hakaret ve nefret söylemi veri seti kullanılarak modeller eğitilmiş ve DistilBERT, BERT, RoBERTa ve XLM-RoBERTa modelleri için sırasıyla 0.74, 0.77, 0.79 ve 0.80 f1 skorları alınmıştır.

Tablo 3. Türkçe hakaret ve nefret söylemi tespiti için oluşturulan modeller ve sonuçları.

Eğitim Verisi	Model Bilgisi	OffensEval 2020 Test Seti		Hakaret ve Nefret Söylemi Test Seti	
		Doğruluk	F1 Skoru	Doğruluk	F1 Skoru
OffensEval 2020 Eğitim Seti	DistilBERT	0.84	0.74	0.90	0.45
	BERT	0.85	0.76	0.91	0.46
	RoBERTa	0.88	0.79	0.93	0.47
	XLM-RoBERTa	0.88	0.79	0.93	0.47
OffensEval 2020 Eğitim Seti + Hakaret ve Nefret Söylemi Eğitim Seti (200K)	DistilBERT	0.85	0.74	0.91	0.46
	BERT	0.86	0.77	0.92	0.47
	RoBERTa	0.89	0.79	0.94	0.47
	XLM-RoBERTa	0.88	0.80	0.95	0.48
OffensEval 2020 Eğitim Seti + Hakaret ve Nefret Söylemi Eğitim Seti (500K)	DistilBERT	0.86	0.75	0.93	0.47
	BERT	0.87	0.77	0.98	0.48
	RoBERTa	0.89	0.80	0.99	0.49
	XLM-RoBERTa	0.90	0.81	0.99	0.49

Tüm bu deneyler sonucunda, Türkçe hakaret ve nefret söylemi için en uygun makine öğrenme modelinin XLM-RoBERTa modeli olduğu belirlenmiştir. Otomatik etiketlenen hakaret ve nefret söylemi veri setinin model eğitimine olumlu katkı sağladığı ve kullanım miktarının artmasıyla model performansının iyileştiği gözlemlenmiştir. Bu nedenle, nihai Türkçe hakaret ve nefret söylemi modelinin eğitiminde bu veri setinin tamamı kullanılmıştır.

- Yapılan çalışmada, Türkçe hakaret ve nefret söylemi örnek çeşitliliğini arttırmak amacıyla etiketsiz Twitter veri setleri kullanılmıştır. Bu veri setindeki tweet'ler, OffensEval 2020 eğitim setiyle eğitilen makine öğrenme modelleri tarafından sınıflandırılmış ve etiketlenmiştir. Elde edilen etiketlenmiş tweet'ler, yeni makine öğrenme modellerinin eğitiminde kullanılmıştır. Ancak, bu tweet'lerin mevcut test veri setleri üzerinde modellerin performansına olan etkileri belirlenememiştir. Daha fazla örnek sayısına sahip elle etiketlenmiş test verilerinin kullanılması, bu veri setinin modellerin performansının daha iyi değerlendirilmesi için gereklidir. Bununla birlikte, mevcut veri setinin modellerin performansını olumsuz etkilememesi ve Türkçe hakaret ve nefret söylemi örnek çeşitliliğini arttırması nedeniyle kullanımına karar verilmiştir.

3.1.2 Nihai Modelin Oluşturulması

Türkçe hakaret ve nefret söylemi tespit modelinin oluşturulması için önceki aşamalarda dönüştürücü modellerin performanslarının karşılaştırılması ve mevcut veri setlerinin model performansına etkisi ile ilgili deneyler gerçekleştirilmiştir. Deneylerin sonucunda, mevcut veri setleri üzerinde XLM-RoBERTa dönüştürücü modelinin en iyi performansı gösterdiği belirlenmiş ve bu model için hiperparametre optimizasyonu yapılmıştır. Bu hiperparametre optimizasyonu sonucunda maksimum metin uzunluğu 128, öğrenme oranı 5e-6 (0.000005), küme(batch) büyüklüğü 16 ve seyretme oranı 0.3 olarak belirlenmiştir. Model eğitimi için elle etiketlenmiş Türkçe OffensEval 2020 eğitim seti ile otomatik olarak etiketlenmiş Türkçe hakaret ve nefret söylemi eğitim setinin uygun ve kullanılabilir durumda olduğu doğrulanmıştır. Bu iki eğitim setine ek olarak, örnek çeşitliliğini artırmak amacıyla OffensEval 2020 eğitim seti ile eğitilen makine öğrenme modeli tarafından sınıflandırılmış ve etiketlenmiş tweet'ler de kullanılarak nihai modelin eğitimi gerçekleştirilmiştir.

Yapılan model performans testlerinde, otomatik olarak etiketlenmiş Türkçe hakaret ve nefret söylemi test setinin içerdiği anahtar kelimelerin model tarafından ezberlendiği tespit edilmiştir. Bu durum, model performans test sonuçlarının yanıltıcı olabileceğine işaret etmektedir. Bu nedenle, otomatik etiketlenmiş Türkçe hakaret ve nefret söylemi test setinin kullanıma uygun olmadığına karar verilmiştir. Bunun yerine, model performansını ölçmek için elle etiketlenmiş Türkçe OffensEval 2020 test setinin kullanılmasına karar verilmiştir. Tablo 4'te, Türkçe hakaret ve nefret söylemi otomatik tespiti için oluşturulan modellerin eğitim veri setleri ve OffensEval 2020 test veri seti üzerinde elde edilen f1 skorları ve doğruluk skorları listelenmektedir.

Tablo 4, farklı eğitim veri setleri ile eğitilen XLM-RoBERTa modellerine dair bilgileri sunmaktadır. Bu modellerin eğitimi Google Colaboratory ve Kaggle platformlarından temin edilen NVidia K80 GPU'ya sahip bilgisayarlarla gerçekleştirilmiştir. Toplamda 8 adet XLM-RoBERTa modeli oluşturulmuş ve elde edilen sonuçlar aşağıda özetlenmiştir:

1. Model: Sadece OffensEval 2020 eğitim veri seti kullanılarak eğitilmiş. 10 iterasyon sonucunda elde edilen model, OffensEval 2020 test verisinde 0.88 doğruluk ve 0.79 f1 skoru elde etmiştir.

2. Model:Türkçe hakaret ve nefret söylemi eğitim veri setinden 100 bin örnek ile eğitilmiştir. 8 iterasyon boyunca süren bu eğitimde elde edilen model, test verisinde 0.78 doğruluk ve 0.67 f1 skoru elde etmiştir.

Tablo 4. Türkçe hakaret ve nefret söylemi otomatik tespiti için oluşturulan modeller.

Model No	Eğitim Verisi	Eğitim Süresi	OffensEval 2020 Test Seti	
			Doğruluk	F1 Skoru
1	OffensEval 2020 Eğitim Seti	10 x 21 dk	0.88	0.79
2	Hakaret ve Nefret Söylemi Eğitim Seti (100K)	8 x 72 dk	0.78	0.67
3	OffensEval 2020 Eğitim Seti Hakaret ve Nefret Söylemi Eğitim Seti (100K)	7 x 91 dk	0.88	0.79
4	OffensEval 2020 Eğitim Seti (x3) Hakaret ve Nefret Söylemi Eğitim Seti (100K)	9 x 129 dk	0.89	0.79
5	OffensEval 2020 Eğitim Seti (x2) Hakaret ve Nefret Söylemi Eğitim Seti (100K) (*)Twitter Veri Seti (100K)	8 x 183 dk	0.88	0.79
6	OffensEval 2020 Eğitim Seti (x3) Hakaret ve Nefret Söylemi Eğitim Seti (250K) (*)Twitter Veri Seti (100K)	8 x 311 dk	0.89	0.81
7	OffensEval 2020 Eğitim Seti (x3) Hakaret ve Nefret Söylemi Eğitim Seti (100K) (*)Twitter Veri Seti (250K)	10 x 321 dk	0.89	0.81
8	OffensEval 2020 Eğitim Seti (x10) Hakaret ve Nefret Söylemi Eğitim Seti (1M) (*)Twitter Veri Seti (1M)	9 x 1640 dk	0.89	0.82

3. Model: OffensEval 2020 eğitim veri seti ve Türkçe hakaret ve nefret söylemi eğitim setinden 100 bin örnek kullanılarak eğitilmiştir. 7 iterasyon sonucunda elde edilen model, test verisinde 0.88 doğruluk ve 0.79 f1 skoru elde etmiştir.

4. Model: Model: OffensEval 2020 eğitim veri setinin 3 katı Türkçe hakaret ve nefret söylemi eğitim seti ile birlikte eğitilmiştir. Bu eğitim süreci toplamda 9 iterasyon ve 1161 dakika sürmüştür. Test verisinde 0.89 doğruluk ve 0.79 f1 skoru elde edilmiştir.

5. Model: OffensEval 2020 eğitim veri setinin 2 katı Türkçe hakaret ve nefret söylemi eğitim seti ve makine öğrenme modeli tarafından sınıflandırılmış 100 bin tweet kullanılarak eğitilmiştir. Bu eğitim süreci toplamda 8 iterasyon ve 1464 dakika sürmüştür. Test verisinde 0.88 doğruluk ve 0.79 f1 skoru elde edilmiştir.

6. Model: OffensEval 2020 eğitim veri setinin 3 katı Türkçe hakaret ve nefret söylemi eğitim seti ve makine öğrenme modeli tarafından sınıflandırılmış 250 bin tweet kullanılarak eğitilmiştir. Bu eğitim süreci toplamda 8 iterasyon ve 2488 dakika sürmüştür. Test verisinde 0.89 doğruluk ve 0.81 f1 skoru elde edilmiştir.

7. Model: OffensEval 2020 eğitim veri setinin 3 katı Türkçe hakaret ve nefret söylemi eğitim seti ve makine öğrenme modeli tarafından sınıflandırılmış 100 bin tweet kullanılarak eğitilmiştir. Bu eğitim süreci toplamda 10 iterasyon ve 3210 dakika sürmüştür. Test verisinde 0.89 doğruluk ve 0.81 f1 skoru elde edilmiştir.

8. Model: OffensEval 2020 eğitim veri setinin 10 katı Türkçe hakaret ve nefret söylemi eğitim seti ve makine öğrenme modeli tarafından sınıflandırılmış 1 milyon metin kullanılarak eğitilmiştir. Bu eğitim süreci toplamda 9 iterasyon ve 14760 dakika sürmüştür. Test verisinde 0.89 doğruluk ve 0.82 f1 skoru elde edilmiştir.

Tablo 4'te listelenmiş olan XLM-RoBERTa modelleri kullanılan eğitim setlerinde büyük farklılıklar bulunmasına rağmen test veri seti üzerinde birbirine yakın performans ölçümleri yapılmıştır. Bu durum, OffensEval 2020 test setinin sınırlı örnek sayısı nedeniyle kapsamlı bir modelin performansını tam olarak değerlendirememesinden kaynaklanmaktadır. Oluşturulan XLM-RoBERTa modelleri arasında en iyi sonuçları veren ve kapsamlı bir eğitim seti kullanan 8 numaralı model, nihai model olarak seçilmiştir. Bu model, Türkçe hakaret ve nefret söylemini otomatik olarak tespit etmek amacıyla geliştirilen bir web arayüzü aracılığıyla kullanıcılara sunulmuştur.

Bu çalışmada oluşturulan nihai model, OffensEval 2020 test veri seti üzerinde 0.89 doğruluk skoru ve

0.82 f1 skoru elde etmiştir. Bu sonuca göre oluşturulan model aynı test seti üzerinde 0.77 f1 skoru elde eden Çöltekin (2020) çalışmasındaki modelden daha iyi performans göstermektedir. Ayrıca, bu model, OffensEval 2020 yarışmasında önerilen en iyi modellere yakın performans göstermiştir.

Bu araştırmada, oluşturulan modellerin performansını değerlendirmek ve diğer çalışmalarla (Çöltekin, 2020; Zampieri ve diğerleri, 2020) karşılaştırmak için OffensEval 2020 test veri seti kullanılmıştır. Ancak, bu çalışmada, milyonlarca hakaret ve nefret söylemi metin örneğiyle eğitilen makine öğrenme modellerinin performansını değerlendirmek ve karşılaştırmak için yalnızca 3528 metin örneği içeren OffensEval 2020 test setinin kullanılması önemli bir eksiklik olarak değerlendirilmektedir. Bununla birlikte, oluşturulan hakaret ve nefret söylemi otomatik tespit modelinin web arayüzü sorgulamalarında Türkçe hakaret ve nefret söylemlerini önemli ölçüde tespit ettiği gözlemlenmiştir, ancak daha sağlıklı bir test için daha kapsamlı veri setlerine ihtiyaç duyulmaktadır.

Bu araştırma, veri seti etiketlemek için önerilen otomatik etiketleme yöntemi, bu yöntemle oluşturulan Türkçe hakaret ve nefret söylemi veri seti, gerçekleştirilen deneyler ve oluşturulan Türkçe hakaret ve nefret söylemi otomatik tespiti modellerini içermektedir. Bu araştırma özgün olmakla birlikte Türkçe hakaret ve nefret söylemi otomatik tespiti alanındaki çalışmalar için öncü bir nitelik taşımaktadır.

4 Sonuç

Bu araştırmada Türkçe hakaret ve nefret söylemlerinin otomatik tespiti için bir makine öğrenme modeli içeren bir sistem önerilmiştir. Elde edilen sonuçlar ve iyileştirme önerileri şunlardır:

- Türkçe dilinde hakaret ve nefret söylemi otomatik tespiti alanında literatürde çok az çalışma bulunmaktadır. Türkçe hakaret ve nefret söylemi ile ilgili etiketli veri seti temin edilecebilecek sadece bir adet araştırma tespit edilmiştir. Söz konusu bu çalışmadaki örnek sayısı da kısıtlı olduğu için Türkçe elle etiketlenmiş veri seti konusunda büyük bir eksiklik olduğu vurgulanmıştır.
- Literatürde hakaret ve nefret söylemi ile ilgili çalışmaların büyük çoğunluğu İngilizce dilinde gerçekleştirilmiştir. Bu çalışmada, Türkçe ve İngilizce arasındaki dil farklılıkları (kaynak kısıtlılığı, dilbilimsel farklılıklar) yanı sıra Türkçe konuşanlar ve İngilizce konuşanlar arasındaki kültürel, sosyal ve politik farklılıklar da önemli zorluklara yol açmıştır.
- Bu çalışmada, Türkçe hakaret ve nefret söylemi modelini geliştirmek için özgün bir yöntem kullanılarak otomatik etiketli bir Türkçe veri seti oluşturulmuştur. Oluşturulan veri seti, eğitim ve test olarak ikiye ayrılmıştır. Yapılan deneyler sonucunda, otomatik etiketli eğitim setinin modelin performansını iyileştirdiği ancak test setinin modelin performansını değerlendirmede yetersiz olduğu tespit edilmiştir.
- Bu çalışmada, hem elle etiketlenmiş OffensEval 2020 Türkçe eğitim seti hem de otomatik etiketlenmiş Türkçe hakaret ve nefret söylemi eğitim seti kullanılarak makine öğrenme modelleri eğitilmiştir. Oluşturulan bu modellerin performansı OffensEval 2020 Türkçe test seti üzerinde karşılaştırılmış ve en yüksek f1 skoru XLM-RoBERTa dönüştürücü modeliyle elde edilmiştir.
- Bu çalışma sonucunda, farklı eğitim veri setlerinin (OffensEval 2020 eğitim seti, 1 milyon metin örneğine sahip Türkçe hakaret ve nefret söylemi eğitim seti ve makine öğrenme modelinin sınıflandırmasıyla etiketlenmiş 1 milyon metin örneği) birlikte kullanılarak XLM-RoBERTa modelinin eğitildiği ve bu modelin OffensEval 2020 test veri seti üzerinde 0.82 f1 skoru elde ettiği belirlenmiştir. Bu performans değeri, Çöltekin (2020) çalışmasında aynı test setini kullanarak elde edilen 0.77 f1 skorunu ve Zampieri ve diğerleri (2020) çalışmasında önerilen birçok modelin performansını aşmaktadır. Oluşturulan Türkçe hakaret ve nefret söylemi otomatik tespit modeli, bir web servisi ve web arayüzü aracılığıyla son kullanıcıya sunulmuştur. Bu araştırma hem oluşturulan ve kullanılan veri setleri ile hem de makine öğrenme modelleri üzerine gerçekleştirdiği deneyler ile Türkçe hakaret ve nefret söylemi alanında gerçekleştirilen en kapsamlı araştırmadır. Gelecekte, bu çalışmanın bu alanda yapılacak araştırmalara öncülük etmesi beklenmektedir.
- Bu çalışmada, milyonlarca hakaret ve nefret söylemi metin örneğiyle eğitilen makine öğrenme modellerinin performansını değerlendirme ve karşılaştırma aşamalarında

yalnızca 3528 metin örneğinden oluşan OffensEval 2020 test setinin kullanılmasının önemli bir eksiklik olduğu vurgulanmıştır. Bu nedenle, bu çalışmada oluşturulan Türkçe hakaret ve nefret söylemi otomatik tespit modelinin performansını daha sağlıklı bir şekilde değerlendirebilmek için gelecek çalışmalarda Türkçe elle etiketlenmiş kapsamlı kapsamlı bir test setinin oluşturulması hedeflenmektedir.

- Çevrimiçi ortamda kullanıcılar tarafından oluşturulan istenmeyen söylemleri (hakaret, küfür, cinsiyetçi, ırkçı, alay etme, trolleme, siber zorbalık, incitici ve küçük düşürücü söylemler) ayrıntılı bir şekilde tespit edebilmek için literatürde büyük kabul gören Zampieri ve diğerleri (2019) tarafından önerilen etiketleme hiyerarşisine uygun olarak elle etiketlenmiş bir Türkçe veri seti oluşturmanın büyük önemi vardır. Böyle bir veri setinin oluşturulması, bu alandaki çalışmaların ilerlemesine ve karşılaştırılmasına büyük bir katkı sağlayacağı düşünülmektedir. Bu nedenle, söz konusu veri setinin oluşturulması, gelecekteki çalışmalarda yapılacaklar listesine eklenmiş ve bu alanda çalışacak olan araştırmacılara önerilmiştir.

Kaynaklar

[1] Sap, M., Card, D., Gabriel, S., Choi, Y., A, N., 2019. The risk of racial bias in hate speech detection. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics , 1668–1678.

[2] Mathew, B., Dutt, R., Goyal, P., Mukherjee, A., 2019. Spread of hate speech in online social media. In Proceedings of WebSci. ACM.

[3] Das, M., Mathew, B., Saha, P., Goyal, P., Mukherjee, A., 2020. Hate speech in online social media. ACM SIGWEB Newsletter, (Autumn) , 1–8.

[4] Rizwan, H., Shakeel, M.H., Karim, A., 2020. Hate-speech and offensive language detection in roman urdu. In Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP) , 2512–2522.

[5] https://www.theguardian.com/world/2018/mar/14/facebook-accused-by-sri-lanka-of-failing-to-control-hate-speech. (15.07.2022)

[6] https://www.reuters.com/investigates/special-report/myanmar-facebook-hate. (15.07.2022)

[7] https://money.cnn.com/2017/06/01/technology/twitter-facebook-hate-speech-europe/index.html. (15.07.2022)

[8] https://help.twitter.com/tr/rules-and-policies/hateful-conduct-policy. (17.07.2022)

[9] https://www.theverge.com/2019/12/16/21021005/google-youtube-moderators-ptsd-accenture-violent-disturbing-content-interviews-video. (17.07.2022)

[10] Wiedemann, G., Ruppert, E., Jindal, R., Biemann, C., 2018. Transfer learning from lda to bilstm-cnn for offensive language detection in twitter. Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018) , 85–94.

[11] Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., Chang, Y., 2016. Abusive language detection in online user content. In Proceedings of the 25th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee , 145–153.

[12] Davidson, T., Warmsley, D., Macy, M., Weber, I., 2017. Automated hate speech detection and the problem of offensive language. The 11th International AAAI Conference on Web and Social Media , 6–7.

[13] Waseem, Z., Hovy, D., 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies , 88–93.

[14] Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., Kumar, R., 2019a. Predicting the type and target of offensive posts in social media. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies .

[15] Çağrı Çöltekin, 2020. A corpus of turkish offensive language on social media. Proceedings of the 12th Language Resources and Evaluation Conference, 6174–6184.

[16] Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., Çağrı Çöltekin, 2020. Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). In Proceedings of SemEval .

[17] Şahi, H., Kılıç, Y., Sağlam, R.B., 2018. Automated detection of hate speech towards woman on twitter. 3rd International Conference on Computer Science and Engineering , 533–536.

[18] Mayda, I., Diri, B., Dalyan, T., 2021. Türkçe tweetler üzerinde makine öğrenmesi ile nefret söylemi tespiti. Avrupa Bilim ve Teknoloji Dergisi , 328–334.

[19] Sezer, T., 2020. Twitter derlemi – ts tweets corpus. URL: https://tanersezer.com/?p=155. (14.09.2020)

[20] Kemik, 2020. Kemik doğal dil İşleme grubu. URL: http://www.kemik.yildiz.edu.tr/verikumelerimiz.html. (10.09.2020)

[21] https://github.com/drsalihkurt/HateSpeechandOffensiveLanguageDetectionInTurkish. (15.05.2023)

[22] Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional

transformers for language understanding. arXiv preprint arXiv:1810.04805.

[23] Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L., 2018. Deep contextualized word representations. arxiv preprint arxiv:1802.05365. Structure.

[24] Howard, J., Ruder, S., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics5 , 328–339.

[25] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., 2018. Improving language understanding by generative pre-training. Technical Report. OpenAI.

[26] 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.

[27] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692

[28] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzman, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V., 2020. Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL) .