Mühendislik Bilimleri ve Tasarım Dergisi
10(2), 504 – 519, 2022
e-ISSN: 1308-6693

Journal of Engineering Sciences and Design
DOI: 10.21923/jesd.973181

Araştırma Makalesi Research Article

TÜRKÇE DİLİNDE YAZILAN BİLİMSEL METİNLERİN DERİN ÖĞRENME TEKNİĞİ UYGULANARAK ÇOKLU SINIFLANDIRILMASI

Mustafa ÖZKAN*, Görkem KAR

Bahçeşehir Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, İstanbul, Türkiye

Anahtar Kelimeler

Makine Öğrenmesi, Doğal Dil İşleme, Derin Öğrenme, Çok Sınıflı Sınıflandırma, BERT.

Öz

Ekim 2018 yılında Google tarafından geliştirilen BERT derin öğrenme tekniği, makine öğrenimi ve doğal dil işleme dünyasında çok popüler oldu. Transformatörlerin Çift Yönlü Kodlayıcı Gösterimleri anlamına gelen BERT, yapay zeka ve makine öğrenimi teknolojilerini bir arada kullanan bir doğal dil işleme tekniği olarak açıklanabilir. Günümüzde, gözetimli öğrenme metodolojisinin bir parçası olan sınıflandırma problemleriyle çokça karşılaşılmaktadır. Sınıflandırmanın temeli eğitilen bir makinenin yeni gelen bir veri hakkında tahminleme yapabilmesine ve sınıflandırabilmesine dayanır. Buradaki amaç bir veri kümesi üzerinde tanımlı olan sınıflar arasında veriyi dağıtabilmektir. Türkçe'nin morfolojisinin zengin ama karmaşık olması, sondan eklemeli bir dil olması ve dil bilgisinden kaynaklanan zorluklar çoklu sınıflandırma problemlerinin çözümünde başlıca sorun teşkil etmekte iken BERT derin öğrenme tekniği ile bu sorun daha kolay çözülebilir hale gelmiştir. Bu çalışmada, son 10 yıl içinde Türkçe dili ile yazılmış akademik araştırma ve bilimsel çalışmalar veri seti olarak kullanıldı. Çoklu sınıflandırma problemlerinde kullanmak üzere, veri setine BERT derin öğrenme tekniği uygulanarak önceden eğitilmiş Türkçe bir BERT modeli üzerinde ince ayar (fine-tuning) yapıldı. Deneylerin sonucunda, eğitilmiş olan sistemin doğruluğu %96 başarım oranına sahip olmuştur.

MULTICLASS CLASSIFICATION OF SCIENTIFIC TEXTS WRITTEN IN TURKISH BY APPLYING DEEP LEARNING TECHNIQUE

Keywords

Machine Learning, Natural Language Processing, Deep Learning, Multiclass Classification, BERT.

Abstract

The BERT deep learning technique, which is developed by Google in October 2018, has become very popular in the world of machine learning and natural language processing. BERT, which stands for Bidirectional Encoder Representations of Transformers, can be explained as a natural language processing technique that uses artificial intelligence and machine learning technologies together. Nowadays, classification problems that are part of the supervised learning methodology are frequently encountered. Classification is based on the ability of a trained machine to predict and classify new data. The purpose is to distribute data between classes defined on a dataset. In Turkish many of the difficulties arise from being an agglutinative language and having a rich but complex morphology. These difficulties cause hard to solving multiclass classification problems. However, it has become more easily solvable with using BERT deep learning technique. We used academic research and scientific studies written in Turkish in the last 10 years as our dataset. We fine-tuned our dataset on a pre-trained Turkish BERT model by applying BERT deep learning technique to use in multiclass classification problems. As a result of experiments, it is seen that the accuracy of the system we have trained has achieved 96% accuracy.

Alıntı / Cite

Özkan, M., Kar, G., (2022). Türkçe Dilinde Yazılan Bilimsel Metinlerin Derin Öğrenme Tekniği Uygulanarak Çoklu Sınıflandırılması, Mühendislik Bilimleri ve Tasarım Dergisi, 10(2), 504-519.

* İlgili yazar / Corresponding author: m.ozkan0093@gmail.com, +90-538-323-6484

Yazar Kimliği / Author ID (ORCID Number)	Makale Süreci / Article Process	
M. Özkan, 0000-0003-4287-9220	**Başvuru Tarihi / Submission Date**	19.07.2021
G. Kar, 0000-0003-0367-4409	**Revizyon Tarihi / Revision Date**	23.12.2021
	Kabul Tarihi / Accepted Date	28.12.2021
	Yayım Tarihi / Published Date	30.06.2022

1. Giriş (Introduction)

Yapay zekanın en dikkat çekici teknolojilerden biri olmasının sebebi bir insana gerek kalmadan veya minimum insan müdahalesiyle işlem yapabilme yeteneğine sahip olmasıdır. Yapay zeka konusuna genelden özele doğru bakıldığı zaman ortaya Makine Öğrenmesi ve Derin Öğrenme kavramları çıkacaktır. Yapay zeka, görevleri yerine getirmek için insan zekasını taklit eden ve topladıkları bilgilere göre kendini yinelemeli olarak iyileştirebilen sistemleri veya makineleri ifade eder. Makine öğrenimi, yapay zekanın bir alt bilim dalıdır. Bilgisayarların algılayıcı verisi ya da veri tabanları gibi veri türlerine dayalı öğrenimini olanaklı kılan algoritmaların tasarım ve geliştirme süreçlerini konu edinen bir bilim dalıdır. Derin öğrenme, bir makine öğrenme yöntemidir ve verilen bir veri kümesi ile çıktıları tahmin edecek yapay zekayı eğitmemize olanak sağlar.

Derin öğrenme yöntemlerinin daha ayrıntılı olarak verilebilecek uygulama alanları arasında aşağıdakiler yer almaktadır (Deng ve Yu, 2014):

- Dil modelleme ve doğal dil işleme (Language modeling and natural language processing)
- Konuşma ve ses işleme (Speech and audio processing)
- Bilgi erişimi (Information retrieval)
- Nesne tanıma ve bilgisayarlı görü (Object recognition and computer vision)
- Çok modelli ve çok görevli öğrenme (Multimodal and multitask learning)

Yukarıda paylaşılan bilgiler ışığında ilk sırada yerini alan doğal dil işleme konusu, metin madenciliği olarak da anılmakta ve yapay zeka alanı içerisinde değerlendirilmektedir. Doğal dil işleme, birçok konuyu bünyesinde barındırmaktadır. Uygulama alanı gittikçe artmaya devam eden doğal dil işleme konusuna başlıca örnek olarak aşağıdakiler verilebilir:

- Metin ayrıştırma
- Metin sınıflandırma
- Bilgi çıkarımı
- Duygu analizi

Yukarıda verilen örneklerden ikinci sıradaki yerini alan “Metin Sınıflandırma” konusu Türkçe dili için belli başlı sorunları da beraberinde ortaya çıkarmaktadır. Türkçe'nin morfolojisinin zengin ama karmaşık olması, sondan eklemeli bir dil olması ve dil bilgisinden kaynaklanan zorluklar metin sınıflandırma yöntemlerinden biri olan çoklu sınıflandırma problemlerinin çözümünde başlıca sorun teşkil etmekte iken BERT (Bidirectional Encoder Representations from Transformers) derin öğrenme tekniği ile bu sorun daha kolay çözülebilir hale gelmiştir.

Bu çalışmada; doğal dil işleme problemlerinin çözümü için en güncel derin öğrenme tekniklerinden biri olan BERT teknolojisi, son 10 yıla ait toplanan Türkçe dili ile yazılmış tez ve patentlerden oluşan bir veri seti ile kullanılarak incelenmiş olup, Türkçe bir kaynak olarak araştırmacıların dikkatine sunulmuştur. Bu çalışmanın göze çarpan aşamaları aşağıda özetlenmiştir.

- 2011- 2021 yılları arasında oluşturulmuş 1560 adet tez ve patent veri seti olarak kullanılmıştır.
- Veri seti bir ön işleme aşamasından geçirilmiş ve etiketleme yapılmıştır.
- Eğitim ve doğrulama kümeleri belirlenmiştir.
- Tokenizer ve model belirlenmiş, en uygun optimizasyon algoritması kullanılmıştır.
- Eğitim gerçekleştirilmiş ve BERT teknolojisi ile Türkçe dilinde çok sınıflı sınıflandırma yapılmıştır.

Çalışmanın kalan bölümleri şu şekilde ilerlemektedir: 2. bölümde materyal ve kullanılan metotlar incelenmiş, 3. bölümde deneysel sonuçlar paylaşılmış, 4. Bölümde konu hakkında tartışma yapılmış, 5. bölümde ise makale özetlenmiş ve sonuçlar verilmiştir.

Not: Bu makaledeki çalışma, Python 3.7.10 versiyonu (What’s New In Python 3.7, 2018) ile çalışma zamanı GPU olarak ayarlanmış Google Colab (Bisong, 2019) ortamı kullanılarak test edilmiştir. Kullanılan fiziksel aygıt Tesla T4 (Jia vd., 2019) modelidir.

BERT derin öğrenme tekniğinin oldukça yeni bir teknoloji olmasından dolayı, Türkçe dili üzerindeki bilimsel çalışmalar oldukça kısıtlıdır. Literatürde BERT derin öğrenme tekniği ile Türkçe dili için gerçekleştirilen farklı çalışmalar incelenmiş ve bu çalışmalara dair ayrıntılar kronolojik olarak aşağıda paylaşılmıştır.

Temmuz 2019 tarihinde gerçekleştirilen bu çalışmada transfer öğrenme (transfer learning) tekniği kullanılarak restoran ve ürün değerlendirmeleri hakkında pozitif ve negatif olarak duygu analizi üzerine bir çalışma gerçekleştirilmiştir. Bu çalışma aynı zamanda ince ayar tekniği ve derin öğrenme mimarisi uygulanarak transfer öğrenmeyi Türk Dili için duygu analizi problemini çözmek için kullanan önemli bir girişim olarak kabul edilmektedir. Çalışma sonucunda transfer öğrenmesi ile gerçekleştirilen eğitim sonucunda elde edilen modelin F1 skoru restoran ve ürün değerlendirmeleri için sırasıyla 0.913 ve 0.842 olarak belirtilmiştir. (Akin ve Yildiz, 2019)

Ekim 2020 tarihinde Türkçe duygu analizi üzerine gerçekleştirilen bu bilimsel çalışmada, Türkçe duygu analizi için BERT bazlı 3 farklı model geliştirilmiştir. Geliştirilen 3 model de ikili sınıflandırma probleminin çözümünde kullanılmıştır. Film ve otel yorumlarından oluşan, pozitif ve negatif olmak üzere 2 etikete ve aynı eğitim ve doğrulama kümesine sahip Türkçe veri kümeleri üzerinde yapılan deneylerde en yüksek başarım oranı 0.9332 olarak elde edilmiştir. (Acikalin vd., 2020)

Ocak 2021 tarihinde gerçekleştirilen bu çalışmada Türkçe metinlerdeki duyguların sınıflandırılması amaçlanmıştır. Bu çalışmada, Türkçe dili ile daha güçlü bir duygu sınıflandırma modeli oluşturmak için önceden eğitilmiş dil modeli yaklaşımı kullanılmıştır. İyi bilinen önceden eğitilmiş dil modellerine bu amaç için ince ayar yapılmıştır. Türkçe duygu sınıflandırması için bu ince ayarlı modellerin performansları, deneysel çalışmalarda geleneksel makine öğrenmesi ve derin öğrenme yöntemlerinin performansları ile kapsamlı bir şekilde karşılaştırılmıştır. Bu çalışma sonucunda önerilen yaklaşımın, Türkçe duygu sınıflandırması için en gelişmiş performansı sağladığı ortaya konulmuştur. (Uçan vd., 2021)

Ocak 2021 tarihinde gerçekleştirilen bu çalışma da bir önceki çalışma gibi ikili sınıflandırma problemini temel almaktadır. Gerçekleştirilen çalışmada, içerisinde deprem, sel, kaza, olumsuz hava olayları gibi felaket durumları hakkında paylaşılan 7613 adet tweet ile sınıflandırma çalışması gerçekleştirilmiştir. Veri seti içerisinde yer alan tweet'lerin bazıları gerçek bir felaket durumunu ifade ederken bazıları bağımsız kelimeler açısından bir felaket olarak nitelenebilecekken bütün itibariyle felaket olmayan bir durumu ifade etmektedir. Felaketler konusunda ve gerçek dışı olarak işaretlenmiş bu veriler Google yapay zeka ekibi tarafından geliştirilen, sinir ağı temelli bir model olan BERT ile sınıflandırılmıştır. %80 eğitim, %20 doğrulama seti olarak ayrılan veri kümesi üzerinde tur sayısı 10 olarak gerçekleştirilen eğitim işlemi sonucunda 0.9888 elde edilmiştir. Doğruluk değeri 0.8794, duyarlılık değeri ise 0.8503 olarak elde edilmiştir. (Sevli ve Kemaloğlu, 2021)

Mayıs 2021 tarihinde bir diğer bilimsel çalışmada ise deneysel bir vaka çalışması gerçekleştirilmiştir. Morfolojik olarak zengin bir dil olan Türkçe için BERT'in etkinliği gösterilmiştir. Geleneksel olarak morfolojik yapısı zor olan diller, verileri makine öğrenimi algoritmalarına uygun olacak şekilde modellemek için yoğun bir dil ön işleme adımlarından geçer. Özellikle, veri seyrekliği veya yüksek boyutlu problemlerin üstesinden gelmek için verimli bir veri modeli elde etmeye ve sözcüklere ayırma, köklere ayırma gibi görevlere ihtiyaç vardır. Bu bağlamda, literatürden duygu analizi, siber zorbalık tespiti, metin sınıflandırma, duygu tanıma ve spam tespiti olmak üzere beş farklı Türkçe NLP araştırma problemi seçilmiş, daha sonra BERT'in deneysel performansı temel makine öğrenimi algoritmalarıyla karşılaştırılmıştır. Son olarak, ağır ön işleme görevleri ortadan kaldırılırken, seçilen NLP problemlerinde temel makine öğrenimi algoritmalarına kıyasla gelişmiş sonuçlar bulunmuştur. (Özçift vd., 2021)

Mayıs 2021 tarihinde gerçekleştirilen bir diğer bilimsel çalışma ise Türk Facebook kullanıcıları için cinsiyet tahminine yönelik keşifsel bir çalışmadır. İkili sınıflandırma problemine çözüm bulan bu çalışmada mevcut çalışmalardan farklı olarak Facebook kullanıcılarının cinsiyet tespiti sadece metin içerikleri kullanılarak değil, aynı zamanda profil bilgileri, ağ yapısı, duvar etkileşimleri ve duvar içerikleri kullanılarak da gerçekleştirilmiştir. BERT teknolojisi ile geliştirilen modelde, Türk sosyal medya kullanıcılarının duvar içerikleri kullanılarak elde edilen cinsiyet tahmininin doğruluk değeri 0.926 olarak elde edilmiştir. Duvar içerikleri ile eğitilen BERT modelinin, çalışma kapsamında gerçekleştirilen diğer tüm makine öğrenmesi sınıflandırıcılarından ve derin öğrenme algoritmalarından daha iyi performans gösterdiği ortaya konulmuştur. Ayrıca BERT modelinin çok daha büyük bir korpus üzerinde eğitildiğinde çok daha iyi sonuçlar verebileceği sonucuna varılmıştır. (Çoban vd., 2021)

Haziran 2021 tarihinde gerçekleştirilen bir diğer bilimsel çalışma ile farklı konulardaki Türkçe etiketli metinler sınıflandırılarak başarıları incelenmiştir. Klasik makine öğrenmesi yöntemleri, derin öğrenme mimarileri ve son zamanlarda popüler hale gelen transformer tabanlı sınıflandırıcı modellerinin başarıları dört farklı veri kümesi üzerinde karşılaştırılmıştır. Bu çalışmada tüm veri kümeleri üzerinde BERT tabanlı sınıflandırıcı modelinin hem klasik hem de derin öğrenme tabanlı sınıflandırıcılardan daha yüksek sınıflandırma başarısı verdiği gözlenmiştir. (Şahin ve Diri, 2021)

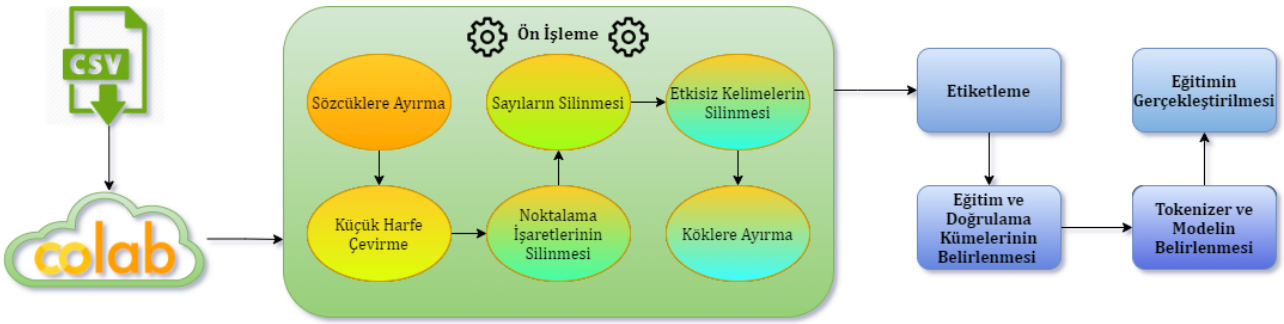
Ortaya koyduğumuz bu çalışma ile metin sınıflandırma yöntemlerinden biri olan "Çok Sınıflı Sınıflandırma" probleminin Türkçe dili için BERT derin öğrenme tekniği kullanılarak çözülmesi hedeflenmiş ve gerçekleştirilmiştir. Birçok eğitimden geçmiş farklı veri setleri ve doğrulama kümesi ile kapsamlı bir başarım tablosu (Tablo 10) elde edilmiştir. Elde edilen bu başarım tablosu ile gelecekteki "Çok Sınıflı Sınıflandırma" problemlerinin Türkçe dili için BERT derin öğrenme tekniği uygulanarak gerçekleştirilecek çalışmalara ışık tutulması ve daha yüksek başarım oranlarına ulaşılması temenni edilmektedir.

2. Materyal ve Yöntem (Material and Method)

Günümüzde lisans eğitimini tamamlayan çoğu öğrenci, daha iyi bir gelecek için ikinci bir yeterlilik almaya karar verir ve lisans üstü eğitimine başlar. Yüksek lisans, lisans öğretimine dayalı eğitim-öğretim ve araştırmanın sonuçlarını ortaya koymayı amaçlayan bir yükseköğretimdir. Yüksek lisans eğitimi, kişinin eğitim aldığı branşta uzmanlaşması ve bu durumu bilimsel bir tezle ispat etmesiyle tamamlanır. Her yıl sayısız mezun veren yüksek lisans programları ile sayısız bilimsel tez akademimize kazandırılmaktadır.

Bir diğer önemli konu ise fikri ve sınai mülkiyet haklarından biri olan patenttir. Bir buluşa sahip olan kişinin, buluşa konu olan ürünü, belirlenen süre boyunca üretme, kullanma, satma ya da ithal etme hakkına sahip olması durumuna patent hakkı adı verilir. Bu hakkın somut göstergesi olan belge ise patent adını taşır. Günümüzde bilgiye dayalı faaliyetlerin artması ile birlikte patentin önemi daha da artmıştır.

Tez ve patentlerin ortak noktaları, ortaya konulan eserin bir kategoriye ait olmasıdır. Hem tez yazarı hem de patent yazarı, çalışmalarının içerdiği kategoriyi kendileri yazarak belirtmektedir. İnternet üzerinden erişime açık olan bu iki alanda da kategorisi belirlenmemiş birçok eser mevcuttur. Günümüz teknolojisinde yapay zeka teknolojisi ile metinlerin kategorilerinin belirlenmesi sağlanabilmektedir. Bu çalışma buradan yola çıkarak özellikle Türkçe dilinde yazılan bilimsel metinlerin BERT derin öğrenme tekniği uygulanarak çoklu sınıflandırılmasını konu edinmiştir. İlk aşama olarak bir veri seti oluşturulmuştur. Oluşturulan veri seti üzerinde ön işleme (Pre-processing) gerçekleştirilmiştir. Ön işlemeden geçirilen veri seti için eğitim ve doğrulama setleri belirlenmiştir. Sonrasında BERT derin öğrenme tekniği uygulanarak bir performans metriği olan F1 skoru elde edilmiştir. Çalışmaya ait aşamalar ve sistemin genel görünümünü içeren diyagram Şekil 1 üzerinde görülebilir. Devamındaki bölümlerde ise çalışmanın ayrıntıları detaylı bir şekilde sunulmuştur.



Şekil 1. Sistem Genel Görünümü (System Overview)

2.1. Metin Sınıflandırma (Text Classification)

Metin sınıflandırma sorunu, $B = \{b_1, b_2, \dots, b_n\}$ kümesindeki her bir belgenin (metin), önceden tanımlanmış $S = \{s_1, s_2, \dots, s_m\}$ kümesindeki sınıflara ait olup olmadığının belirlenmesidir. Yani her $(b_j, s_i) \in B \; x \; S$ çifti için doğru ya da yanlış biçiminde bir mantıksal değer üretilmesi gerekmektedir. (Tantuğ, 2016) Metin sınıflandırma, verilerden kolayca iç görü elde etmeye olanak sağladığı için yapay zeka dünyasının gün geçtikçe daha da önemli bir parçası haline gelmiştir. Veriler belirli bir şekilde düzenlenmedikçe bu verilerden değer elde etmek zordur. Metin etiketleme veya metin kategorizasyonu olarak da bilinen metin sınıflandırması, veri olarak ele alınan metni organize gruplar halinde kategorize etme işlemidir. Bir Doğal Dil İşleme alt konusu olan metin sınıflandırma ile metin analiz edilir ve ardından içeriğine göre önceden tanımlanmış bir dizi etikete veya kategoriye atanabilir. Bu atamanın sonucunda metin kategorilere ayrılmış olur ve düzensiz veri düzenlenerek kullanılmak üzere bir değer elde edilir.

Metin sınıflandırması için kullanılan başlıca yöntemler bir sonraki konu başlığında ayrıntılı bir şekilde ele alınmıştır.

2.1.1. Sınıflandırma Yöntemleri (Classification Methods)

Genel olarak iki tür sınıflandırma problemi vardır: ikili problem ve çok sınıflı problem. İkili problem, bir tahmin sonucunun evet veya hayır kararıyla belirlenmesi gereken bir durum iken, çoklu sınıflandırma problemi, tahmin edilen bir sonucun birden çok sonuç olarak belirlendiği bir durumdur. (Kraipeerapun, 2009) Sınıflandırma problemlerinin çözümü için İkili Sınıflandırma, Çok Sınıflı Sınıflandırma ve Çok Etiketli Sınıflandırma yöntemleri kullanılır. Ayrıntılar aşağıdaki alt başlıklarda listelenmektedir.

2.1.1.1. İkili Sınıflandırma (Binary Classification)

İkili sınıflandırma, adından da anlaşılacağı üzere verileri olası iki kategoriden birine sınıflandırma problemlerinin çözümünde kullanılır. Bu sınıflandırma yönteminde, sorulacak soruların sadece iki cevabı olmalıdır. Bu tür sınıflandırmalar için ünlü filozof Aristoteles'in mantığı geçerlidir. Bu mantığın temelinde, sorulara karşılık gelen iki ana cevaba dahil olup olmama durumu vardır.

Bu sınıflandırmaya örnek olarak: E-posta spam tespiti, belirli tıbbi durum tespiti, duygu analizi (yalnızca pozitif ve negatif kategori kabul edilmiştir) düşünülebilir.

2.1.1.2. Çok Sınıflı Sınıflandırma (Multi-class Classification)

Çok sınıflı sınıflandırmada, ikili sınıflandırmadan farklı olarak sınıf sayısı ikiden fazladır. Bu sınıflandırma problemlerinin çözümünde, verilecek cevap ikiden fazla değer içinden seçilerek cevaplanacaksa buna çok sınıflı sınıflandırma (Multi-class classification) problemi denir. Buradaki önemli noktalardan biri, verilen cevap maksimum bir sınıfa dahil olabilir.

Örneğin; elimizde bir e-ticaret sitesinde satılan ürüne yapılmış olan yorumlar olsun. Bu yorumların da 3 kategorisi olsun: Fiyat, kullanım ve kalite. Bu yorumlar belli bir sınıf altında toplanmak istendiği zaman çok sınıflı sınıflandırma yöntemi kullanılmalıdır.

2.1.1.3 Çok Etiketli Sınıflandırma (Multi-label Classification)

Çoklu etiketli sınıflandırma problemlerinin çözümünde de bir önceki yöntemde anlatılan çok sınıflı sınıflandırma gibi verilecek cevap, ikiden fazla değer içinden seçilerek verilecektir. Bu sınıflandırma yönteminin çok sınıflı sınıflandırma yönteminden farkı verilen cevabın birden fazla sınıfa dahil olabilmesidir. Çoklu etiketli sınıflandırmanın amacı, tek bir örnek için bir dizi ilgili etiket ataması yapabilmektir.

Örneğin; bir önceki yöntemde bahsettiğimiz, satın alınmış olan yorumların kategorileri şu şekildedir: Fiyat, kullanım ve kalite.

Bu yorumların içinde birden fazla sınıfa dahil olan yorumlar olabilir. Sadece kullanımdan bahsedilebileceği gibi, fiyat ve kaliteden de aynı anda bahsedilen yorumlar içerebilir ve bu yorumları belirli bir sınıfa atamak istenmeyebilir. Bu tarz problemlerde doğru çıktı, çoklu etiketli sınıflandırma yöntemi kullanılarak elde edilmelidir. Bu çalışma boyunca yukarıda ayrıntıları paylaşılan sınıflandırma yöntemlerinden Çok Sınıflı Sınıflandırma kullanılmıştır. Bu sınıflandırma yönteminin kullanılmasının sebebi, ileride ayrıntıları verilecek olan veri setinin sınıflandırılmasının ikiden fazla değer içinden seçilerek maksimum bir sınıfa dahil olmasından kaynaklanmaktadır. Kısaca, kullanılacak olan veri seti beş farklı sınıftan oluşacak ve maksimum bir sınıfa dahil olacaktır.

Veri seti iki farklı kategoriden daha fazla kategoriye sahip olduğu için İkili Sınıflandırma yöntemi kullanılmamıştır. Veri seti içindeki herhangi bir verinin birden fazla kategoriye aynı anda dahil olma ihtimali olmadığı için de Çoklu Etiketli Sınıflandırma yöntemi kullanılmamıştır.

2.2. BERT (Bidirectional Encoder Representations from Transformers)

"Bidirectional Encoder Representations from Transformers" ifadelerinin baş harflerinden oluşan BERT algoritması, Ekim 2018'de Google tarafından geliştirilen doğal dil işleme (NLP) ön eğitimi için Transformer (derin öğrenme modeli) tabanlı bir makine öğrenimi tekniğidir. (Devlin vd., 2018)

Google, 2015 yılında Rankbrain (Schachinger, 2017) algoritmasını makine öğrenmesi ile destekleyerek arama sonuçlarında insan mantığına en yakın ve doğru cevapları filtrelemeyi sağlamıştır. 2019 yılı ile beraber BERT güncellemesini yayınlayarak, sorgu kelimelerini ayrı ayrı işlemek yerine tüm cümleyi incelemeye başlamış ve tüm

kelimeleri mantıksal biçimde değerlendirerek en tutarlı sonuçlara ulaşmayı sağlamıştır.

BERT algoritması, geleneksel soldan sağa ve sağdan sola dil işleme modelleri yerine "İki Yönlü Dil İşleme" özelliğini kullanmaktadır. Soldan sağa ve sağdan sola giden yüzeysel çift yönlü dil işlemesinin aksine, BERT her bir kelimenin diğerine olan ilişkisini anlamaya çalışan daha karmaşık bir "Maskeli Dil Modeli" ve "Sıradaki Cümle Tahmini" öğrenme stratejilerini kullanır.

BERT tabanlı mimariler şu anda birçok NLP görevinde son teknoloji ürün olarak performans sağlar. (Devlin vd., 2018) Soru cevaplama, sıralı etiketleme, duyarlılık analizi ve çıkarım dahil olmak üzere farklı yapıya, ana hatlara ve karmaşıklığa sahip bir dizi problemlerin çözümünde kullanılır ve yukarıda bahsedilen iki temel öğrenme stratejisini temelinde barındırır. BERT temelini oluşturan bu iki öğrenme stratejisi konunun devamında ayrıntılı olarak ele alınmıştır.

2.2.1. Maskeli Dil Modelleme (Masked Language Modelling)

BERT, kelimeleri bir dizide rastgele maskeleyen ve çift yönlü temsilleri öğrenmek için kullanılabilen maskeli dil modelinin yardımıyla kelimeler arasındaki ilişkiyi anlamayı öğrenir. (Lee vd., 2020) Kelime dizileri BERT modeline aktarılmadan önce, belirli bir cümle için $x = (x_1, x_2, \dots, x_n)$, rastgele %15'lik kısmı özel bir sembol $[M]$ ile değiştirilir. Daha sonra model, dizideki diğer maskelenmemiş kelimelerin oluşturduğu %85'lik kısmın bağlama dayanarak maskelenen kelimelerin orijinal değerini tahmin etmeye çalışır. K'yı maskelenmiş konumlar kümesi, χ_{K}'yı maskelenmiş simgeler kümesi olarak, $\chi_{\setminus \mathrm{K}}$'yı maskelemeden sonraki cümle ve θ'yı model olarak tanımlayacak olursak, Maskeli dil modelleme Formül 1'in maksimuma çıkarılması ile modeli ön eğitimden geçirir (Song vd., 2020):

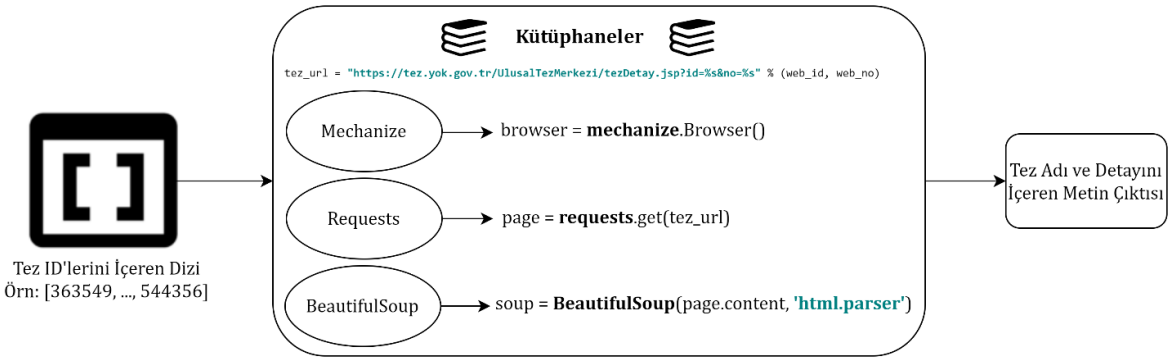
$$\log P\left(\chi_{\mathrm{K}} | \chi_{\setminus \mathrm{K}}; \theta\right) \approx \sum_{k \in \mathrm{K}} \log P\left(\chi_{\mathrm{K}} | \chi_{\setminus \mathrm{K}}; \theta\right) \qquad (1)$$

2.2.2 Sıradaki Cümle Tahmini (Next Sentence Prediction)

BERT eğitim sürecinde, modele cümle çiftlerini girdi olarak alır. Çiftteki ikinci cümlenin belge içerisinde sonraki cümle olup olmadığını tahmin etmeyi öğrenir. Modelin eğitimi sırasında girdilerin %50'lik kısmında ikinci cümlenin orijinal belge içerisinde takip eden cümle olduğu çiftler alınırken, diğer %50'lik kısımda ise ikinci cümle rastgele seçilir. Burada modelin rastgele seçilen cümlenin ilk cümleye bağlı olmadığını tespit etmesi beklenir. (Sevli ve Kemaloğlu, 2021)

2.3. Veri Seti Oluşturma (Dataset Creation)

Yazarları tarafından Yükseköğretim Kurulu Ulusal Tez Merkezi Veri Tabanında arşivlenmesine ve internet üzerinden tam metin erişime açılmasına izin verilen tezler açık erişime sunulmaktadır. Ulusal Tez Merkezi aracılığıyla Bilgisayar Mühendisliği Ana Bilim Dalı, Yazılım Mühendisliği Ana Bilim Dalı ve Elektrik-Elektronik Mühendisliği Ana Bilim Dalı altında son 10 yıl içerisinde yazılmış olan 1110 adet tez başlık ve özet kısımlarını içermek üzere Python dilinde yazılan script ile toplanmıştır. Bu script içinde "Mechanize", "Requests" ve "BeautifulSoup4" kütüphanelerinden faydalanılmıştır. Mechanize, John J. Lee tarafından geliştirilen, web siteleri üzerinde browser gibi davranarak işlemler yapılmasını sağlar. (Lee, 2013) Requests, web üzerindeki isteklerin yönetilmesini sağlar. (Chandra ve Varanasi, 2015) BeautifulSoup4, HTML ve XML dosyalarından veri çekmek için kullanılır. (Richardson, 2007) Jsp teknolojisi ile oluşturulmuş Ulusal Tez Merkezi'nin url'si (https://tez.yok.gov.tr/UlusalTezMerkezi/tezDetay.jsp?id=%s&no=%s) üzerinden gerekli tez id'si parametre olarak verilmiş, Requests kütüphanesinin get metodu kullanılarak bir döngü içinde istenen tüm tez verileri toplanmıştır. Kütüphanelerin kullanım şekli ve veri setinin oluşturulmasına ait aşamalar diyagram Şekil 2 üzerinde görülebilir.



Şekil 2. Veri Setinin Oluşturulması (Creating Dataset)

Bu çalışma, çoklu sınıf sınıflandırma konusunu temel aldığı için mümkün olduğunca birbirinden farklı kategori belirlenmeye çalışılmış ve aşağıdaki 5 kategori alanı belirlenmiştir:

1. Yapay Zeka ve Makine Öğrenimi
2. Nesnelerin İnterneti (IoT)
3. Görüntü İşleme
4. Artırılmış ve Sanal Gerçeklik
5. Haberleşme ve Sinyalleşme

1110 adet tez içeren ve herhangi bir kategoriye ait olduğunu belirtmeyen veri seti tek tek, titizlikle analiz edilerek yukarıdaki 5 kategorinin hangisine dahil olduğu belirlenmiştir. 732 adet tez yukarıda belirtilen 5 farklı kategorinin birine dahil olmuş fakat 378 adet tez bu kategorilerden birine dahil olmamıştır. Her kategori için en az 100 adet veri içermesini hedeflenmiştir. Az önce bahsedilmiş olan 378 adet tezin kendi aralarında oluşturabileceği ortak kategori sayısının en az 100 adet veri içermemesinden dolayı belirlenen 5 kategorinin dışında ayrı bir kategori eklenmemiştir.

732 adet verinin eğitim için yeterli olmayacağı düşünülerek veri setini arttırma yoluna gidilmiş ve Türkiye Cumhuriyeti Sanayi ve Teknoloji Bakanlığı'na bağlı, özel bütçeli, bağımsız yasal bir kuruluş olan Türk Patent ve Marka Kurumu'nun veri tabanından yararlanılmıştır. İlgili kurumun internet sayfası üzerinden detaylı arama yaparak 2011- 2021 yılları içerisinde yukarıda belirtilen 5 kategoriye ait olan toplam 828 adet patent veri setine dahil edilmiştir. Böylelikle veri seti 1560 adet tez ve patent çalışmalarının başlık ve özet kısımlarını içerecek şekilde son haline gelmiştir.

Bu çalışmanın devamında verilecek olan bütün istatistikler ve gerçekleştirilen çalışmalar, başlık ve özet kısımlarını beraber içeren veri seti dahilinde gerçekleştirilmiştir. Sadece başlık ve sadece özet bilgilerini içeren veri setleri ile ayrıca iki eğitim daha geçekleştirilmiştir. Bahsedilen iki eğitimin birbirinden farkı doğrulama büyüklüklerinin %20 ve %15 olarak belirlenmesidir. Bu iki eğitim sonucunda elde edilen değerlerden %20'lik doğrulama büyüklüğü içeren eğitimin sonucu direkt "Sonuç Değerlerinin Elde Edilmesi" bölümünde Tablo 9 üzerinde paylaşılacaktır. %15'lik doğrulama büyüklüğü içeren eğitimin sonucu ise "Tartışma" bölümünde Tablo 10 üzerinde paylaşılacaktır.

2.4. Veri İstatistikleri (Data Stats)

Veri ön işleme üzerinde çalışmaya başlanılmadan önce, mevcut verinin istatistikleri incelenmiştir. Veri seti içerisinde 1560 adet bilimsel metin bulunduğu daha önce belirtilmiştir. Eldeki verinin kategoriye ilişkin dağılımı Tablo 1'de paylaşılmıştır.

Tablo 1. Veri Kategori İlişkisi (Data Category Relationship)

KATEGORİ	METİN SAYISI
Yapay Zeka ve Makine Öğrenimi	598
Görüntü İşleme	438
Haberleşme ve Sinyalleşme	217
Nesnelerin İnterneti	183
Artırılmış ve Sanal Gerçeklik	124

Tablo 1'den anlaşılacağı üzere veri setinin dengesiz (imbalanced) durumda olduğu görülmektedir. BERT, ek veri artırma olmadan dengesiz sınıfları işleyebilir. (Madabushi vd., 2020) Bu sebeple veri setini dengelemek üzerine ayrı bir çalışma yapılmamıştır.

2.5. Veri Ön İşleme (Data Preprocessing)

Metni veri olarak kullanan her niceliksel çalışma, kelimelerin sayılara nasıl dönüştürüleceğine dair kararlar gerektirir. Toplu olarak 'ön işleme' olarak bilinen bu kararlar, analiz edilecek girdileri, sonraki modelin yorumlanabilirliğini veya sonuçlarını olumsuz bir şekilde etkilemeyecek şekilde daha az karmaşık hale getirmeyi amaçlamaktadır (Denny ve Spirling, 2018) .

Yapısal olmayan veri, önceden tanımlı bir veri modeline sahip olmayan ya da önceden tanımlı bir modele uyarlanamayan verileri ifade etmek için kullanılır. Genellikle metin ağırlıklı olmakla birlikte içerisinde tarih, sayı, koordinat verisi gibi farklı türlerde veriler de içerebilir. Bir önceki aşamada oluşturulan veri seti de yapısal olmayan veriler içermektedir. Yapısal olmayan verilerden anlam çıkartmak için verileri işlenebilir hale getirmek gerekmektedir. Bu işlemler aşağıda detaylı bir şekilde anlatılmıştır.

Aşağıda verilen örnek girdinin her ön işleme aşamasındaki çıktısı Tablo 2 üzerinde görülebilir.

Örnek girdi: BU BULUŞ YAYA GEÇİDİ İHLALİNİ TESPİT EDEN BİR SİSTEM (1) İLE İLGİLİDİR.

Tablo 2. Ön İşleme Çıktısı (Pre-Processing Output)

ÖN İŞLEME AŞAMASI	ÇIKTI
Sözcüklere Ayırma	'BU', 'BULUŞ', 'YAYA', 'GEÇİDİ', 'İHLALİNİ', 'TESPİT', 'EDEN', 'BİR', 'SİSTEM', '(', '1', ')', 'İLE', 'İLGİLİDİR', '. '
Küçük Harfe Çevirme	'bu', 'buluş', 'yaya', 'geçidi', 'ihlalini', 'tespit', 'eden', 'bir', 'sistem', '(', '1', ')', 'ile', 'ilgilidir', '. '
Noktalama İşaretlerinin Silinmesi	'bu', 'buluş', 'yaya', 'geçidi', 'ihlalini', 'tespit', 'eden', 'bir', 'sistem', '1', 'ile', 'ilgilidir'
Sayıların Silinmesi	'bu', 'buluş', 'yaya', 'geçidi', 'ihlalini', 'tespit', 'eden', 'bir', 'sistem', 'ile', 'ilgilidir'
Etkisiz Kelimelerin Silinmesi	'buluş', 'yaya', 'geçidi', 'ihlalini', 'tespit', 'sistem', 'ilgilidir'
Köklere Ayırma	'buluş', 'ya', 'geçidi', 'ihlalini', 'tespit', 'siste', 'ilgi'

2.5.1. Sözcüklere Ayırma (Tokenization)

Sözcüklere ayırma, bütün bir yazıyı oluşturan her bir sözcüğü ayırma işlemidir. Metin kelime kelime parçalanarak dizilere kaydedilir. Böylelikle, her bir kelime bir token (belirteç) olmuş olur.

NLTK (Natural Language Toolkit), insan dili verileriyle çalışacak Python programları oluşturmak için geliştirilmiş bir platformdur. (Tuzcu, 2020) Bu çalışmada NLTK kütüphanesinden "Punkt Word Tokenizer" kullanılarak bütün veri tokenlarına ayrılmıştır.

2.5.2. Küçük Harfe Çevirme (Lowercase Conversion)

Bu aşamada veri setinin içindeki metinlerin büyük küçük harf ayrımını ortadan kaldırmak amacıyla, tokenlarına ayrılmış olan her bir kelime küçük harfe çevrilmiştir.

2.5.3. Noktalama İşaretlerinin Silinmesi (Removing Punctuation)

Bu aşamada veri seti içerisindeki her bir metnin içerdiği noktalama işaretlerinin kaldırılması gerçekleştirilmiştir. String kütüphanesinden "punctuation" kullanılmıştır. Veri seti içerisinden silinen noktalama işaretleri şunlardır: {!, ", #, $, %, &, ', (,), *, +, ,, -, ., /, :, ;, <, =, >, ?, @, [, \,], ^, _, `, {, |, }, ~}

2.5.4. Sayıların Silinmesi (Removing Numbers)

Bu aşamada Regex kütüphanesinden faydalanarak, basit bir regex tanımı ile veri seti içindeki metinlerden sayılar kaldırılmıştır. Kullanılan regex tanımı: [0-9]+

2.5.5. Etkisiz Kelimelerin Silinmesi (Removing Stop Words)

Metin içerisinde geçen ve anlamda herhangi bir değişiklik yapmayan kelimelere etkisiz kelime denir. Bu kelimeler cümleden çıkarıldığında anlamda bir değişikliğe sebep olmadıkları için özellikle veri seti içerisinde bulunması gereksizdir. Bu sebeple, NLTK kütüphanesinden "stopwords" ile TRSTOP kütüphanesini birleştirerek, Türkçe dilindeki etkisiz 210 adet kelime veri setinden kaldırılmıştır. Veri setinin herhangi bir ön işleme aşamasından geçmeden önce içerdiği en çok kullanılan ilk 5 kelime Tablo 3'te gösterilmiştir.

Tablo 3. En Çok Kullanılan Kelimeler (Most Used Words)

KELİME	METİNLERDE GÖZLENME FREKANSI
ve	9624
bir	6637
ile	3701
için	3539
Bu	2903

Veri setinden kaldırılan 210 adet etkisiz kelime Şekil 3 üzerinde görülebilir:

Şekil 3. Etkisiz Kelimeler (Stop Words)

2.5.6. Kök Bulma (Stemming)

İngilizcede "Stemming" olarak isimlendirilen kök bulma işlemi, metinde geçen kelimelerdeki eklerin atılarak kelime köklerinin kaydedilmesidir. Yazının başında belirtildiği gibi Türkçe sondan eklemeli bir dildir, ancak İngilizcede bu durum söz konusu değildir. Dolayısıyla Türkçenin durumdan duruma farklılık gösterebilen ekleri İngilizcede görülmez. Kelimeleri köklere ayırma işlemi Türkçe dili için mükemmel olmaktan uzak olmasına rağmen metin verisi üzerinde çalışma gerçekleştiren yapay zeka çalışmalarında kullanımı avantaj sağlamaktadır. Bu aşamada Snowballstemmer modülündeki "TurkishStemmer" kütüphanesi kullanılarak veri setine Türkçe "stemming" uygulanmıştır. Kullanılan bu kütüphane ile "stemmer" bir kelime için tek aday döndürmektedir. Kullanım şekli aşağıda paylaşılmıştır:

```
1. from snowballstemmer import TurkishStemmer
2. turkStem=TurkishStemmer()
3. pp['turkish_stem'] = pp['words'].apply(lambda x: [turkStem.stemWord(word) for word in x])
```

2.6. Etiketleme (Labelling)

Daha önce belirlenmiş olan 5 kategori birbirinden farklı 5 adet sayıya etiketlenmiştir. Bu etiketler ileride yanlış yapılan tahminlerin hangi kategoriye ait olduğunun görülmesinde yardımcı olacaktır. Etiketlenen kategoriler ve değerleri Tablo 4 ile gösterilmiştir.

Tablo 4. Etiket Bilgileri (Label Informations)

KATEGORİ	ETİKET
Artırılmış ve Sanal Gerçeklik	1
Görüntü İşleme	2
Haberleşme ve Sinyalleşme	3
Nesnelerin İnterneti	4
Yapay Zeka ve Makine Öğrenimi	5

2.7. Eğitim ve Doğrulama Kümelerinin Belirlenmesi (Identifying Training and Validation Sets)

Veriyi eğitim ve doğrulama olarak ayırmak için gereken kütüphaneler yüklendikten sonra doğrulama büyüklüğü belirlenir. Doğrulama büyüklüğü önce %20 seçilerek, veri setinin %80'i eğitim için kullanılmıştır. Bir sonraki modelde ise doğrulama büyüklüğü %15 seçilerek, veri setinin %85'i eğitim için kullanılmıştır.

Veri ayrıldıktan sonra eğitim kümesinde mi yoksa doğrulama kümesinde mi olduğu görülebilir. Kategoriler, bu kategorilere atanan etiketler ve hangi kümede kaç veri olduğunun istatistikleri tablo olarak çıkarılabilir. Doğrulama büyüklüğü %20 olarak belirlenen veri setinin ayrıntıları Tablo 5 üzerinde paylaşılmıştır.

Tablo 5. Eğitim ve Doğrulama Kümesi (Training and Validation Set)

KATEGORİ	ETİKET	VERİ KÜMESİ	VERİ SAYISI
Artırılmış ve Sanal Gerçeklik	1	Eğitim	99
		Doğrulama	25
Görüntü İşleme	2	Eğitim	350
		Doğrulama	88
Haberleşme ve Sinyalleşme	3	Eğitim	174
		Doğrulama	43
Nesnelerin İnterneti	4	Eğitim	147
		Doğrulama	36
Yapay Zeka ve Makine Öğrenimi	5	Eğitim	478
		Doğrulama	120

Toplam Eğitim kümesi: 1248
Toplam Doğrulama kümesi: 312

Not: Doğrulama büyüklüğü %15 olan modelin F1 skoru direkt "Tartışma" bölümünde Tablo 10 üzerinde paylaşılacaktır. Bu çalışmanın içeriği doğrulama büyüklüğü %20 olarak belirlenmiş olan model üzerinden anlatılacak, çalışmanın devamında verilen bütün istatistik ve sonuçlar "Tartışma" bölümüne kadar bu modele ait olacaktır.

2.8. Tokenizer ve Model Belirleme (Tokenizer and Model Determination)

BERT derin öğrenme tekniğini kullanmak için iki ana objeye ihtiyaç vardır. Bunlardan biri Tokenizer, diğeri ise ince ayar yapılacak önceden eğitilmiş bir model. Tokenizer, önceden sahip olunan kelime haznesinin kullanılarak metnin ögelerine ayrılması işleminde kullanılan araç olarak tanımlanabilir. Bu çalışma da Türkçe dili üzerinde gerçekleştiği için, Tokenizer ve model olarak Türkçe doğal dil işleme topluluğunun oluşturduğu, Türkçe metinler kullanılarak ön eğitilmiş olan uncased BERTurk modeli kullanılmıştır. Bu model, toplamda 128 bin adet Türkçe kelime içerip 35 GB büyüklüğe sahiptir. 128 GB RAM ve 8 çekirdeğe sahip TPU v3-8 yongası ile eğitilmiştir. (BERTurk, 2020) Kullanılan modele https://huggingface.co/dbmdz/bert-base-turkish-128k-uncased adresi üzerinden ulaşılabilir.

2.9. Optimize Edici Belirleme (Optimizer Determination)

Gerçekleştirilecek olan derin öğrenme çalışmasında doğrusal olmayan probleme optimum çözüm aranmıştır. Burada optimumdan kasıt en uygun ve en verimli çözüm olarak düşünülebilir. Bu çalışmada aranan optimum değerin bulunması için AdamW optimizasyon yöntemi kullanılmıştır. Adam, derin sinir ağlarını eğitmek için özel olarak dizayn edilmiş, uyarlanabilir bir öğrenme hızı (learning rate) optimizasyon algoritmasıdır (Kingma ve Ba, 2014). Loshchilov & Hutter, Adam optimizasyon algoritmasını daha düzenli hale getirmek için AdamW'yi önerdiler ve ortaya koydukları çalışma ile AdamW optimizasyon algoritması ile eğitilen modellerin Adam optimizasyon algoritması ile eğitilen modellere göre daha az aşırı öğrenme(overfitting) problemi yaşadığını göstermişlerdir. (Loshchilov ve Hutter, 2017)

3. Deneysel Sonuçlar (Experimental Results)

3.1. Eğitimin Gerçekleştirilmesi (Training Phase)

BERT derin öğrenme algoritmasını geliştiren mühendisler tarafından, gerçekleştirilecek olan eğitimin aşağıda belirtilen hiper parametreler ile kullanılması önerilmiştir (Google Research Bert, 2018):

- Eğitim Tur Sayısı (Epoch): 4
- Yığın (Batch) Sayısı: 8, 16, 32, 64, 128
- Öğrenme Hızı (Learning Rate): 3e-4, 1e-4, 5e-5, 3e-5

Öneriler eşliğinde, bu çalışmada 5e-5 öğrenme hızı ile toplam 4 kez eğitim aşamasından geçmekte, girdiler 32'şer 32'şer alınarak modeli beslemektedir. Google Colab üzerinde GPU ile çalışıldığından dolayı her eğitim döngüsü ortalama 1 dakika 6 saniye süren zaman içerisinde gerçekleşmiş, eğitimin tamamlanması ise 4 dakika 24 saniye gibi kısa bir süre içerisinde gerçekleştirilmiştir. Her eğitim aşamasıyla beraber eğitim kaybı, doğrulama kaybı ve F1 skoru bilgileri kaydedilmiş ve her eğitim için oluşturulan model kaydedilmiştir. Eğitim aşamasından geçmiş her modelin ayrıntılı bilgileri bir sonraki bölümde paylaşılmıştır.

3.2. Sonuç Değerlerinin Elde Edilmesi (Obtaining Result Values)

F1 skoru değeri Kesinlik (Precision) ve Duyarlılık (Recall) değerlerinin harmonik ortalamasını göstermektedir. (Opitz ve Burst, 2019) Her bir modelin F1 skoru değerlerinin bulunabilmesi için Python dili ile yazılmış bir makine öğrenmesi kütüphanesi olan "Scikit-learn" modülü kullanılmıştır. Scikit-learn, orta ölçekli denetimli ve denetimsiz problemler için çok çeşitli son teknoloji makine öğrenimi algoritmalarını entegre eden bir Python modülüdür. (Pedregosa vd.,2011) Bu modüle ait olan "sklearn.metrics.f1_score" fonksiyonu, her tur sayısında gerçekleştirilen eğitim işlemi sonucunda kendi yazdığımız fonksiyon içerisinde çağrılmış ve her modelin F1 skoru elde edilmiştir. "sklearn.metrics.f1_score" fonksiyonu F1 skorunun formülünü temel almıştır. Fonksiyona ait ayrıntılı bilgiye https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html#sklearn.metrics.f1_score adresinden ulaşılabilir. Elde edilen F1 skorunun doğruluğu, "sklearn.metrics.precision_score" ve "sklearn.metrics.recall_score" fonksiyonları kullanılarak F1 skorunun formülü ile kontrol edilmiştir. F1 skorunun formülü Formül 2'de paylaşılmıştır.

$$F1\ Score\ = 2 * \frac{Precision * Recall}{Precision + Recall} \qquad (2)$$

Tablo 6 üzerinden de görüldüğü üzere eğitimin son aşaması olan 4. adımda oluşturulan 4 numaralı model en yüksek F1 skorunu elde etmiş ve 0.9551'lik bir başarıya ulaşmıştır.

Tablo 6. Eğitim Çıktıları (Training Outcomes)

MODEL	EĞİTİM KAYBI	DOĞRULAMA KAYBI	F1 SKORU	EĞİTİM SÜRESİ
1	1.1734	0.4738	0.8083	0:01:07
2	0.3121	0.2948	0.9127	0:02:13
3	0.1538	0.2477	0.9357	0:03:19
4	0.0969	0.1687	0.9551	0:04:24

Her bir modelin kategoriler özelindeki doğruluk, kesinlik, duyarlılık ve F1 skoru değerlerinin bulunması için "Scikit-learn" modülüne ait olan"sklearn.metrics.classification_report" fonksiyonu kullanılmıştır. 4 numaralı modelin performans metrik sonuçları Tablo 7 ile paylaşılmıştır:

Tablo 7. Model Performans Metrik Sonuçları (Model Performance Metric Results)

ETİKET	KATEGORİ	DOĞRULUK DEĞERİ	KESİNLİK DEĞERİ	DUYARLILIK DEĞERİ	F1 SKORU
1	Artırılmış ve Sanal Gerçeklik	0.96	1.00	0.96	0.98
2	Görüntü İşleme	0.94	1.00	0.94	0.97
3	Haberleşme ve Sinyalleşme	0.97	1.00	0.98	0.99
4	Nesnelerin İnterneti	0.97	1.00	0.97	0.99
5	Yapay Zeka ve Makine Öğrenimi	0.95	1.00	0.95	0.97

Eğitim öncesi ön işleme aşamasından sonra her bir kategori için gerçekleştirilen eğitim ve doğrulama kümeleri sayılarını içeren tablonun, 4. modelin elde ettiği doğruluk değerinin sonuçları ile güncellenmiş hali Tablo 8 üzerinde görülebilir:

Tablo 8. Model Sınıflandırma Sonuçları (Model Classification Results)

ETİKET	KATEGORİ	DOĞRULAMA SETİ	DOĞRU SINIFLANDIRILAN ÖRNEK SAYISI
1	Artırılmış ve Sanal Gerçeklik	25	24
2	Görüntü İşleme	88	83
3	Haberleşme ve Sinyalleşme	43	42
4	Nesnelerin İnterneti	36	35
5	Yapay Zeka ve Makine Öğrenimi	120	114

Eğitime başlamadan önce doğrulama büyüklüğü %20 olarak belirlenmiş ve böylelikle toplam doğrulama kümesi sayısının 312 olduğu gösterilmiştir. En yüksek F1 skoruna sahip 4. Modelde 312 adet doğrulama kümesi sayısının 14 tanesi yanlış kategoriye sınıflandırılmıştır.

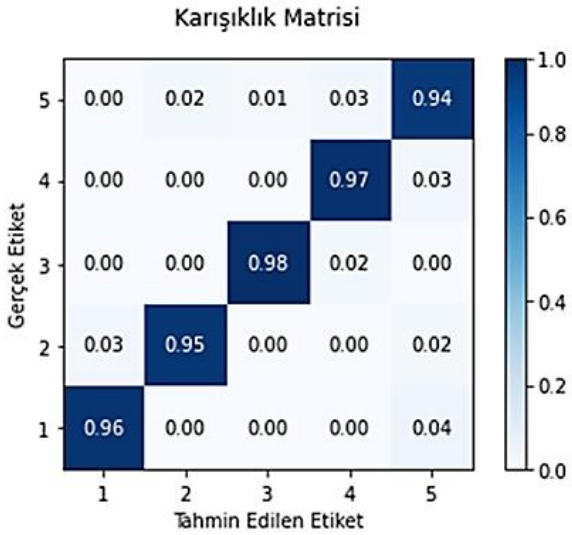
Sadece başlık ve sadece özet bilgilerini içeren veri seti ile doğrulama büyüklüğü %20 olarak gerçekleştirilen eğitimin sonuçları ön işleme aşamasından geçmiş ve ön işleme aşamasından geçmemiş şekilde ayrıntılı olarak elde edilmiştir.

Sonuç olarak elde edilen değerler ile Tablo 9'daki gibi bir başarım tablosu ortaya çıkmıştır:

Tablo 9. Başarım Tablosu (Achievement Chart)

EĞİTİM	DOĞRULAMA	VERİ SETİ İÇERİĞİ	ÖN İŞLEMESİZ F1 SKORU	ÖN İŞLEMELİ F1 SKORU
%80	%20	Başlık	0.701	0.730
%80	%20	Özet	0.903	0.925
%80	%20	Başlık & Özet	0.914	0.955

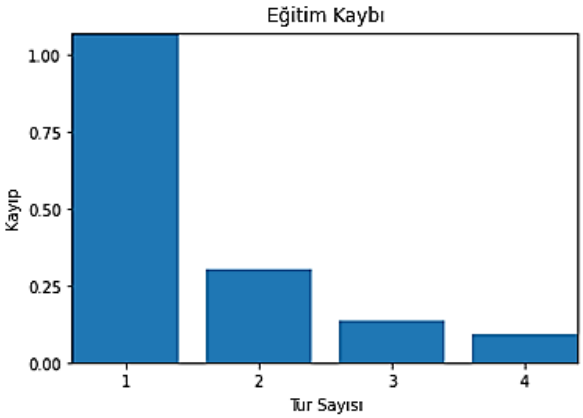
Yanlış tahminde bulunulan değerler ile ortaya çıkan karışıklık matrisi Şekil 4'te paylaşılmıştır:



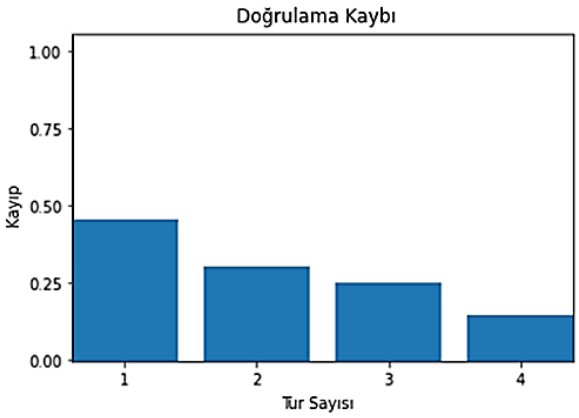
Şekil 4. Karışıklık Matrisi (Confusion Matrix)

Not: Etiket numaraları ilgili kategorileri göstermektedir.

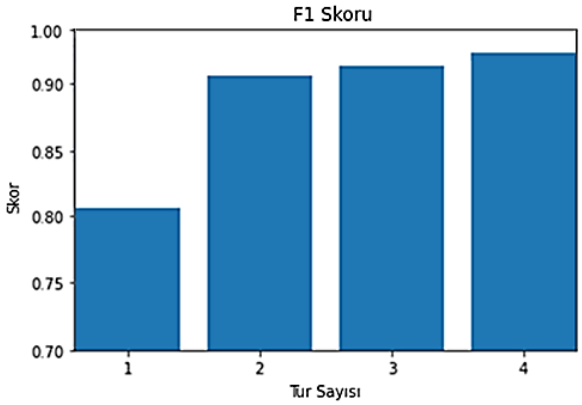
Eğitim sonucu ortaya çıkan her modelin eğitim kaybı, doğrulama kaybı ve F1 skoruna ait grafikler Şekil 5, Şekil 6 ve Şekil 7'de paylaşılmıştır:



Şekil 5. Eğitim Kaybı (Training Loss)



Şekil 6. Doğrulama Kaybı (Validation Loss)



Şekil 7. F1 Skoru (F1 Score)

4. Tartışma (Discussion)

Bilgiye erişimin geçmişe kıyasla çok daha kolay ve pratik hale geldiği günümüzde, bilimsel araştırmalarda elde edilen bulguları, bilimsel bir buluşu ve araştırmaların sonuçlarını ortaya koymaya yarayan bilimsel metinlerin önemi ve sayısı paralel bir şekilde artmaktadır. Her bilimsel metnin kendi alanına, varsa kullanılan teknolojiye hatta ve hatta içerdikleri anahtar kelimelere göre sınıflandırılması, bilgiye erişimin en elverişli şekilde kullanılabilmesi için elzemdir. Doğal dillerin doğaları gereği karmaşık bir yapıya sahip ve net kuralları olmamalarının yanında özellikle Türkçe dilinin morfolojik yapısının karmaşık olması, sınıflandırma problemlerinin çözümünde sorun teşkil etmekte iken, bu sorun Google yapay zeka ekibi mühendisleri tarafından geliştirilen BERT derin öğrenme tekniği ile günümüzde en aza indirgenmiştir.

Bu çalışmada Yükseköğretim Kurulu'na bağlı Ulusal Tez Merkezi ve Türkiye Cumhuriyeti Sanayi ve Teknoloji Bakanlığı'na bağlı Türk Patent ve Marka Kurumu'nun veri tabanından yararlanılmıştır. Python dilinde yazılan script ile veri seti oluşturulmuş ve analiz edilmiştir. Veri seti içerisinde yer alan bilimsel metinler eğitim başarısını arttırmak adına öncelikle bir ön işleme aşamasından geçirilmiş, sonrasında belirlenmiş olan 5 kategoriye göre etiketlenmiştir. Veriler sinir ağı temelli bir model olan BERT ile sınıflandırılmıştır. %80 eğitim ve %20 doğrulama seti olarak ayrılan veri kümesi üzerinde tur sayısı 4 olarak gerçekleştirilen eğitim işlemi sonucunda 0.9551 F1 skoru elde edilmiştir.

Veri Seti Oluşturma başlığında bahsedilen, sadece başlık ve sadece özet bilgilerini içeren veri setleri ile ayrıca iki eğitim daha gerçekleştirilmiştir. Bu iki veri kümesi de %80 eğitim ve %20 doğrulama seti olarak ayrılmıştır. Sadece başlık bilgisini içeren eğitim 0.7306 F1 skoru, sadece özet bilgisini içeren eğitim 0.9259 F1 skoru elde etmiştir.

Eğitim ve Doğrulama Kümesinin Belirlenmesi başlığında bahsedilen, %85 eğitim ve %15 doğrulama seti olarak ayrılan veri kümesi üzerinde tur sayısı 4 olarak gerçekleştirilen eğitim işlemi sonucunda ise 0.9613 F1 skoru elde edilmiştir.

Veri Seti Oluşturma başlığında bahsedilen, sadece başlık ve sadece özet bilgilerini içeren veri setleri ile ayrıca iki eğitim daha gerçekleştirilmiştir. Bu iki veri kümesi de %85 eğitim ve %15 doğrulama seti olarak ayrılmıştır. Sadece başlık bilgisini içeren eğitim 0.7622 F1 skoru, sadece özet bilgisini içeren eğitim 0.9401 F1 skoru elde etmiştir.

Hem %20 hem de %15 doğrulama seti olarak ayrılan veri kümesi üzerinde gerçekleştirilen bütün çalışmalar, veri setinin ön işleme aşamasından geçmemiş hali ile yapıldığından daha önce bahsedilmiştir. Tablo 9 üzerinde görülen başarım sonuçları doğrulama büyüklüğü %15 olarak gerçekleştirilen eğitim sonucunda ortaya çıkan değerler ile aşağıdaki Tablo 10 ile genişletilmiştir.

Sonuç olarak elde edilen bütün değerler ile Tablo 10'daki gibi bir başarım tablosu ortaya çıkmıştır:

Tablo 10. Genişletilmiş Başarım Tablosu (Extended Achievement Chart)

EĞİTİM	DOĞRULAMA	VERİ SETİ İÇERİĞİ	ÖN İŞLEMESİZ F1 SKORU	ÖN İŞLEMELİ F1 SKORU
%80	%20	Başlık	0.701	0.730
%80	%20	Özet	0.903	0.925
%80	%20	Başlık & Özet	0.914	0.955
%85	%15	Başlık	0.754	0.762
%85	%15	Özet	0.919	0.940
%85	%15	Başlık & Özet	0.940	0.961

5. Sonuç (Result)

Bu çalışma dahilinde BERT derin öğrenme tekniği kullanılarak bu çalışmaya özel oluşturulan veri seti üzerinde çalışmalar gerçekleştirilmiş, Tablo 10 üzerinden de anlaşıldığı üzere elde edilen sonuçlar ile yüksek bir performans gösterilmiştir. Türkçe dilinin Çok Sınıflı Sınıflandırma problemlerinin çözümünde kullanılmak üzere, %80 eğitim %20 doğrulama seti ve %85 eğitim %15 doğrulama seti olarak ayrılan veri kümesi üzerinde tur sayısı 4 olarak gerçekleştirilen eğitimler sonucunda model ortaya çıkarılmıştır. Eğitilen bu modelden %80 eğitim %20 doğrulama seti büyüklüğüne sahip olanı 0.955'lik bir başarım oranına sahip olurken, %85 eğitim %15 doğrulama seti büyüklüğüne sahip olan diğer model 0.961'lik bir başarım oranına sahip olmuştur. Daha önce Türkçe dili için BERT derin öğrenme tekniği ile gerçekleştirilen Çok Sınıflı Sınıflandırma çalışmasından (Uçan vd., 2021) F1 skoru olarak 0.04 daha iyi performans ortaya konulmuştur. Daha sonra gerçekleştirilecek olan çalışmalarda veri setinin genişletilmesi, kategorilerin arttırılması, ön işleme aşamasının daha çok detaylandırılması ve Türkçe dilindeki kelime köklerinin daha başarılı bir şekilde ayrıştırılması ile başarıyı arttıracak çalışmalar gerçekleştirilebilir.

Çıkar Çatışması (Conflict of Interest)

Yazarlar tarafından herhangi bir çıkar çatışması beyan edilmemiştir. No conflict of interest was declared by the authors.

Kaynaklar (References)

Acikalin, U. U., Bardak, B., & Kutlu, M. (2020). Turkish Sentiment Analysis Using BERT. In 2020 28th Signal Processing and Communications Applications Conference (SIU) (pp. 1-4). IEEE.

Akin, S. E., & Yildiz, T. (2019, July). Sentiment Analysis through Transfer Learning for Turkish Language. In 2019 IEEE International Symposium on INnovations in Intelligent SysTems and Applications (INISTA) (pp. 1-6). IEEE.

BERTurk. (2020). https://github.com/stefan-it/turkish-bert. (Erişim Tarihi:30.01.2021)

Bisong, E. (2019). Google colaboratory. In Building Machine Learning and Deep Learning Models on Google Cloud Platform (pp. 59-64). Apress, Berkeley, CA.

Chandra, R. V., & Varanasi, B. S. (2015). Python requests essentials. Packt Publishing Ltd.

Çoban, Ö., İnan, A., & Özel, S. A. (2021). Facebook Tells Me Your Gender: An Exploratory Study of Gender Prediction for Turkish Facebook Users. Transactions on Asian and Low-Resource Language Information Processing, 20(4), 1-38.

Deng, L., & Yu, D. (2014). Deep learning: methods and applications. Foundations and trends in signal processing, 7(3–4), 197-387.

Denny, M. J., & Spirling, A. (2018). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. Political Analysis, 26(2), 168-189.

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.

Google Research Bert. (2018). https://github.com/google-research/bert (Erişim Tarihi:07.02.2021)

Jia, Z., Maggioni, M., Smith, J., & Scarpazza, D. P. (2019). Dissecting the NVidia Turing T4 GPU via microbenchmarking. arXiv preprint arXiv:1903.07486.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kraipeerapun, P. (2009). Neural network classification based on quantification of uncertainty (Doctoral dissertation, Murdoch University).

Lee, J. J. (2013). Mechanize: Stateful programmatic web browsing in Python. http://wwwsearch.sourceforge.net/mechanize/ (Erişim Tarihi:17.01.2021)

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234-1240.

Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.
Madabushi, H. T., Kochkina, E., & Castelle, M. (2020). Cost-sensitive BERT for generalisable sentence classification with imbalanced data. arXiv preprint arXiv:2003.11563.
Opitz, J., & Burst, S. (2019). Macro f1 and macro f1. arXiv preprint arXiv:1911.03347.
Özçift, A., Akarsu, K., Yumuk, F., & Söylemez, C. (2021). Advancing natural language processing (NLP) applications of morphologically rich languages with bidirectional encoder representations from transformers (BERT): an empirical case study for Turkish. Automatika, 1-13.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.
Richardson, L. (2007). Beautiful soup documentation. https://www. crummy. com/software/BeautifulSoup/bs4/doc/. (Erişim Tarihi:15.01.2021)
Schachinger, K. (2017). A Complete Guide to the Google RankBrain Algorithm. Search Engine Journal.
Sevli, O , Kemaloğlu, N . (2021). Olağandışı Olaylar Hakkındaki Tweet'lerin Gerçek ve Gerçek Dışı Olarak Google BERT Modeli ile Sınıflandırılması . Veri Bilimi , 4 (1) , 31-37 .
Song, K., Tan, X., Qin, T., Lu, J., & Liu, T. Y. (2020). Mpnet: Masked and permuted pre-training for language understanding. arXiv preprint arXiv:2004.09297.
Şahin, G., & Diri, B. (2021, June). The Effect of Transfer Learning on Turkish Text Classification. In 2021 29th Signal Processing and Communications Applications Conference (SIU) (pp. 1-4). IEEE.
Tantuğ, A. C. (2016). Metin Sınıflandırma. Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, 5(2).
Tuzcu, S. (2020). Çevrimiçi Kullanıcı Yorumlarının Duygu Analizi ile Sınıflandırılması. Eskişehir Türk Dünyası Uygulama ve Araştırma Merkezi Bilişim Dergisi, 1(2), 1-5.
Uçan, A., Dörterler, M., & Akçapınar Sezer, E. (2021). A study of Turkish emotion classification with pretrained language models. Journal of Information Science, 0165551520985507.
What's New In Python 3.7. (2018). https://docs.python.org/3.7/whatsnew/3.7.html (Erişim Tarihi:18.04.2021)