

DistilBERT Tabanlı Metin Sınıflandırmada Veri Verimliliği: Farklı Veri Boyutlarında Bir Performans Analizi

Data Efficiency in DistilBERT-Based Text Classification: A Performance Analysis Across Different Training Data Sizes

Hülya SAYGILI^{1*} 

Öz

Bu çalışmada, DistilBERT tabanlı metin sınıflandırma modelinde eğitim veri kümesi boyutunun model performansı ve genelleme kapasitesi üzerindeki etkisi deneysel olarak incelenmiştir. Çalışma kapsamında, farklı dilsel yapılar ve görev karakteristiklerini temsil edecek şekilde seçilen beş ayrı veri seti (AG News, IMDb, DBpedia, TTC-4900 ve Türkçe e-ticaret ürün yorumları veri seti) üzerinde kapsamlı deneyler gerçekleştirilmiştir. Deneylerde, tüm hiperparametreler sabit tutulmuş ve yalnızca eğitim veri miktarı kademeli olarak değiştirilerek küçük veri boyutlarından büyük ölçekli veri yapılarına kadar model davranışı analiz edilmiştir. Model performansı doğruluk, Macro F1 skoru ve kayıp metrikleri üzerinden değerlendirilmiştir. Eğitim ve doğrulama veri kümelerinin sonuçları karşılaştırılarak genelleme eğilimleri ortaya konulmuştur. Elde edilen bulgular, eğitim veri boyutunun model başarımı üzerinde belirleyici bir etkiye sahip olduğunu, ancak bu etkinin görev türüne, dil yapısına ve veri gürültüsüne bağlı olarak değişkenlik gösterdiğini göstermektedir. Özellikle büyük ölçekli ve yapısal veri setlerinde daha yüksek doğruluk, daha düşük kayıp ve daha küçük genelleme farkı ile erken performans doygunluğu gözlemlenirken, öznel ve gürültülü metinlerden oluşan veri setlerinde model performansındaki iyileşmenin daha sınırlı olduğu belirlenmiştir. Bu sonuçlar, DistilBERT tabanlı modelde veri verimliliğinin yalnızca model mimarisiyle değil, aynı zamanda görev ve veri karakteristiği ile birlikte değerlendirilmesi gerektiğini ortaya koymaktadır.

Anahtar Kelimeler: DistilBERT, Veri boyutu, Metin sınıflandırma, Genelleme kapasitesi, Doğal dil işleme.

Abstract

In this study, the impact of training dataset size on the performance and generalization capability of a DistilBERT-based text classification model was experimentally investigated. Within the scope of the research, comprehensive experiments were conducted on five distinct datasets (AG News, IMDb, DBpedia, TTC-4900, and the Turkish E-commerce Product Reviews dataset), selected to represent diverse linguistic structures and task characteristics. In the experiments, all hyperparameters were kept constant, and model behavior was analyzed across datasets ranging from small to large sizes by incrementally varying only the amount of training data. Model performance was evaluated in terms of accuracy, Macro F1-score and loss. Generalization trends were identified by comparing the performance on the training and validation sets. The findings indicate that while training data size has a decisive impact on model performance, this effect varies depending on the task type, linguistic structure, and data noise. In particular, large-scale and structured datasets achieved earlier performance saturation, characterized by higher accuracy, lower loss, and a smaller generalization gap, whereas performance improvements were more gradual and limited in datasets composed of subjective and noisy texts. These results demonstrate that data efficiency in the DistilBERT-based model should be evaluated not only in terms of the model architecture but also in conjunction with the task and data characteristics.

Keywords: DistilBERT, Data size, Text classification, Generalization capacity, Natural language processing.

¹Karayolları 8.Bölge Müdürlüğü, Elazığ, Türkiye

*Sorumlu Yazar/Corresponding Author: hulya-saygili@hotmail.com

1. Giriş

Yapay zeka ve doğal dil işleme alanındaki gelişmeler, özellikle Transformer tabanlı önceden eğitilmiş dil modellerinin (BERT, RoBERTa, DistilBERT vb.) kullanıma sunulmasıyla hız kazanmıştır. Transformer tabanlı dil modelleri, büyük metin koleksiyonları üzerinde ön eğitim (pre-training) süreciyle dilin anlamsal ve bağlamsal özelliklerini öğrenen derin sinir ağı modelleridir (Arzu ve Aydoğan, 2023). Dilin anlamsal ve yapısal özelliklerini geniş veri kitlelerinden öğrenen bu modeller, sınıflandırma ve duygu analizi gibi spesifik görevlere uyarlandıklarında üstün performans göstermektedir.

Literatürde, özellikle büyük dil modellerinin ön eğitim (pre-training) sürecinde eğitim verisi miktarının model performansı üzerinde belirleyici bir etkisi olduğu gösterilmiştir (Kaplan vd., 2020). Eğitim verisinin miktarı ile model başarımı arasındaki ilişki çoğunlukla doğruluk değerleri üzerinden ele alınmaktadır. Transformer tabanlı modellerin karmaşık yapısı ve yüksek parametre sayısı, özellikle küçük veri kümeleriyle gerçekleştirilen eğitimlerde genelleme performansının sınırlı kalmasına yol açabilmektedir. Eğitim başarımı artarken doğrulama başarımının aynı oranda iyileşmediği ve eğitim ile doğrulama başarımları arasında belirgin farkların olduğu gözlemlenmektedir. Bu nedenle, eğitim veri boyutunun yalnızca performans üzerindeki etkisinin değil, aynı zamanda modelin öğrenme sürecini ve genelleme davranışını nasıl şekillendirdiğinin de ayrıntılı biçimde incelenmesi gerekmektedir.

Çalışmanın kapsamı, tek bir dil ve görevle sınırlı tutulmamış; İngilizce ve Türkçe dillerinde nesnel haber sınıflandırma, öznel duygu analizi ve ontolojik kategorizasyon gibi farklı karakteristiklere sahip beş ayrı veri seti üzerinden karşılaştırmalı bir analiz gerçekleştirilmiştir. Bu yaklaşım, veri miktarının etkisinin dilsel yapı ve görev türüne göre nasıl değiştiğini deneysel olarak ortaya koymayı amaçlamaktadır.

Bu çalışmada, DistilBERT tabanlı bir metin sınıflandırma modelinde eğitim veri kümesi boyutunun model başarımı ve genelleme kapasitesi üzerindeki etkisi deneysel olarak incelenmiştir. Farklı büyüklüklerde eğitim veri kümeleri kullanılarak gerçekleştirilen deneylerde, tüm hiperparametreler sabit tutulmuş ve yalnızca kullanılan eğitim örneği sayısı kademeli olarak değiştirilmiştir. Model başarımı doğruluk, Macro F1 skoru ve kayıp metrikleri üzerinden değerlendirilmiştir. Eğitim ve doğrulama veri kümelerinden elde edilen sonuçlar karşılaştırmalı biçimde analiz edilerek modelin genelleme davranışı ortaya konulmuştur. Bu çalışma, DistilBERT tabanlı metin sınıflandırma modelinde eğitim veri boyutunun etkisini dil, görev türü ve veri yapısı bağlamında ele alarak, farklı senaryolar altında veri verimliliğine ilişkin kapsamlı bir deneysel değerlendirme sunmaktadır.

Makalenin kalan bölümleri şu şekilde yapılandırılmıştır: Bölüm 2’de ilişkili literatüre yer verilmiştir. Bölüm 3’te materyal ve yöntem ele alınmıştır. Bölüm 4’te, deneysel çalışmalar sonucunda elde edilen bulgular sunulmaktadır. Son olarak, Bölüm 5’te bu sonuçlar değerlendirilerek çalışmanın genel çıkarımları özetlenmiştir.

2. İlişkili çalışmalar

Kaplan ve ark. (2020), dil modeli performansının çapraz entropi kaybı açısından model boyutu, eğitim veri miktarı ve hesaplama bütçesi ile güç yasası (power-law) ilişkisi sergilediğini göstermiştir. Çalışmada, sabit bir hesaplama bütçesi altında daha büyük modellerin daha az veriyle eğitilmesinin hesaplama açısından verimli olabileceği gösterilmiştir. Hoffmann ve ark. (2022), model boyutu ölçeklenirken eğitim veri miktarının yeterince artırılmamasının büyük dil modellerinin yetersiz eğitilmesine yol açtığını ortaya koymuşlardır. Bu çalışmada, hesaplama açısından optimal eğitimin, model parametre sayısı ile kullanılan token miktarının birlikte ve orantılı olarak ölçeklendirilmesiyle sağlandığı gösterilmiştir. Bu yaklaşım doğrultusunda geliştirdikleri Chinchilla modelinin, benzer bütçeye sahip daha büyük modellerden daha yüksek genelleme başarımı elde ettiği rapor edilmiştir. Bu çalışmalar, büyük dil modellerinin ön eğitim (pre-training) sürecinde veri miktarı, model boyutu ve hesaplama bütçesi arasındaki ilişkinin önemini ortaya koymaktadır. Buna karşın, bu çalışmada önceden eğitilmiş DistilBERT modeli üzerinde ince ayar (fine-tuning) uygulanmıştır. Dolayısıyla bu çalışmanın amacı, ölçeklenme yasalarını doğrudan doğrulamak değil, farklı eğitim veri boyutlarının ince ayar sürecindeki model performansı ve genelleme davranışı üzerindeki etkisini deneysel olarak incelemektir.

Schwartz ve ark. (2020), derin öğrenme alanında kullanılan hesaplama miktarının son yıllarda hızlı biçimde arttığını ve bunun önemli bir çevresel ve ekonomik maliyet oluşturduğunu vurgulamıştır. Çalışmada, derin öğrenme araştırmalarında başarımların ölçütlerinin yalnızca doğruluk gibi metriklerle sınırlandırılmaması, bunun yerine hesaplama verimliliğinin de değerlendirme ölçütü olarak ele alınması gerektiği savunulmuştur.

Genel modellerin yanı sıra dile özgü mimariler de literatürde geniş yer bulmaktadır. Arzu ve Aydoğan (2023), Transformer tabanlı mimarilerin Türkçe duygu sınıflandırma görevlerindeki başarımını incelemiştir. Bu kapsamda, TRSAv1 veri seti kullanılarak gerçekleştirilen bir çalışmada, farklı BERTurk ve ELECTRA tabanlı önceden eğitilmiş modeller duygu sınıflandırma görevinde değerlendirilmiştir. Modeller, farklı eğitim stratejileri altında eğitilmiş ve performansları doğruluk ve F1-skor metrikleri üzerinden karşılaştırılmıştır. Elde ettikleri deneysel sonuçlar, Transformer tabanlı modellerin Türkçe duygu sınıflandırma görevlerinde etkili bir biçimde kullanılabildiğini göstermişlerdir.

Eşidir (2025), sosyal medya verileri üzerinde gerçekleştirdiği çalışmada, Transformer tabanlı DistilBERT modelinin duygu sınıflandırma görevlerinde geleneksel makine öğrenmesi yöntemlerine göre daha yüksek performans sunduğunu göstermiştir. Özellikle dengesiz sınıf dağılımına sahip veri setlerinde DistilBERT'in genelleme başarımının daha yüksek olduğunu rapor etmiştir. Bu bulgular, Transformer tabanlı yaklaşımların metin sınıflandırma problemlerindeki etkinliğini desteklemektedir.

3. Materyal ve Metot

Bu bölümde, araştırmada kullanılan veri setlerinin nitelikleri, seçilen model mimarisinin teknik detayları, deney tasarımı ve performans değerlendirme süreçleri açıklanmaktadır.

3.1. Veri Setleri

Çalışmada kullanılan veri setleri, dilsel çeşitlilik ve görev zorluğu açısından heterojen bir yapı sunacak şekilde seçilmiştir. Kullanılan veri setlerinin istatistiksel özetleri Tablo 1'de sunulmuştur:

Tablo 1. Deneysel çalışmalarda kullanılan veri setlerinin özellikleri

Veri Seti	Dil	Görev Türü	Toplam Örnek Sayısı	Sınıf Sayısı
AG News	İngilizce	Haber Sınıflandırma	127.600	4
TTC-4900	Türkçe	Haber Sınıflandırma	4.900	6
IMDB	İngilizce	Duygu Analizi	50.000	2
E-ticaret	Türkçe	Duygu Analizi	11021	2
DBpedia	İngilizce	Metin Sınıflandırma	630.000	14

AG News: Dünya, Spor, İş ve Bilim/Teknoloji başlıklarında dengeli dağılmış haber metinlerinden oluşmaktadır (Zhang vd., 2015).

TTC-4900: Türkçedeki haber sınıflandırma başarısını ölçmek için kullanılan; siyaset, ekonomi, teknoloji, spor, dergi ve sağlık kategorilerini içeren veri setidir (Kılınç vd., 2015).

IMDB: Sinema eleştirilerini içeren bu set, uzun cümle yapıları ve öznel ifadeler barındırması nedeniyle modelin duygu yakalama kabiliyetini test eder (Maas vd., 2011).

E-ticaret ürün yorumları: Hepsiburada, Trendyol ve N11 gibi e-ticaret platformlarından e-ticaret platformu üzerinden kullanıcılar tarafından oluşturulan Türkçe ürün yorumlarından oluşmaktadır. Veri setinde 6799 olumlu, 6978 olumsuz ve 1393 nötr yorum bulunmaktadır (Çabuk, 2022). E-ticaret veri setinde nötr sınıf deney kapsamı dışında bırakılmış, analizler olumlu ve olumsuz yorumlar üzerinden iki sınıflı duygu analizi görevi olarak yürütülmüştür.

DBpedia: Wikipedia'nın yapılandırılmış bilgi kaynaklarından türetilmiş kısa metinlerden oluşan ve metinlerin ait oldukları anlamsal varlık türlerinin (örneğin şirket, kişi, kurum, yer vb.) sınıflandırılmasını amaçlayan çok sınıflı bir metin sınıflandırma veri setidir (Lehman vd., 2015).

3.2. Model Mimarisi

3.2.1. DistilBERT Mimarisi

DistilBERT; BERT (Bidirectional Encoder Representations from Transformers) modelinin damıtılmış, daha küçük ve daha hızlı bir varyantıdır. Orijinal BERT modelinin boyutunu %40 oranında azaltırken performansının yaklaşık %97'sini korumak üzere geliştirilmiştir. DistilBERT modelde belirteç tipi gömülmeleri ve havuzlayıcı katmanlar kaldırılarak katman sayısı iki kat azaltılmıştır (Sanh vd.,2019). Modelin seçilme nedenleri şunlardır:

- **Hesaplama Verimliliği:** BERT modeline kıyasla %60 daha hızlı olması, farklı veri büyüklüklerinde çok sayıda deneyin gerçekleştirilmesini mümkün kılmıştır.
- **Küçük Veri Rejimi Uyumu:** Literatürde, damıtılmış modellerin sınırlı veri setlerinde bile güçlü anlamsal temsiller sunduğu gösterilmiştir (Tang, R. vd.,2019).
- **Dil Desteği:** İngilizce veri setleri için distilbert-base-uncased, Türkçe veri setleri için ise 104 dili destekleyen ve Türkçe üzerinde yüksek başarıyı kanıtlanmış olan distilbert-base-multilingual-cased modelleri kullanılmıştır (Ünal vd., 2023).

3.2.2. Deneysel Tasarım ve Eğitim Prosedürü

Eğitim veri boyutunun model performansı ve genelleme kapasitesi üzerindeki etkisini incelemek amacıyla kademeli veri artırımı stratejisi uygulanmıştır. Deney süreci şu adımlardan oluşmaktadır:

- **Veri Örneklemesi:** Çalışma kapsamında deneysel tutarlılığı ve literatürdeki diğer çalışmalarla doğrudan karşılaştırılabilirliği sağlamak amacıyla, veri setlerinin orijinal kaynaklarında yer alan standart eğitim ve test bölünmeleri korunmuştur. Tablolarda "Tam Veri Seti" olarak belirtilen büyüklükler, deneylerde kullanılan en büyük eğitim örneklemelerini temsil etmektedir. AG News, IMDb ve E-ticaret veri setlerinde bu büyüklükler ilgili veri setlerinin orijinal eğitim kümelerine karşılık gelirken, DBpedia veri setinde hesaplama maliyetini makul düzeyde tutabilmek amacıyla orijinal 560.000 örnekten oluşan eğitim kümesi yerine rastgele seçilmiş 56.000 örnekten oluşan eğitim alt kümesi kullanılmıştır. TTC-4900 veri setinde ise çalışmada kullanılan veri dosyası toplam 4900 örnek içermekte olup, standart %80-%20

eğitim-test bölünmesi sonucunda en büyük eğitim örneklemini 3920 örnekten oluşmuştur. Kademeli alt örneklemeler (500, 1000, 3000, 5000 ve 10000 örnek) eğitim kümeleri içerisinde rastgele örnekleme yöntemiyle oluşturulmuştur. Deneylerin tekrarlanabilirliğini ve sonuçların güvenilirliğini değerlendirmek amacıyla her deney beş farklı rastgelelik tohumu (seed = 42, 52, 62, 72 ve 82) kullanılarak bağımsız olarak yürütülmüştür. Veri setlerinin orijinal test kümeleri ise modellerin nihai genelleme performansını değerlendirmek amacıyla eğitim sürecinden tamamen bağımsız tutulmuştur.

- **Model Mimarisi:** Çalışmada, hesaplama maliyeti düşük, eğitim süresi kısa ve literatürde yaygın olarak kullanılan bir Transformer modeli olması nedeniyle DistilBERT tercih edilmiştir. Böylece farklı eğitim veri boyutlarının model performansı üzerindeki etkisi, model mimarisinden kaynaklanabilecek farklılıklar en aza indirilerek incelenmiştir.
- **Sabit Hiperparametreler:** Veri boyutunun etkisini net bir şekilde gözlemleyebilmek adına tüm deneylerde hiperparametre seti sabit tutulmuştur. Bu kapsamda; öğrenme oranı 2×10^{-5} , yığın boyutu 32 (AG News veri seti için 64) ve eğitim iterasyon sayısı veri setine uygun şekilde sabit tutulmuştur.
- **Donanım ve Optimizasyon:** Eğitimler NVIDIA GPU donanımı üzerinde gerçekleştirilmiştir. Hesaplama süresini optimize etmek ve bellek verimliliğini artırmak amacıyla Karma Hassasiyetli Eğitim (Mixed Precision Training- FP16) tekniği ve AdamW optimizasyon algoritması kullanılmıştır. Karma Hassasiyetli Eğitim (Mixed Precision Training), model eğitiminde FP32 (Full Precision) yerine FP16 (Half Precision) kullanarak bellek tüketimini minimize eden ve hesaplama hızını artıran bir tekniktir. Bu yöntem, model ağırlıklarını ve aktivasyonlarını 16-bit hassasiyetinde tutarken, ağırlık güncellemelerini 32-bit hassasiyetinde gerçekleştirerek hem eğitim süresinden tasarruf sağlar hem de modelin nihai doğruluğunun korunmasını garanti eder (Micikevicius vd., 2017).
- **Değerlendirme Stratejisi:** Modelin her bir veri dilimindeki başarımı, eğitim sürecine dahil edilmeyen bağımsız test kümesi üzerinde doğruluk, Macro F1 skoru ve kayıp değerleri kullanılarak değerlendirilmiştir. Deneysel sonuçların rastgele başlangıç koşullarından etkilenme düzeyini değerlendirmek amacıyla her deney beş farklı rastgelelik tohumu (seed = 42, 52, 62, 72 ve 82) kullanılarak bağımsız olarak tekrarlanmış ve performans metrikleri ortalama $\pm$ standart sapma (Ort $\pm$ SS) biçiminde raporlanmıştır. Ayrıca eğitim ve test performansları birlikte incelenerek modelin farklı veri boyutlarındaki genelleme davranışı analiz edilmiştir.
- **Yazılım Ortamı ve Kütüphaneler:** Deneysel çalışmaların tutarlılığı ve tekrarlanabilirliğini sağlamak amacıyla, tüm süreç Anaconda paket yönetim sistemi üzerinde oluşturulan izole bir sanal ortamda yürütülmüştür. Yazılım geliştirme dili olarak Python 3.12.7 tercih edilmiş;

model eğitimi ve veri işleme süreçleri için Hugging Face Transformers, PyTorch ve Scikit-learn kütüphanelerinden yararlanılmıştır. GPU hızlandırmalı hesaplamalar için NVIDIA CUDA 11.x araç kiti kullanılarak donanım optimizasyonu sağlanmıştır.

4. Bulgular

Bu bölümde, seçilen beş farklı veri seti üzerinde gerçekleştirilen deneylerden elde edilen sonuçlar sunulmaktadır. Her bir veri boyutu için doğruluk ve kayıp değerleri analiz edilmiştir.

Tablo 2 incelendiğinde, eğitim veri boyutu arttıkça hem eğitim hem de test doğruluğunun düzenli biçimde yükseldiği görülmektedir. En düşük veri boyutu olan 500 örnekte test doğruluğu 0.7788 ± 0.0695 ve Macro F1 skoru 0.7641 ± 0.0953 olarak elde edilirken, tam eğitim kümesi kullanıldığında bu değerler sırasıyla 0.9452 ± 0.0008 ve 0.9452 ± 0.0008 seviyesine ulaşmıştır. Bu durum, eğitim veri miktarındaki artışın modelin sınıflandırma başarısını ve genelleme performansını önemli ölçüde iyileştirdiğini göstermektedir.

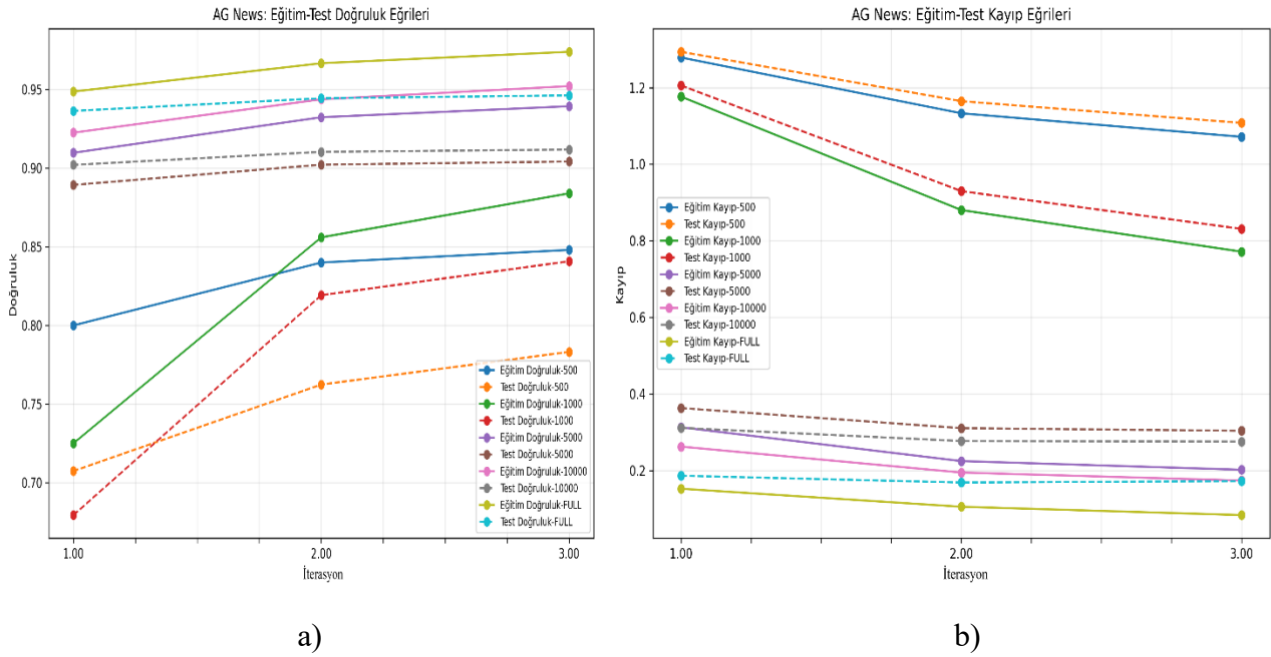
Eğitim veri boyutunun artmasıyla birlikte eğitim ve test kayıplarının düzenli olarak azaldığı gözlenmiştir. Eğitim kaybı 1.0985 ± 0.0200 değerinden 0.0837 ± 0.0007 'ye, test kaybı ise 1.1268 ± 0.0120 değerinden 0.1707 ± 0.0020 'ye düşmüştür. Bu sonuçlar, modelin daha fazla eğitim verisi ile daha kararlı bir öğrenme süreci gerçekleştirdiğini göstermektedir.

Genelleme farkı (eğitim doğruluğu – test doğruluğu) tüm veri boyutlarında düşük seviyelerde kalmış olup, tam veri kümesi kullanıldığında 0.0288 ± 0.0007 ile en düşük değerine ulaşmıştır. Ayrıca küçük veri boyutlarında standart sapma değerlerinin daha yüksek, büyük veri boyutlarında ise oldukça düşük olması, sınırlı veri koşullarında model performansının rastgele başlangıç koşullarından daha fazla etkilendiğini; eğitim veri miktarı arttıkça ise modelin daha kararlı ve tekrarlanabilir sonuçlar ürettiğini göstermektedir.

Tablo 2. AG NEWS Sonuçlar

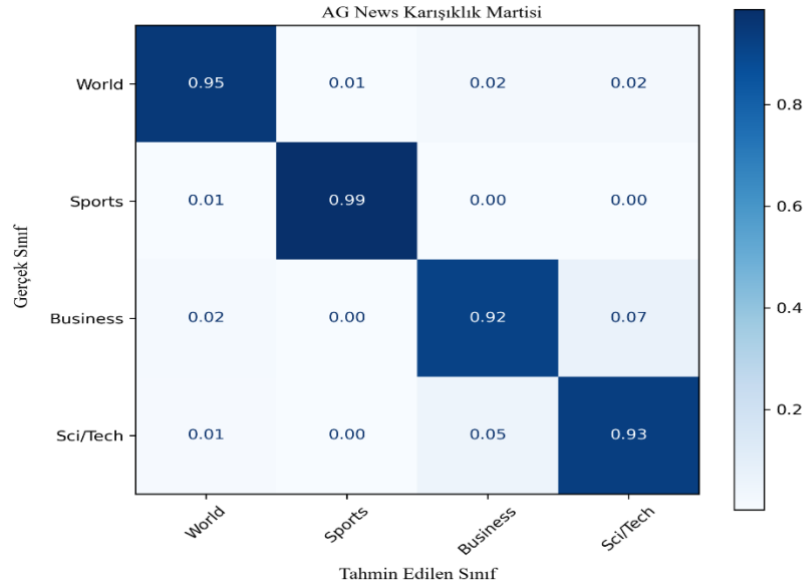
Veri Boyutu	Eğitim Doğruluk (Ort±SS)	Test Doğruluk (Ort±SS)	Macro F1 (Ort±SS)	Eğitim Kaybı (Ort±SS)	Test Kaybı (Ort±SS)	Genelleme Farkı (Ort±SS)
500	0.8200 ±0.0751	0.7788 ±0.0695	0.7641 ±0.0953	1.0985 ±0.0200	1.1268 ±0.0120	0.0412 ±0.0283
1000	0.8862 ±0.0041	0.8563 ±0.0062	0.8553 ±0.0057	0.7658 ±0.0270	0.7976 ±0.0229	0.0299 ±0.0077
5000	0.9407 ±0.0033	0.9029 ±0.0023	0.9027 ±0.0022	0.2163 ±0.0084	0.3047 ±0.0039	0.0377 ±0.0040
10000	0.9503 ±0.0038	0.9128 ±0.0016	0.9127 ±0.0016	0.1728 ±0.0076	0.2676 ±0.0069	0.0375 ±0.0038
120000	0.9740 ±0.0003	0.9452 ±0.0008	0.9452 ±0.0008	0.0837 ±0.0007	0.1707 ±0.0020	0.0288 ±0.0007

Şekil 1'de AG News veri seti için farklı eğitim veri boyutlarında elde edilen eğitim-test doğruluk ve kayıp eğrileri sunulmaktadır. Eğitim veri miktarı arttıkça modelin hem eğitim hem de test doğruluğunun düzenli olarak yükseldiği görülmektedir. Özellikle 500 ve 1000 örnek kullanılan deneylerde ilk iterasyonlarda belirgin performans artışı gözlenirken, 5000 örnek ve üzerindeki veri boyutlarında doğruluk değerleri daha yüksek seviyelerde kararlı hâle gelmiştir. Tam eğitim kümesi kullanıldığında model yaklaşık %95 test doğruluğuna ulaşarak en yüksek performansı elde etmiştir. Benzer şekilde eğitim ve test kayıp değerleri tüm veri boyutlarında iterasyon ilerledikçe düzenli olarak azalmış, büyük veri kümelerinde daha düşük kayıp değerlerine ulaşılması modelin parametrelerini daha etkin optimize edebildiğini ve genelleme performansının iyileştiğini göstermektedir. Bununla birlikte ikinci ve üçüncü iterasyon arasında performans artışının sınırlı kalması, modelin kısa sürede yakınsadığını ve üç iterasyon sonunda kararlı bir öğrenme sürecine ulaştığını göstermektedir.



Şekil 1. AG News veri setine ait doğruluk ve kayıp grafikleri.

Şekil 2'de verilen normalize edilmiş karmaşıklık matrisi incelendiğinde, modelin tüm sınıflarda yüksek sınıflandırma başarısı gösterdiği görülmektedir. En yüksek doğru sınıflandırma oranı Sports (%99) sınıfında elde edilirken, World (%95), Sci/Tech (%93) ve Business (%92) sınıflarında da yüksek doğruluk sağlanmıştır. Yanlış sınıflandırmaların büyük bölümü Business ile Sci/Tech sınıfları arasında yoğunlaşmıştır. Bunun temel nedeni, bu iki sınıfa ait haberlerin teknoloji şirketleri, finansal teknoloji ve ekonomik gelişmeler gibi benzer kavramsal içerikler barındırmasıdır. Genel olarak karmaşıklık matrisi, modelin sınıflar arasında yüksek ayırt edicilik sergilediğini ve elde edilen yüksek doğruluk ile Macro F1 skorlarını desteklediğini göstermektedir.



Şekil 2. AG News veri setine ait karmaşıklık matrisi

IMDb veri setine ait deney sonuçları Tablo 3'te sunulmuştur. Bulgular, eğitim veri boyutu arttıkça model performansının düzenli olarak iyileştiğini göstermektedir. En düşük veri boyutu olan 500 örnekte test doğruluğu %72,28 ve Macro F1 skoru 0,7227 olarak elde edilirken, eğitim veri miktarının artırılmasıyla birlikte bu değerler sürekli yükselmiş ve tam eğitim kümesinde (25.000 örnek) sırasıyla %87,52 ve 0,8752 değerlerine ulaşmıştır. Bu durum, DistilBERT modelinin daha fazla eğitim örneğiyle metinlerin anlamsal özelliklerini daha etkin öğrenebildiğini ve sınıflar arasındaki ayırt edici temsilleri daha başarılı oluşturabildiğini göstermektedir.

Eğitim kaybı veri boyutunun artmasına bağlı olarak 0,5493'ten 0,0603'e kadar önemli ölçüde azalırken, test kaybı da 0,6018'den 0,3115'e gerilemiştir. Kayıp değerlerindeki bu düşüş, model parametrelerinin daha kararlı optimize edildiğini ve öğrenme sürecinin veri miktarından olumlu yönde etkilendiğini göstermektedir. Bununla birlikte genelleme farkının 0,0667'den 0,1135'e yükselmesi, eğitim başarısındaki artışın test performansına aynı oranda yansımadığını ve özellikle büyük veri boyutlarında modelin eğitim verisini daha iyi öğrenmesine rağmen sınırlı düzeyde aşırı öğrenme eğilimi göstermeye başladığını ortaya koymaktadır.

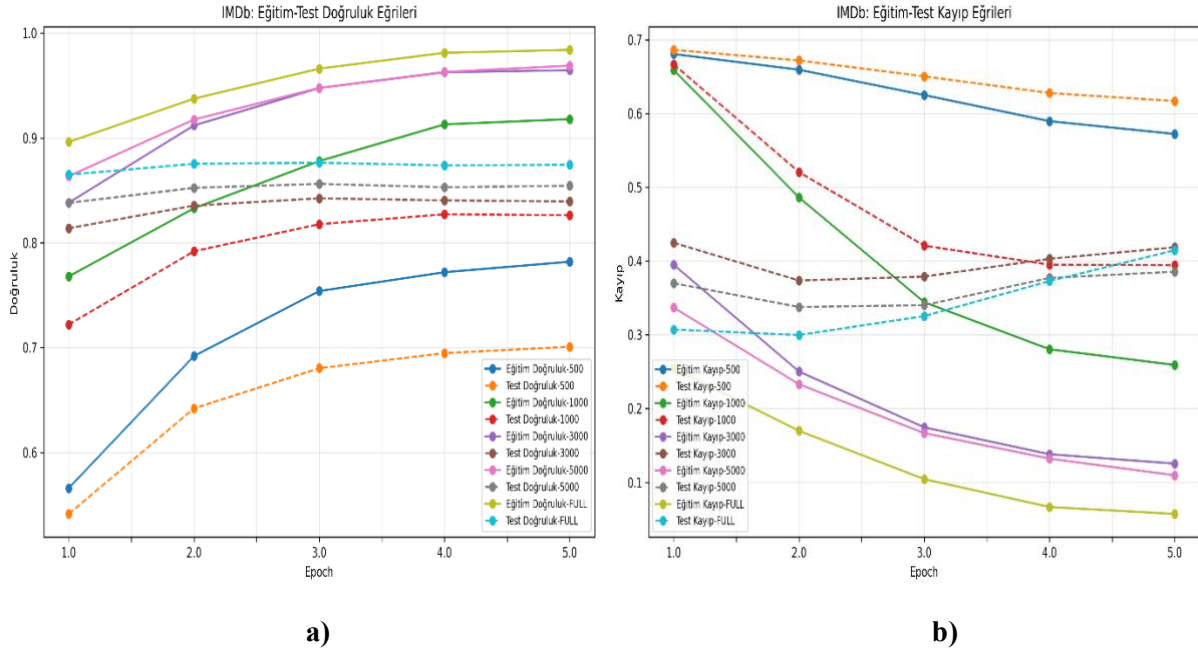
Ayrıca, tüm deneylerde standart sapma değerlerinin oldukça düşük olması (özellikle tam eğitim kümesinde test doğruluğu için $\pm 0,0009$) modelin farklı rastgelelik tohumları altında benzer sonuçlar ürettiğini ve deneylerin yüksek düzeyde kararlılık ve tekrarlanabilirlik sağladığını göstermektedir.

Tablo 3. IMDB Sonuçlar

Veri Boyutu	Eğitim Doğruluk (Ort±SS)	Test Doğruluk (Ort±SS)	Macro F1 (Ort±SS)	Eğitim Kaybı (Ort±SS)	Test Kaybı (Ort±SS)	Genelleme Farkı (Ort±SS)
500	0.7895 ±0.0346	0.7228 ±0.0258	0.7227 ±0.0259	0.5493 ±0.0745	0.6018 ±0.0197	0.0667 ±0.0243
1000	0.8934 ±0.0067	0.8216 ±0.0155	0.8215 ±0.0155	0.2827 ±0.0221	0.4342 ±0.0186	0.0718 ±0.0153
3000	0.9491 ±0.0039	0.8429 ±0.0058	0.8429 ±0.0058	0.1312 ±0.0106	0.3901 ±0.0144	0.1062 ±0.0071
5000	0.9646 ±0.0021	0.8525 ±0.0043	0.8525 ±0.0043	0.0964 ±0.0065	0.3728 ±0.0112	0.1121 ±0.0046
25000	0.9887 ±0.0004	0.8752 ±0.0009	0.8752 ±0.0009	0.0603 ±0.0013	0.3115 ±0.0031	0.1135 ±0.0009

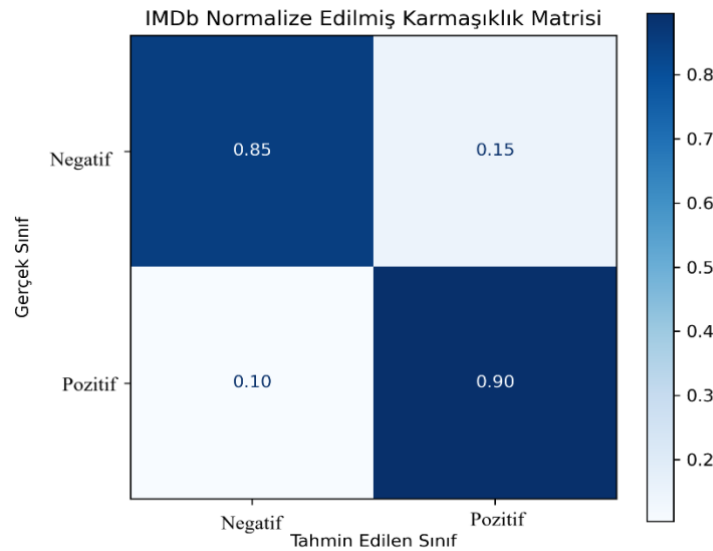
Şekil 3.a incelendiğinde, tüm veri boyutlarında eğitim ve test doğruluklarının iterasyonlar boyunca düzenli olarak arttığı görülmektedir. Özellikle 500 ve 1000 örnekli deneylerde ilk iki iterasyonda belirgin performans kazanımları elde edilirken, sonraki iterasyonlarda artış hızının azaldığı gözlenmiştir. Veri miktarının artmasıyla model daha yüksek doğruluk değerlerine ulaşmış; tam eğitim kümesinde eğitim doğruluğu yaklaşık %98, test doğruluğu ise %87–88 düzeyinde kararlı hale gelmiştir. Eğitim ve test doğrulukları arasındaki farkın sınırlı kalması, modelin farklı veri boyutlarında istikrarlı bir öğrenme süreci sergilediğini göstermektedir.

Şekil 3.b'de ise tüm veri boyutlarında eğitim kaybının iterasyonlarla birlikte düzenli olarak azaldığı görülmektedir. En belirgin düşüş 1000, 3000 ve 5000 örnekli deneylerde ilk üç iterasyonda gerçekleşmiştir. Test kaybı başlangıçta azalırken yüksek veri boyutlarında son iterasyonlarda durağanlaşmış ve sınırlı değişimler göstermiştir. Buna karşın test doğruluğunun büyük ölçüde korunması, modelin hafif düzeyde aşırı öğrenme eğilimi göstermesine rağmen genelleme yeteneğini koruduğunu ortaya koymaktadır.



Şekil 3. IMDB veri setine ait doğruluk ve kayıp grafikleri.

Şekil 4 ile gösterilen normalize edilmiş karmaşıklık matrisi, modelin her iki sınıfı da yüksek doğrulukla ayırt edebildiğini göstermektedir. Negatif yorumların yaklaşık %85'i doğru sınıflandırılırken %15'i pozitif olarak tahmin edilmiştir. Pozitif yorumlarda ise doğru sınıflandırma oranı yaklaşık %90 olup yalnızca %10'luk bir kısmı negatif sınıfa atanmıştır. Sonuçlar modelin pozitif yorumları ayırt etmede negatif yorumlara göre daha başarılı olduğunu göstermektedir. Genel olarak yanlış sınıflandırmaların düşük seviyede kalması, DistilBERT modelinin IMDB veri kümesinde duygu analizi görevinde başarılı bir sınıflandırma performansı sergilediğini doğrulamaktadır. Ayrıca iki sınıf arasındaki hata dağılımının dengeli olması, modelin belirli bir sınıfa sistematik olarak yönelmediğini ve dengeli bir karar mekanizması geliştirdiğini göstermektedir.



Şekil 4. IMDB veri setine ait karmaşıklık matrisi

Tablo 4 incelendiğinde, eğitim veri boyutunun artmasının model performansı üzerinde belirgin bir iyileşme sağladığı görülmektedir. Özellikle 500 ve 1000 örnekten oluşan düşük veri rejimlerinde hem test doğruluğu hem de Macro F1 skoru değerleri görece düşük kalırken, standart sapma değerlerinin yüksek olması model performansının rastgele başlangıç koşullarına daha duyarlı olduğunu göstermektedir. Bunun temel nedeni, 14 sınıftan oluşan DBpedia veri kümesinde sınıf başına düşen örnek sayısının yetersiz olması ve modelin sınıflar arasındaki anlamsal ayrımları yeterince öğrenememesidir.

Eğitim veri boyutunun 5000 örneğe çıkarılmasıyla birlikte test doğruluğu ve Macro F1 skoru değerlerinde belirgin bir sıçrama meydana gelmiş ve her iki metrik de yaklaşık %98 seviyesine ulaşmıştır. Veri miktarı 10000 ve 56000 örneğe çıkarıldığında performans bir miktar daha artmış olsa da elde edilen kazanımların sınırlı kaldığı görülmektedir. Bu durum, modelin yaklaşık 5000 eğitim örneğinden sonra veri verimliliği açısından doyum noktasına yaklaşmaya başladığını göstermektedir.

Eğitim ve test kayıp değerleri veri miktarının artmasıyla birlikte düzenli olarak azalmış, buna paralel olarak eğitim ve test doğrulukları arasındaki genelleme farkı büyük veri boyutlarında %1'in oldukça altında kalmıştır. Özellikle 5000 ve üzeri örneklerde genelleme farkının 0.003–0.006 aralığında gerçekleşmesi, modelin aşırı öğrenme eğiliminin oldukça düşük olduğunu ve eğitim verisinden bağımsız olarak test verisi üzerinde başarılı bir genelleme gerçekleştirdiğini ortaya koymaktadır. Ayrıca büyük veri boyutlarında standart sapmanın oldukça düşük seviyelere inmesi, elde edilen sonuçların farklı rastgelelik tohumları altında tutarlı ve tekrarlanabilir olduğunu göstermektedir.

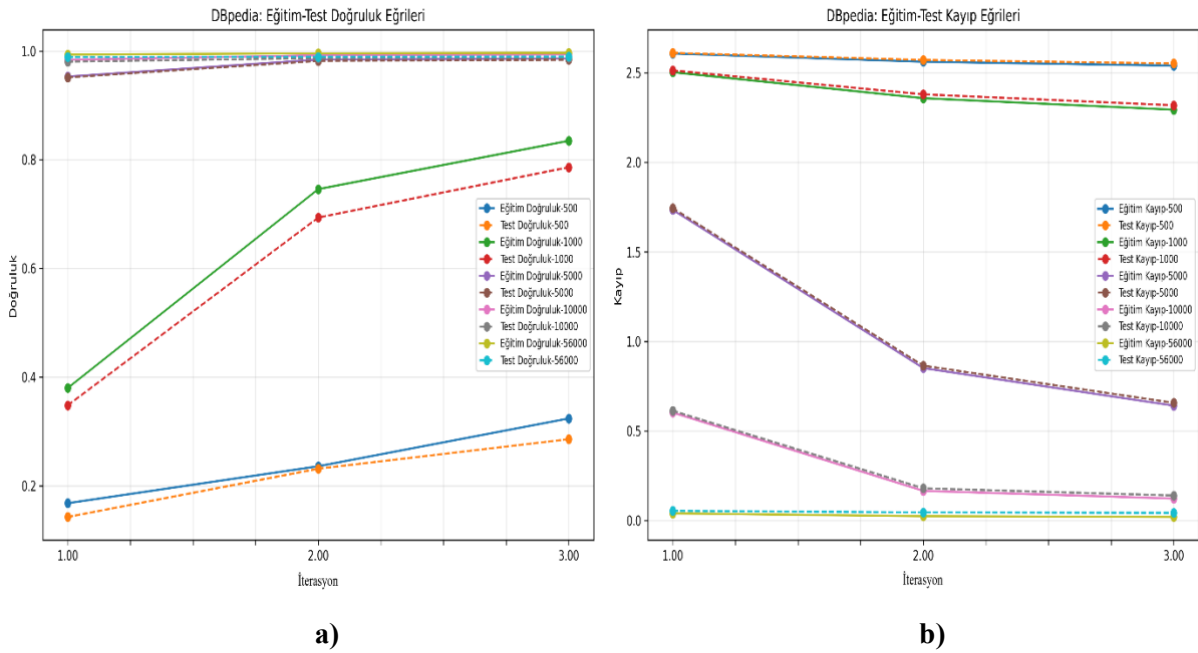
Tablo 4. DBpedia veri setine ait sonuçlar

Veri Boyutu	Eğitim Doğruluk (Ort±SS)	Test Doğruluk (Ort±SS)	Macro F1 (Ort±SS)	Eğitim Kaybı (Ort±SS)	Test Kaybı (Ort±SS)	Genelleme Farkı (Ort±SS)
500	0.3608 ±0.0850	0.3234 ±0.0883	0.2852 ±0.0973	2.5328 ±0.0120	2.5458 ±0.0135	0.0374 ±0.0180
1000	0.6968 ±0.1155	0.6486 ±0.1242	0.6195 ±0.1342	2.3374 ±0.0397	2.3601 ±0.0351	0.0482 ±0.0147
5000	0.9877 ±0.0026	0.9844 ±0.0009	0.9844 ±0.0009	0.6360 ±0.0432	0.6494 ±0.0424	0.0034 ±0.0020
10000	0.9939 ±0.0013	0.9876 ±0.0006	0.9876 ±0.0006	0.1202 ±0.0046	0.1381 ±0.0035	0.0063 ±0.0012
56000	0.9965 ±0.0002	0.9901 ±0.0008	0.9902 ±0.0008	0.0202 ±0.0013	0.0426 ±0.0016	0.0063 ±0.0009

Şekil 5.a incelendiğinde, tüm veri boyutlarında eğitim süresince hem eğitim hem de test doğruluklarının düzenli olarak arttığı görülmektedir. Düşük veri rejimlerinde (500 ve 1000 örnek) modelin ilk iterasyonlarda öğrenme hızı yüksek olmakla birlikte doğruluk değerleri sınırlı kalmıştır.

Veri miktarı 5000 örnek ve üzerine çıkarıldığında ise eğitim ve test doğruluklarında belirgin bir artış gerçekleşmiş, özellikle 10000 ve 56000 örnek kullanıldığında her iki doğruluk değeri de %99 seviyesine yaklaşmıştır. Eğitim ve test doğruluk eğrilerinin büyük veri boyutlarında birbirine oldukça yakın seyretmesi, modelin yüksek genelleme başarısına sahip olduğunu ve aşırı öğrenme eğiliminin oldukça düşük seviyede kaldığını göstermektedir.

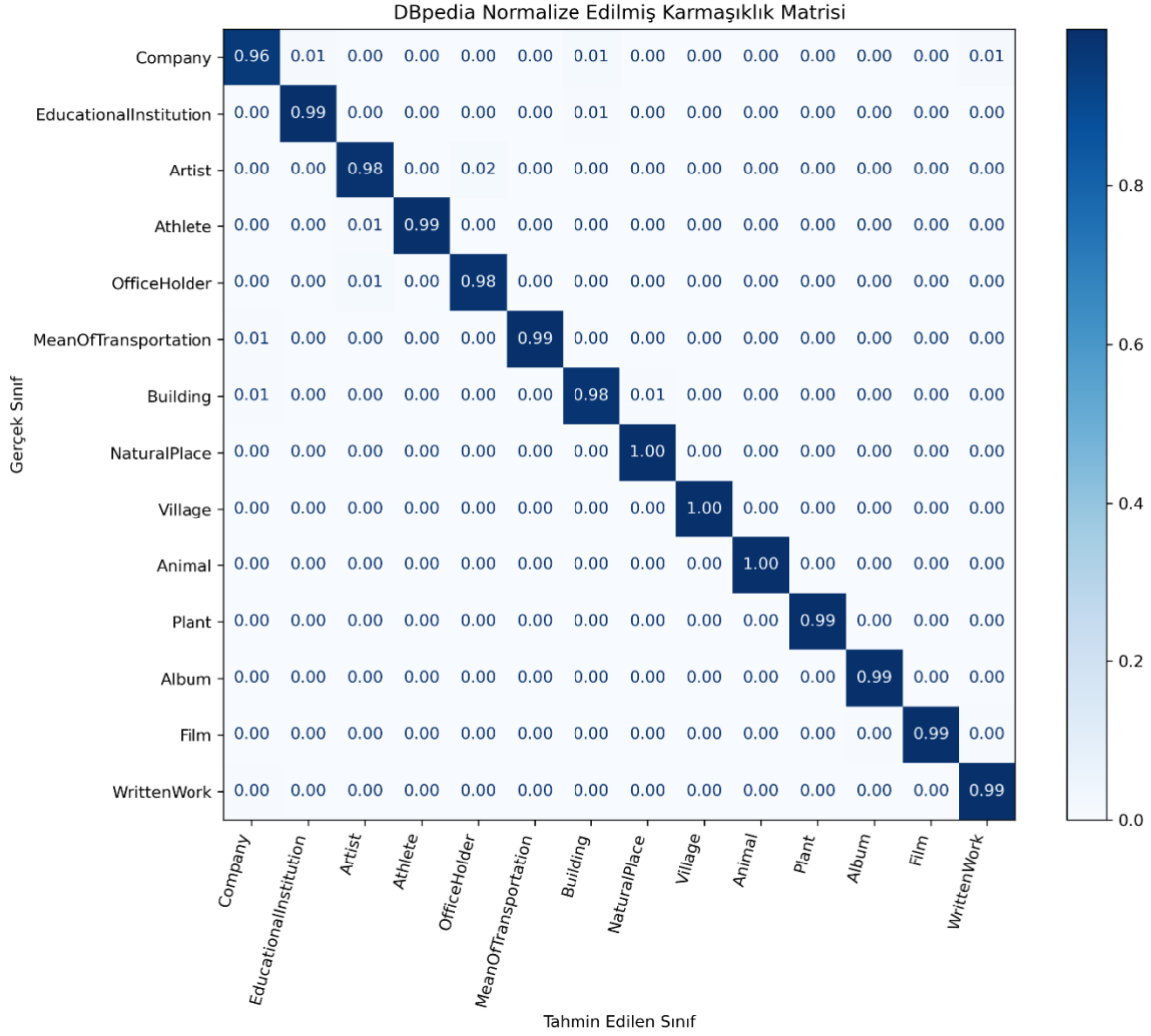
Şekil 5.b'de eğitim kaybının tüm veri boyutlarında iterasyon ilerledikçe düzenli olarak azaldığı görülmektedir. En belirgin kayıp azalması 5000 ve 10000 örnekli deneylerde ilk iki iterasyon içerisinde gerçekleşmiş 56000 örnek kullanıldığında ise başlangıçtan itibaren oldukça düşük kayıp değerlerine ulaşarak eğitim süreci kararlı bir şekilde devam etmiştir. Test kaybı da eğitim kaybını yakından takip etmiş ve veri miktarı arttıkça belirgin biçimde azalmıştır. Özellikle 56000 örnek kullanılan deneyde eğitim ve test kayıplarının oldukça düşük ve birbirine yakın seyretmesi, modelin hem etkili bir şekilde öğrenebildiğini hem de yüksek genelleme performansı sergilediğini ortaya koymaktadır.



Şekil 5. DBpedia veri setine ait doğruluk ve kayıp grafikleri

Şekil 6'da verilen normalize edilmiş karmaşıklık matrisi, modelin 14 sınıfın tamamını oldukça yüksek doğrulukla sınıflandırabildiğini göstermektedir. Doğru sınıflandırma oranları çoğu sınıfta %98–100 aralığında gerçekleşmiş olup özellikle NaturalPlace, Village ve Animal sınıflarında hata yapılmadığı görülmektedir. Yanlış sınıflandırmaların büyük bölümü anlamsal olarak birbirine yakın sınıflar arasında gerçekleşmektedir. Örneğin Artist ile OfficeHolder, Company ile Building ve Building ile NaturalPlace sınıfları arasında düşük oranlı karışıklıklar oluşmuştur. Bununla birlikte diyagonal üzerindeki yüksek değerler ve diyagonal dışındaki çok düşük hata oranları, DistilBERT modelinin DBpedia veri kümesindeki

çok sınıflı metin sınıflandırma görevinde oldukça başarılı olduğunu ve sınıflar arasındaki anlamsal farklılıkları etkili biçimde öğrendiğini göstermektedir.



Şekil 6. DBpedia veri setine ait karmaşıklık matrisi

Tablo 5 incelendiğinde, eğitim veri miktarının artırılmasının model performansını tüm değerlendirme ölçütleri açısından istikrarlı biçimde iyileştirdiği görülmektedir. Eğitim doğruluğu 500 örnekte %80.48 iken 3920 örnekten oluşan eğitim kümesinde %96.38'e yükselmiştir. Benzer şekilde test doğruluğu da %72.35'ten %89.12'ye ulaşarak modelin daha fazla eğitim verisiyle daha başarılı genelleme yaptığını göstermektedir. Macro F1 skorunun 0.7148'den 0.8908'e yükselmesi, doğruluk artışının tüm sınıflara dengeli şekilde yansıdığını ve modelin sınıflar arasında taraflı bir öğrenme gerçekleştirmediğini ortaya koymaktadır.

Eğitim ve test kayıp değerleri veri miktarının artmasıyla birlikte düzenli olarak azalmıştır. Eğitim kaybı 0.6915'ten 0.0790'a, test kaybı ise 1.2580'den 0.3753'e gerileyerek modelin daha kararlı ve güvenilir tahminler ürettiğini göstermektedir. Ayrıca genelleme farkının düşük seviyelerde korunması, veri miktarındaki artışın modelin genelleme performansını desteklediğini ve aşırı

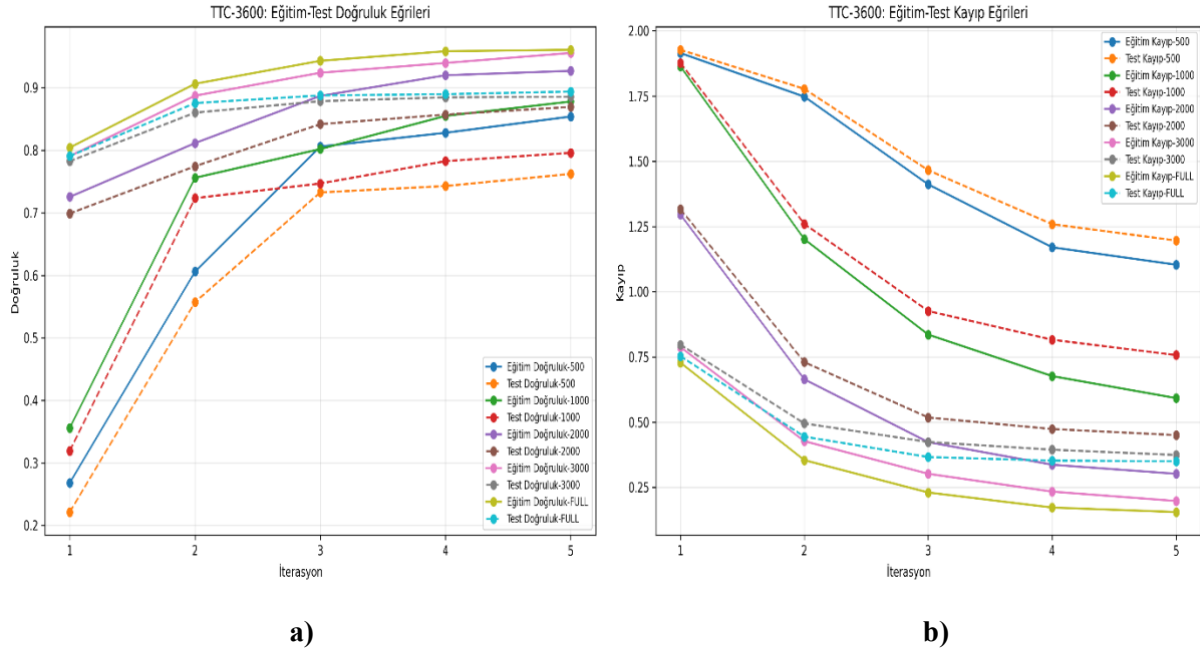
öğrenme eğilimini sınırlı tuttuğunu göstermektedir. Bununla birlikte, tüm deneylerin beş farklı rastgele başlangıç (seed) ile gerçekleştirilmiş olması ve standart sapma değerlerinin oldukça düşük bulunması, elde edilen sonuçların rastlantısal başlangıçlardan önemli ölçüde etkilenmediğini ve modelin farklı çalıştırmalarda tutarlı sonuçlar ürettiğini göstermektedir. Genel olarak değerlendirildiğinde, eğitim veri miktarındaki artışın hem sınıflandırma başarısını hem de modelin genelleme kabiliyetini belirgin biçimde iyileştirdiği söylenebilir.

Tablo 5. TTC-4900 Sonuçlar

Veri Boyutu	Eğitim Doğruluk (Ort±SS)	Test Doğruluk (Ort±SS)	Macro F1 (Ort±SS)	Eğitim Kaybı (Ort±SS)	Test Kaybı (Ort±SS)	Genelleme Farkı (Ort±SS)
500	0.8048 ± 0.0324	0.7235 ± 0.0374	0.7148 ±0.0419	0.6915 ±0.0603	1.2580 ±0.0996	0.0813 ±0.0257
1000	0.8824 ±0.0099	0.8045 ± 0.0108	0.8022 ±0.0111	0.4238 ±0.0238	0.7189 ±0.0271	0.0779 ±0.0109
2000	0.9390 ±0.0080	0.8633 ± 0.0043	0.8627 ±0.0044	0.2090 ±0.0174	0.4733 ±0.0155	0.0757 ±0.0114
3000	0.9556 ±0.0025	0.8816 ± 0.0057	0.8813 ±0.0058	0.1202 ±0.0086	0.4053 ±0.0190	0.0740 ±0.0070
3920	0.9638 ±0.0023	0.8912 ± 0.0049	0.8908 ±0.0050	0.0790 ±0.0058	0.3753 ±0.0172	0.0726 ±0.0046

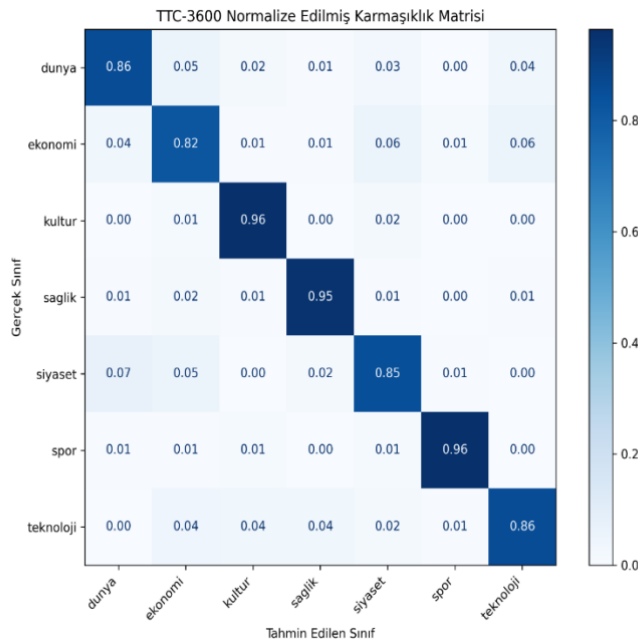
Şekil 7.a incelendiğinde, tüm veri boyutlarında eğitim süresince hem eğitim hem de test doğruluklarının düzenli olarak arttığı görülmektedir. Özellikle 500 ve 1000 örnekten oluşan düşük veri rejimlerinde ilk iki iterasyonda belirgin doğruluk artışları elde edilmiş, sonraki iterasyonlarda ise öğrenme hızının kademeli olarak azaldığı gözlenmiştir. Eğitim veri miktarı arttıkça model daha yüksek doğruluk değerlerine ulaşmış; 3920 örnekten oluşan en büyük eğitim örneklemini kullanıldığında eğitim doğruluğu yaklaşık %96, test doğruluğu ise %89 seviyesine ulaşmıştır. Büyük veri boyutlarında eğitim ve test doğruluk eğrilerinin birbirine yakın seyretmesi, modelin daha başarılı genelleme yaptığını ve öğrenme sürecinin kararlı olduğunu göstermektedir.

Şekil 7.b'de tüm veri boyutlarında eğitim kaybının iterasyon ilerledikçe düzenli olarak azaldığı görülmektedir. En belirgin kayıp azalması özellikle 1000 ve üzerindeki veri boyutlarında ilk üç iterasyon içerisinde gerçekleşmiştir. Test kaybı da benzer şekilde azalmış ve büyük veri kümelerinde eğitim kaybını yakından takip etmiştir. En büyük eğitim örnekleminde hem eğitim hem de test kayıplarının düşük seviyelere ulaşması, modelin daha kararlı öğrenme gerçekleştirdiğini ve aşırı öğrenme eğiliminin sınırlı kaldığını göstermektedir.



Şekil 7. TTC-4900 veri setine ait doğruluk ve kayıp grafikleri.

Şekil 8'de verilen normalize edilmiş karmaşıklık matrisi, modelin sınıfların büyük çoğunluğunu yüksek doğrulukla ayırt edebildiğini göstermektedir. Kültür (%96), sağlık (%95) ve spor (%96) kategorileri en yüksek doğru sınıflandırma oranlarına sahip sınıflar olurken, dünya (%86), siyaset (%85), teknoloji (%86) ve ekonomi (%82) sınıflarında diğer kategorilere göre daha fazla karışıklık meydana gelmiştir. Yanlış sınıflandırmaların büyük bölümü içerik bakımından birbirine yakın olan dünya–ekonomi, ekonomi–siyaset ve siyaset–dünya kategorileri arasında gerçekleşmiştir. Bu durum söz konusu haber kategorilerinin ortak kavramlar ve benzer sözcük dağılımlarına sahip olmasından kaynaklanmaktadır. Buna karşın diyagonal üzerindeki yüksek değerler ve diyagonal dışındaki düşük hata oranları, modelin TTC-4900 veri kümesinde genel olarak başarılı ve dengeli bir sınıflandırma performansı sergilediğini göstermektedir.



Şekil 8. TTC-4900 veri setine ait karmaşıklık matrisi

Tablo 6 incelendiğinde, eğitim veri miktarının artırılmasının model performansını tüm değerlendirme ölçütleri açısından olumlu yönde etkilediği görülmektedir. Eğitim doğruluğu 500 örnek için ortalama 93.24 ± 2.81 iken tam veri setinde (11021 örnek) 99.73 ± 0.02 seviyesine ulaşmıştır. Benzer şekilde test doğruluğu ve Macro F1 skoru değerleri de veri miktarı arttıkça sürekli yükselmiş, modelin daha fazla eğitim verisi ile daha başarılı ve dengeli bir sınıflandırma gerçekleştirdiği gözlenmiştir.

Eğitim ve test kayıp değerleri veri boyutunun artmasıyla birlikte düzenli olarak azalmış, bu durum modelin parametrelerini daha etkin öğrenebildiğini ve tahmin güvenilirliğinin arttığını göstermektedir. Özellikle tam veri setinde elde edilen düşük kayıp değerleri, modelin eğitim sürecinin kararlı biçimde yakınsadığını ortaya koymaktadır.

Genelleme farkı incelendiğinde ise 500 örnekte ortalama 0.0265 ± 0.0075 olan farkın tam veri setinde 0.0151 ± 0.0022 'ye düştüğü görülmektedir. Eğitim ve test performansları arasındaki farkın azalması, veri miktarındaki artışın aşırı öğrenme eğilimini sınırlandırdığını ve modelin daha güçlü bir genelleme yeteneği kazandığını göstermektedir.

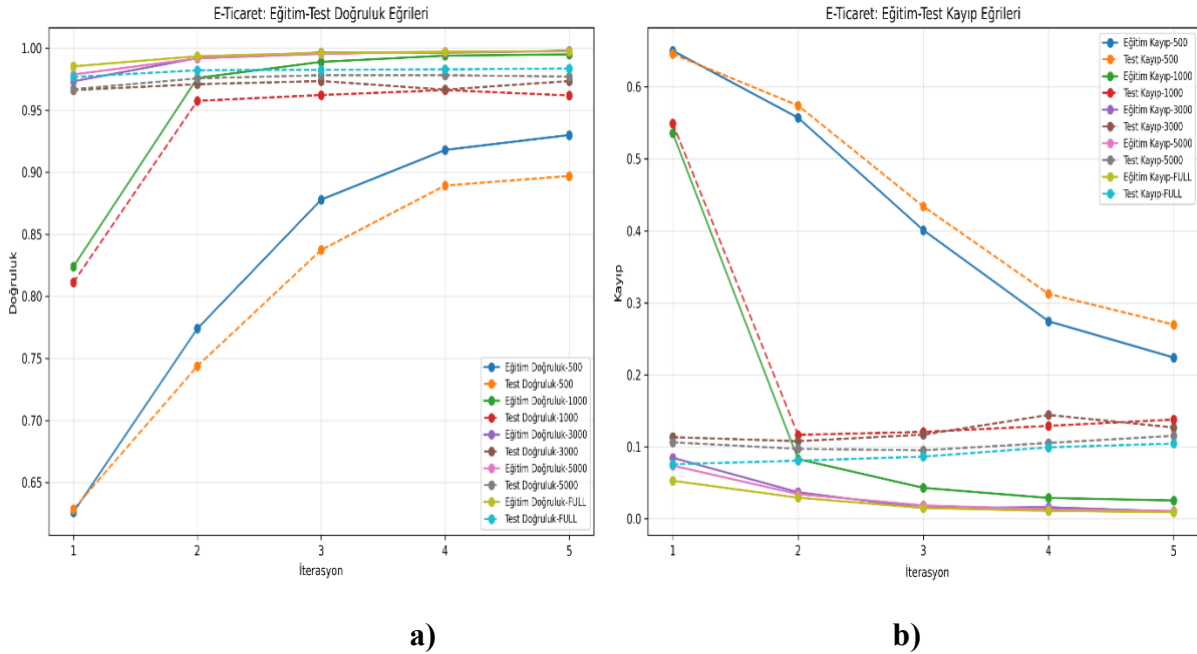
Ayrıca beş farklı rastgele başlangıç (seed) üzerinde elde edilen standart sapma değerlerinin oldukça düşük olması, model performansının başlangıç koşullarından büyük ölçüde etkilenmediğini ve deneylerin yüksek tekrarlanabilirlik ile kararlılık sergilediğini ortaya koymaktadır. Genel olarak sonuçlar, Türkçe e-ticaret ürün yorumlarının duygu sınıflandırılmasında veri miktarının artırılmasının hem doğruluk hem de Macro F1 skoru başarısını artırırken genelleme performansını da iyileştirdiğini göstermektedir.

Tablo6. E-ticaret Sonuçlar

Veri Boyutu	Eğitim Doğruluk (Ort±SS)	Test Doğruluk (Ort±SS)	Macro F1 (Ort±SS)	Eğitim Kaybı (Ort±SS)	Test Kaybı (Ort±SS)	Genelleme Farkı (Ort±SS)
500	0.9324 ±0.0281	0.9059 ± 0.0276	0.9054 ± 0.0278	0.1872 ±0.0720	0.2333 ±0.0617	0.0265 ±0.0075
1000	0.9956 ±0.0019	0.9661 ± 0.0042	0.9661 ± 0.0042	0.0218 ± 0.0104	0.1225 ±0.0144	0.0295 ±0.0035
3000	0.9967 ±0.0010	0.9742 ± 0.0025	0.9742 ± 0.0025	0.0145 ± 0.0055	0.1127 ±0.0148	0.0225 ±0.0030
5000	0.9966 ±0.0007	0.9779 ± 0.0023	0.9779 ± 0.0023	0.0148 ± 0.0038	0.1041 ±0.0118	0.0187 ±0.0028
11021	0.9973 ±0.0002	0.9822 ± 0.0021	0.9822 ± 0.0021	0.0110 ± 0.0013	0.0956 ±0.0123	0.0151 ±0.0022

Şekil 9.a incelendiğinde, tüm veri boyutlarında eğitim süreci boyunca hem eğitim hem de test doğruluklarının düzenli olarak arttığı görülmektedir. Özellikle 500 ve 1000 örnekten oluşan düşük veri rejimlerinde ilk iki iterasyonda doğrulukta belirgin bir artış gerçekleşmiş, sonraki iterasyonlarda ise artış hızının azaldığı gözlenmiştir. Veri miktarının artmasıyla birlikte model daha yüksek doğruluk seviyelerine ulaşmış; 3000 örnek ve üzerindeki deneylerde eğitim doğruluğu %99 seviyesine yaklaşırken test doğruluğu %97–98 bandında kararlı bir yapı sergilemiştir. Tam veri seti kullanıldığında eğitim doğruluğu yaklaşık %99.8'e ulaşırken test doğruluğunun %98 civarında sabitlenmesi, modelin yüksek sınıflandırma başarısını koruduğunu ve eğitim ile test performansı arasındaki farkın oldukça düşük kaldığını göstermektedir.

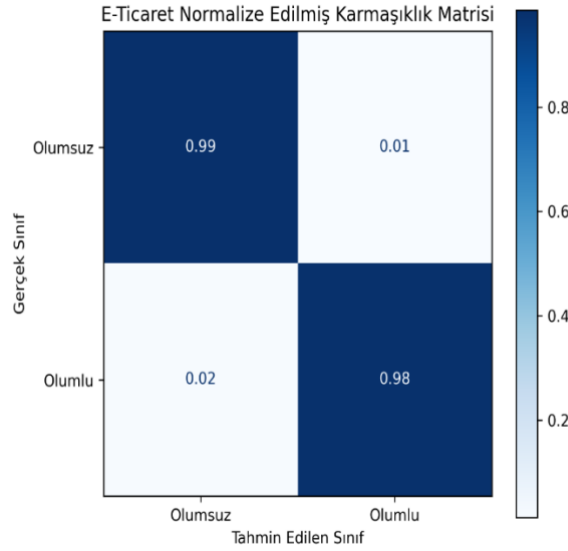
Şekil 9.b incelendiğinde, tüm veri boyutlarında eğitim kaybının iterasyon ilerledikçe düzenli biçimde azaldığı görülmektedir. En belirgin kayıp azalması özellikle 500 ve 1000 örnekle deneylerde ilk üç iterasyon içerisinde gerçekleşmiş, daha büyük veri kümelerinde ise eğitim kaybı kısa sürede çok düşük seviyelere inmiştir. Test kaybı da genel olarak eğitim sürecinde azalma eğilimi göstermiş ve büyük veri boyutlarında düşük değerlerde kararlı hale gelmiştir. Özellikle 3000 örnek ve üzerindeki deneylerde eğitim ve test kayıpları arasındaki fark oldukça sınırlı kalmış, bu durum modelin yalnızca eğitim verisini ezberlemediğini, aynı zamanda yeni örnekler üzerinde de başarılı biçimde genelleme yapabildiğini göstermektedir.



Şekil 9. E-ticaret veri setine ait doğruluk ve kayıp grafikleri.

Şekil 10'da verilen normalize edilmiş karışıklık matrisi, modelin E-ticaret veri setindeki duygu sınıflarını genel olarak yüksek doğrulukla ayırt edebildiğini göstermektedir. Olumsuz yorumların yaklaşık %99'u doğru sınıflandırılırken yalnızca %1'i olumlu olarak tahmin edilmiştir. Benzer şekilde

olumlu yorumların yaklaşık %98'i doğru sınıflandırılmış, yalnızca %2'lik kısmı olumsuz sınıfa atanmıştır. Yanlış sınıflandırma oranlarının her iki sınıf için de oldukça düşük olması, modelin sınıflar arasında dengeli bir karar mekanizması geliştirdiğini ve belirli bir sınıfa yönelik sistematik bir yanlılık göstermediğini ortaya koymaktadır. Elde edilen sonuçlar, DistilBERT tabanlı yaklaşımın E-ticaret ürün yorumlarının duygu analizinde yüksek doğruluk ve güçlü genelleme performansı sergilediğini doğrulamaktadır.



Şekil 10. E-ticaret veri setine ait karmaşıklık matrisi

Tüm veri setleri birlikte değerlendirildiğinde, eğitim veri boyutunun model performansı ve genelleme yeteneği üzerinde belirleyici bir etkiye sahip olduğu görülmektedir. Veri miktarının artmasıyla birlikte tüm veri kümelerinde eğitim ve test doğrulukları yükselmiş, kayıp değerleri azalmış ve genelleme farkı belirgin biçimde küçülmüştür. Bununla birlikte bu iyileşmenin düzeyi, veri setinin yapısına ve sınıflandırma görevinin zorluk derecesine bağlı olarak farklılık göstermektedir. Yapısal olarak daha belirgin sınıf ayrımlarına sahip DBpedia ve E-ticaret veri kümelerinde yüksek doğruluk ve düşük genelleme farkına kısa sürede ulaşılrken, AG News, IMDb ve TTC-4900 gibi içerik benzerliği ve anlamsal örtüşmenin daha fazla olduğu veri kümelerinde veri miktarı artsa dahi sınıflar arası karışıklık tamamen ortadan kalkmamıştır. Buna rağmen tüm deneylerde veri boyutunun artırılması modelin daha kararlı öğrenmesine, daha düşük kayıp değerleri üretmesine ve daha güçlü bir genelleme performansı sergilemesine katkı sağlamıştır. Bu bulgular, eğitim veri miktarının yalnızca nihai doğruluk değerlerini değil, aynı zamanda modelin öğrenme dinamiklerini, sınıflar arası ayrıştırma başarısını ve genelleme kapasitesini doğrudan etkileyen temel faktörlerden biri olduğunu ortaya koymaktadır.

5. Sonuçlar

Bu çalışmada, DistilBERT tabanlı bir metin sınıflandırma modelinde eğitim veri kümesi boyutunun model başarımı ve genelleme davranışı üzerindeki etkisi, farklı dil (İngilizce-Türkçe) ve görev yapılarına sahip beş ayrı veri seti üzerinden incelenmiştir. Elde edilen bulgular, eğitim verisinin miktarının yalnızca doğruluk değerlerini değil, aynı zamanda modelin öğrenme sürecinin kararlılığını ve doğrulama performansını da doğrudan etkilediğini göstermektedir.

Tüm deney setlerinde, eğitim veri miktarının artmasıyla birlikte doğrulama doğruluğunun kararlı biçimde yükseldiği, eğitim ve doğrulama kayıplarının azaldığı ve genelleme farkının belirgin biçimde küçüldüğü gözlemlenmiştir. Düşük veri boyutlarında (500–1000 örnek), modelin eğitim doğruluğunun hızla yükselmesine rağmen doğrulama performansının daha düşük seviyelerde kalması, önceden eğitilmiş DistilBERT modelinin sınırlı veri koşullarında aşırı öğrenmeye daha yatkın olduğunu göstermektedir.

AG News ve özellikle 14 sınıflı yapıya sahip DBpedia veri setlerinde, modelin yeterli veri miktarına ulaşıldığında hızla doygunluğa eriştiği, sınıflar arasındaki karışıklığın oldukça düşük seviyelere indiği ve genelleme performansını yüksek doğrulukla koruduğu gözlemlenmiştir. Buna karşın IMDb gibi uzun dizili ve öznel metinlerden oluşan duygu analizi görevlerinde, veri miktarının artmasına rağmen doğruluk artışının diğer veri kümelerine kıyasla daha sınırlı gerçekleştiği ve genelleme farkının tamamen ortadan kalkmadığı gözlemlenmiştir.

Türkçe veri setleri olan TTC-4900 ve E-ticaret veri setleri üzerinde elde edilen sonuçlar, eklemeli dil yapısına sahip ve gürültülü (kullanıcı yorumları gibi) gerçek dünya verilerinde veri boyutunun etkisinin daha kritik olduğunu ortaya koymuştur. E-ticaret veri setinde veri boyutu arttıkça doğrulama doğruluğunun %98 seviyelerine ulaşması, eğitim ve doğrulama kayıp eğrilerinin birbirine yaklaşması ve normalize edilmiş karmaşıklık matrisinde her iki sınıf için de oldukça düşük hata oranlarının elde edilmesi, önceden eğitilmiş DistilBERT modellerinin yeterli eğitim verisi sağlandığında Türkçe duygu analizi görevlerinde yüksek genelleme performansı sergilediğini göstermektedir.

Sonuç olarak, bu çalışma eğitim veri kümesi boyutunun DistilBERT tabanlı modellerin başarımı ve genelleme kapasitesi üzerinde belirleyici bir role sahip olduğunu ortaya koymaktadır. Elde edilen bulgular, daha büyük veri kümelerinin yalnızca daha yüksek doğruluk değerleri sağlamadığını, aynı zamanda daha kararlı ve dengeli bir öğrenme süreci sunduğunu göstermektedir. Bu nedenle, DistilBERT tabanlı modellerin farklı dil ve görev türlerinde etkin biçimde kullanılabilmesi için eğitim veri miktarının, model karmaşıklığı ve problem yapısı dikkate alınarak ele alınması gerektiği değerlendirilmektedir.

Yazarların Katkısı

Makalenin tamamı tek bir yazar tarafından tasarlanmış ve oluşturulmuştur.

Çıkar Çatışması Beyanı

Yazar herhangi bir çıkar çatışması olmadığını beyan etmiştir.

Araştırma ve Yayın Etiği Beyanı

Yazar, bu çalışmanın Araştırma ve Yayın Etiği kurallarına uygun olduğunu beyan eder.

Kaynaklar

- Arzu, M., Aydoğan, M. (2023). Türkçe Duygu Sınıflandırma İçin Transformers Tabanlı Mimarilerin Karşılaştırılmalı Analizi. Computer Science, IDAP-2023: International Artificial Intelligence and Data Processing Symposium(IDAP-2023), 1-6. <https://doi.org/10.53070/bbd.1350405>
- Kaplan, J., McCandlish, S., Henighan, T.J., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D. (2020). Scaling Laws for Neural Language Models. *ArXiv, abs/2001.08361*.
- Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D.D., Hendricks, L.A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., Driessche, G.V., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J.W., Vinyals, O., Sifre, L. (2022). Training Compute-Optimal Large Language Models. *ArXiv, abs/2203.15556*.
- Schwartz, R., Dodge, J., Smith, N., Etzioni, O. (2019). Green AI. 10.48550/arXiv.1907.10597.
- Eşidir, K. A. (2025). DistilBERT ve Geleneksel Makine Öğrenmesi Modelleri ile Duygu Analizi: Performans Karşılaştırması. International Journal of Pure and Applied Sciences, 11(2), 444-459. <https://doi.org/10.29132/ijpas.1604869>
- Sanh, V., Debut, L., Chaumond, J., Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv, abs/1910.01108*.
- Tang, R., Lu, Y., Liu, L., Mou, L., Vechtomova, O., Lin, J.J. (2019). Distilling Task-Specific Knowledge from BERT into Simple Neural Networks. *ArXiv, abs/1903.12136*.
- Ünal, M.C., Aygün, B., Gerek, A. (2023). Comparison of Pre-trained Language Models for Turkish Address Parsing. *ArXiv, abs/2306.13947*.
- Micikevicius, P., Narang, S., Alben, J., Diamos, G. F., Elsen, E., García, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., Wu, H. (2017). *Mixed precision training*. arXiv preprint arXiv:1710.03740.
- Çabuk, M.(2022). <https://www.kaggle.com/datasets/mujdatcabuk/eticaret-urun-yorumlari>. E.T.: 18.01.2026
- Zhang, X., Zhao, J.J., LeCun, Y. (2015). Character-level Convolutional Networks for Text Classification. *Neural Information Processing Systems*.
- Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., Potts, C. (2011, June). Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (pp. 142-150).
- Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Bizer, C. (2015). DBpedia: A large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 6(2), 167-195.
- Kılınç, D., Özçift, A., Bozyigit, F., Yıldırım, P., Yücalar, F., Borandag, E. (2015). TTC-3600: A new benchmark dataset for Turkish text categorization. *Journal of Information Science*, 43(2), 174-185. <https://doi.org/10.1177/0165551515620551>