LLM ile Veri Arttırma

Müşerref Özkan
Veri Bilimi ve Büyük Veri
235B7053

***Özet*— Bu çalışma, metin sınıflandırması için büyük dil modelleri ile eğitim ve test kümeleri üzerinde veri arttırma işleminin uygulanması ve etkilerinin değerlendirilmesi amacıyla gerçekleştirilmiştir. google/flan-t5-large, tuner007/pegasus_paraphrase, eugenesiow/bart-paraphrase ve eugenesiow/bart-paraphrase modelleri uygulanarak veri arttırma işlemi gerçekleştirilmiştir. Eğitim kümesi için veriler iki, üç ve beş katına çıkarılarak başarıları değerlendirilmiştir. Eğitim kümesi veriler için üç ve beş katına çıkarılarak tahmin aşamasında kararlar oylama yöntemi ile birleştirilmiştir. Sınıflandırma işlemi XGBoost ve Destek Vektör Makinesi algoritmaları ile gerçekleştirilmiştir.**

I. GİRİŞ

Bu çalışma, metin sınıflandırması probleminde büyük dil modellerinin (LLM) eğitim ve test verilerindeki veri arttırma yöntemleri ile birlikte kullanımını incelemek amacıyla gerçekleştirilmiştir. Farklı veri arttırma yöntemlerinin etkilerini gözlemlemek için hem eğitim hem de test kümelerine çeşitli LLM modelleri uygulanmıştır. Çalışmada kullanılan veri setleri, kullanıcı yorumları ve puanlarından oluşan iki farklı kategoride ele alınmıştır: gıda ürünleri ve otel rezervasyon yorumları.

Veri arttırma yöntemleriyle eğitilen modellerin sınıflandırma performansları, XGBoost ve Destek Vektör Makinesi (SVM) algoritmaları kullanılarak değerlendirilmiştir. Farklı LLM'ler aracılığıyla üretilen metinler ile veri setlerinin çeşitli oranlarda genişletilmesi, sınıflandırma başarılarının artırılması hedeflenmiştir. Performans değerlendirmesi sürecinde Accuracy ve F1-Score metrikleri kullanılmış ve sonuçlar detaylı bir şekilde görselleştirilmiştir. Çalışmanın kodu burada bulunmaktadır.

II. YÖNTEM

A. Veri Kümesi

Yapılan uygulamanın ilk veri kümesi, gıda ürünleri için kullanıcı yorumlarını içermektedir ve "Text" ve "Score" özelliklerinden oluşmaktadır. "Score" değişkeni hedef değişken olup beş sınıftan (0,1,2,3,4) oluşmaktadır. Veri kümesi, 100 adet veriden oluşmaktadır ve sınıf dağılımı dengelidir. Veri kümesi %80 eğitim, %20 test kümesi olmak üzere ikiye ayrılmıştır.

İkinci veri kümesi ise otel rezervasyonları yorumları ve skorları bilgilerini bulunmaktadır. Veri kümesi, "Review" ve "Rating" özelliklerinden oluşmaktadır. Hedef değişken beş sınıftan (0,1,2,3,4) oluşmaktadır. Veri kümesi, 120 adet veriden oluşmaktadır ve sınıf dağılımı dengelidir. Veri kümesi %80 eğitim, %20 test kümesi olmak üzere ikiye ayrılmıştır.

B. Temsil Yöntemleri ve Model

Temsil yöntemi olarak all-MiniLM-L12-v2 modeli kullanılmıştır. Bu model, yüksek doğruluk ve düşük hesaplama maliyeti sağlayarak metinlerin anlamını kompakt ve etkili bir şekilde temsil etmek için tercih edilmiştir. Model, kullanıcı girdilerinin anlamını derinlemesine kavrayarak benzerlik ve bağlam açısından zengin bir vektör uzayında temsil eder.

Ürün yorumlarının veri eğitimi sürecinde, yüksek doğruluk ve verimli işlem kapasitesi sunan XGBoost algoritması kullanılmıştır. XGBoost, gradyan artırmalı ağaç tabanlı bir makine öğrenmesi algoritması olup, büyük veri setlerinde hızlı ve etkili modelleme sağlamak amacıyla tercih edilmiştir. Modelin hiperparametre ayarlamaları, optimum performans elde etmek için dikkatlice gerçekleştirilmiştir.

Otel yorumlarının veri eğitim sürecinde Destek Vektör Makinesi(SVM) algoritması kullanılmıştır. SVM modeli, küçük ve orta ölçekli veri setlerinde sınıflandırma performansını artırmak için tercih edilmiştir. SVM, veriyi yüksek boyutlu bir özellik uzayına projelendirerek doğrusal olarak ayrılabilir hale getirir ve optimal bir ayırma hiper düzlemi oluşturur.

C. Veri Arttırma Yöntemleri

Proje kapsamında hem eğitim hem de test kümesini arttırmak için dört adet LLM ile veri arttırma yöntemi uygulanmıştır. Ürün yorumları verisi için tüm arttırma yöntemleri kullanılırken otel yorumları veri kümesi için tuner007/pegasus_paraphrase yöntemi uygulanmamıştır. Kullanılan yöntemler aşağıda listelenmiştir.

google/flan-t5-large modeli, metin tabanlı görevlerde genel amaçlı dil anlama ve üretimi için kullanılan bir T5 (Text-to-Text Transfer Transformer) modelidir. Geniş çapta ince ayar yapılmış olması sayesinde, metinlerin anlamını koruyarak farklı ifadelerle yeniden yazılmasına olanak tanır.

tuner007/pegasus_paraphrase, modeli, özellikle metin özetleme amacıyla tasarlanmış bir transformer modelidir. Ancak paraphrasing (yeniden ifade etme) görevinde de etkili sonuçlar sağlamaktadır. Model, belirli bir metni daha sade veya farklı bir yapıyla yeniden oluşturarak veri çeşitliliğini artırmak için kullanılmıştır. Modelin kapasitesi, eğitim verisinden öğrenilen geniş dil bağlamlarına dayanmaktadır.

eugenesiow/bart-paraphrase, modeli, hem metin oluşturma hem de metin iyileştirme görevlerinde kullanılan bir encoder-decoder yapısına sahiptir. Bu model, metinlerin anlam bütünlüğünü koruyarak daha akıcı ve çeşitli biçimlerde yeniden yazılmasını sağlar. Augmentation sürecinde, BART modeli kullanılarak veri çeşitliliği artırılmış ve modelin daha genel ve genelleştirilebilir hale gelmesi sağlanmıştır.

Gemini 1.5 flash modeli, Google tarafından geliştirilen gelişmiş bir büyük dil modeli (LLM) olup, doğal dil anlama ve üretimi konusunda yüksek performans sunmaktadır. Bu proje kapsamında, modelin API üzerinden kullanımı sağlanmış ve metinlerin anlamını koruyarak farklı varyasyonlarının üretilmesi için kullanılmıştır. API kullanımı sayesinde ölçeklenebilirlik ve hız avantajı elde edilmiştir.

Yukarıdaki yöntemler kullanılarak test kümesi 3 ve 5 katına çıkarılmış, eğitim kümesi ise 2,3 ve 5 katına çıkarılmıştır.

D. Performans Değerlendirme

Eğitim ve test kümelerinde yapılan veri arttırmanın etkisini ölçebilmek için her iki küme içinde Accuracy, F1-Score, Confusion Matrix çıktıları elde edilerek değerlendirme yapılmıştır. Her iki veri kümesi içinde sınıflar dengeli olduğundan ve çıktılarda F1 Score sonuçları ve değişimleride tutarlı olduğundan sonuç görselleştirme aşamasında yalnızca Accuracy metriği kullanışmıştır.

Tahminlerin birleştirilmesi oylama yöntemiyle gerçekleştirilmiştir.

III. Deneysel Sonuçlar

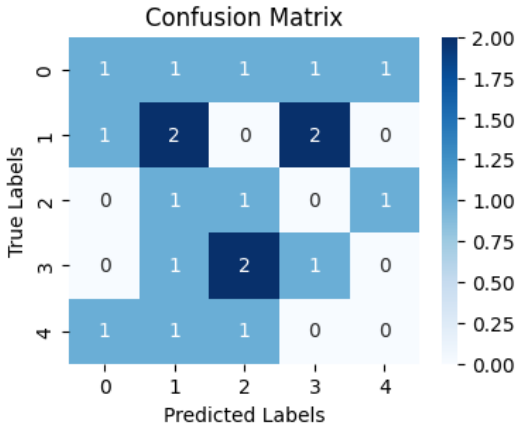
Eğitim ve test kümesi için uygulanan veri arttırma yöntemleri çıktıları ve model başarı metrikleri üzerinden değerlendirme yapılmıştır.

Yukarıda belirtilen modeller için eğitim kümelerinin orijinal metinleri için bir kez metin üreterek veri iki katına, iki kez üreterek üç katına, dört kez üreterek beş katına çıkarılmıştır. Veri üretim aşamasında en iyi çıktıyı alabilmek için modellerin farklı parametre varyasyonları denenmiş ve çeşitli promptlar ile iyileştirmeler yapılmaya çalışılmıştır. Çoğunlukla aşağıdaki iki prompt ifade kulalnılmıştır.

“Generate four diverse paraphrases for the following text: ”

“Paraphrase the following text while maintaining its original meaning: ”

Deneysel sonuçların paylaşılmasına ürün yorumları veri kümesiyle başlanacaktır. Kümenin orijinal metin ile eğitilip test edilmesi sonucunda modelin accuracy değeri 0.25’dir.



Şekil 1. Orijinal Eğitim Kümesi İçin Karmaşıklık Matrisi

Aşağıda örnek orijinal metin ve google/flan-t5-large modeli ile üretilen metinlerin örnekleri verilmiştir.

Original: i love this tea for dieting its the best i will always purchase this tea for me and my family

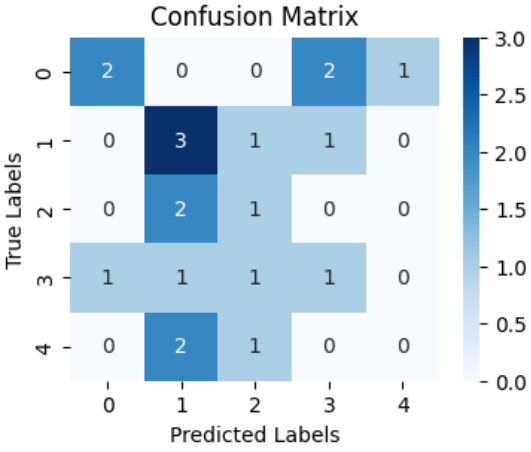
Augmented 1: i am still reviewing this product and will continue to do so

Augmented 2: i love this tea for dieting its the best i will always purchase this tea for me and my family

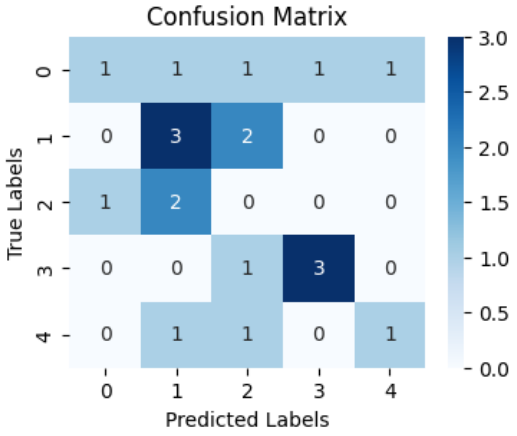
Augmented 3: i love this tea for dieting its the best i will always purchase this tea for me and my family

Augmented 4: it's nice tea for dieting it's the best i will always purchase this tea for me and my family

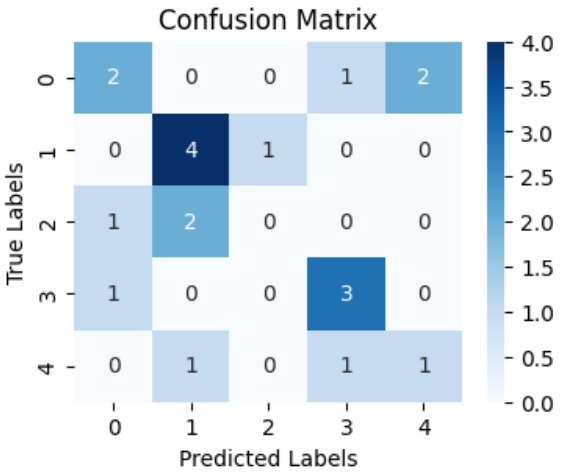
Üretilen metinler ile eğitim kümesi önce iki katına, sonra üç katına ve en son 5 katına çıkarılarak her bir arttırılmış veri kümesi için sonuçlar test edilmiştir. google/flan-t5-large modeli ile veri kümesi iki katına çıkarıldığında accuracy 0.35, üç katına çıkarıldığında 0.40 ve beş katına çıkarıldığında 0.50 olarak elde edilmiştir.



Şekil 2. İki Katına Çıkarılmış Eğitim Kümesi İçin Karmaşıklık Matrisi



Şekil 3. Üç Katına Çıkarılmış Eğitim Kümesi İçin Karmaşıklık Matrisi

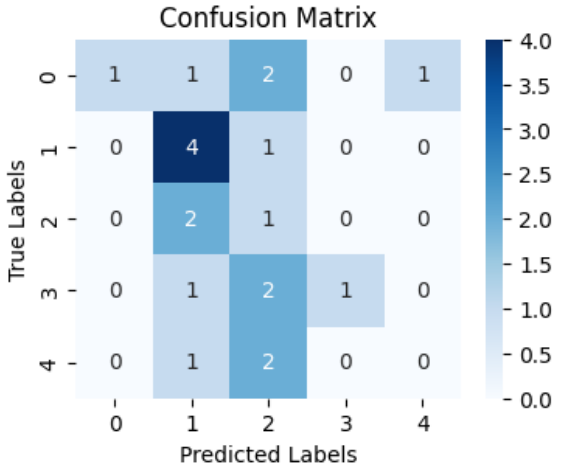


Şekil 4. Beş Katına Çıkarılmış Eğitim Kümesi İçin Karmaşıklık Matrisi

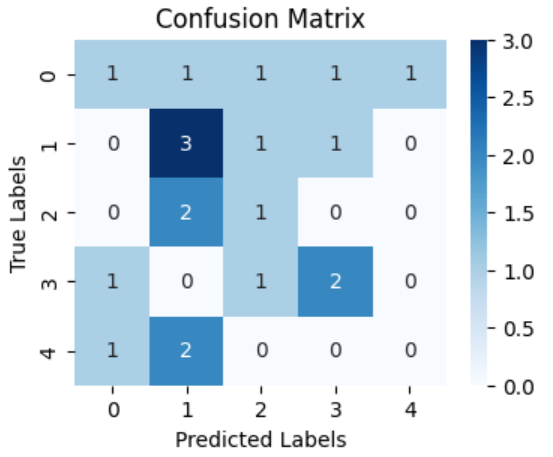
Orijinal ve tuner007/pegasus_paraphrase modeli çıktıları aşağıda verilmiştir.

Original: i love this tea for dieting its the best i will always purchase this tea for me and my family
Augmented 1: I love this tea, and will always purchase it for me and my family.
Augmented 2: I love this tea for dieting and will always buy it for me and my family.
Augmented 3: I love this tea, and will always purchase it for myself and my family.
Augmented 4: I love this tea, and will always purchase it for my family.

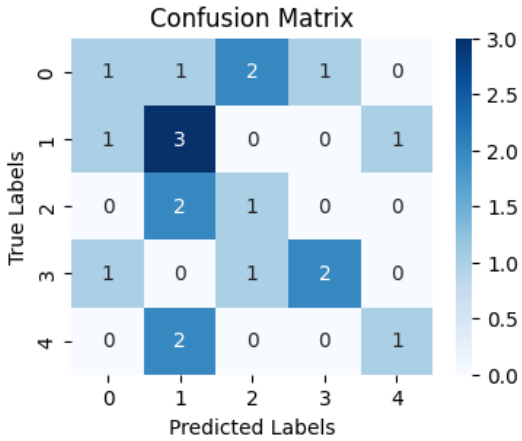
Model ile veri kümesi iki katına çıkarıldığında accuracy 0.35, üç katına çıkarıldığında 0.35 ve beş katına çıkarıldığında 0.40 olarak elde edilmiştir.



Şekil 5. İki Katına Çıkarılmış Eğitim Kümesi İçin Karmaşıklık Matrisi



Şekil 6. Üç Katına Çıkarılmış Eğitim Kümesi İçin Karmaşıklık Matrisi

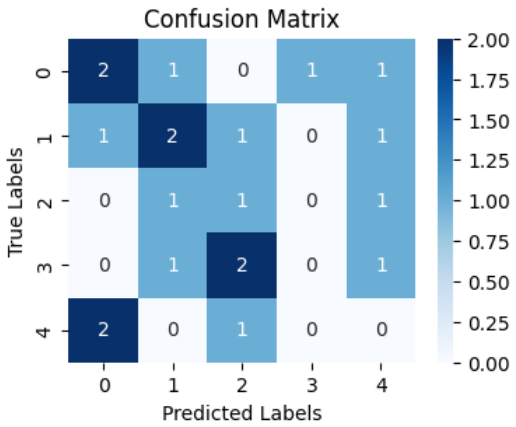


Şekil 7. Beş Katına Çıkarılmış Eğitim Kümesi İçin Karmaşıklık Matrisi

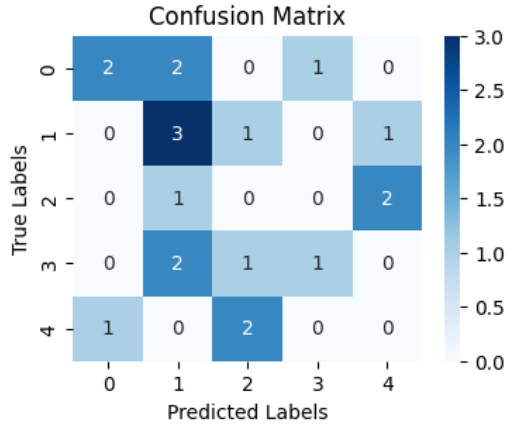
Orijinal ve eugenesiow/bart-paraphrase modeli çıktıları aşağıda verilmiştir.

Original: i love this tea for dieting its the best i will always purchase this tea for me and my family
Augmented 1: i love this tea for dieting its the best i will always purchase this tea in India for me and my family
Augmented 2: i love this tea for dieting its the best i will always purchase this tea in India for me and my family
Augmented 3: i love this tea for dieting its the best i will always purchase this tea in India for me and family
Augmented 4: i love this tea for dieting its the best i will always purchase this tea. for me and my family

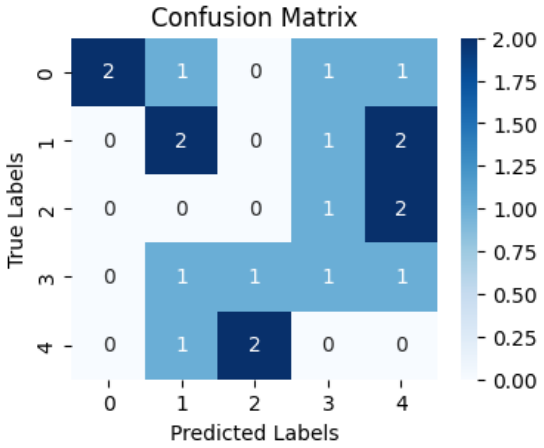
Model ile veri kümesi iki katına çıkarıldığında accuracy 0.30, üç katına çıkarıldığında 0.40 ve beş katına çıkarıldığında 0.35 olarak elde edilmiştir.



Şekil 8. İki Katına Çıkarılmış Eğitim Kümesi İçin Karmaşıklık Matrisi



Şekil 9. Üç Katına Çıkarılmış Eğitim Kümesi İçin Karmaşıklık Matrisi



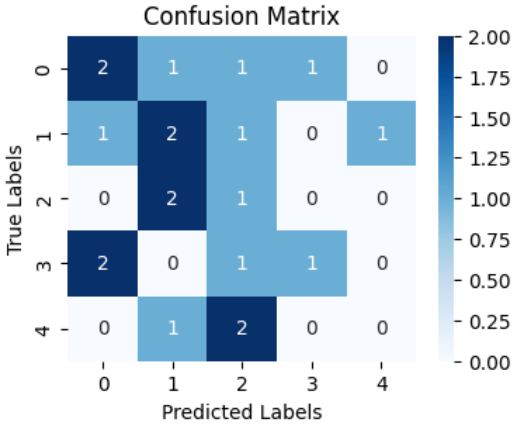
Şekil 10. Beş Katına Çıkarılmış Eğitim Kümesi İçin Karmaşıklık Matrisi

Orijinal ve gemini 1.5 flash modeli çıktıları aşağıda verilmiştir.

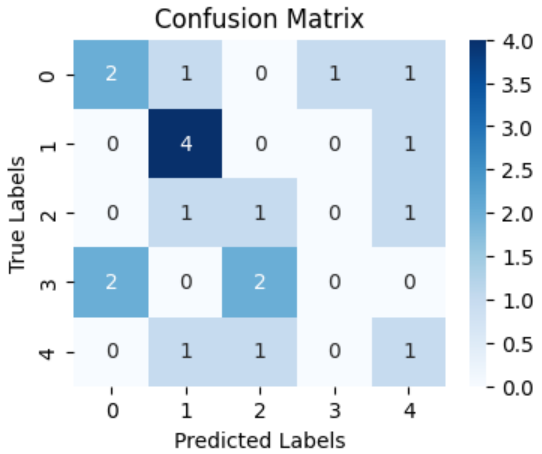
Orignal: i love this tea for dieting its the best i will always purchase this tea for me and my family

Augmented 1: This tea's the best for dieting; I'll always buy it for my family and me.
Augmented 2: I love this tea for dieting—the best! My family and I will always buy it.
Augmented 3: For dieting, this tea is unbeatable. It's a staple for my family and me.
Augmented 4: My family and I will always buy this tea; it's the best for dieting

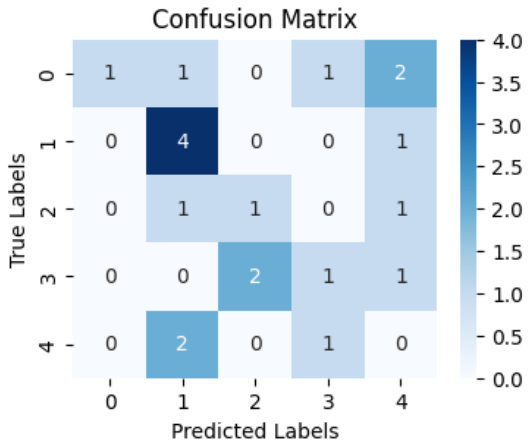
Model ile veri kümesi iki katına çıkarıldığında accuracy 0.30, üç katına çıkarıldığında 0.40 ve beş katına çıkarıldığında 0.35 olarak elde edilmiştir.



Şekil 11. İki Katına Çıkarılmış Eğitim Kümesi İçin Karmaşıklık Matrisi

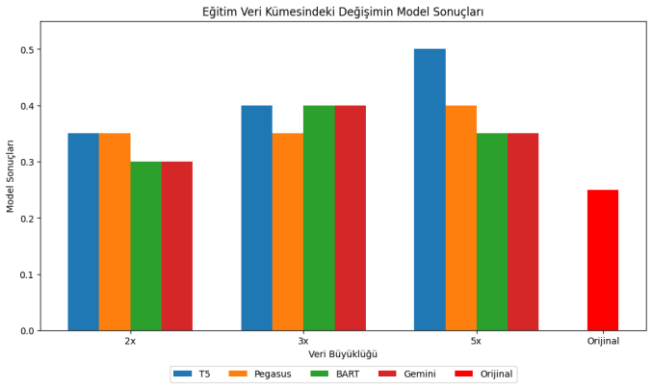


Şekil 12. Üç Katına Çıkarılmış Eğitim Kümesi İçin Karmaşıklık Matrisi



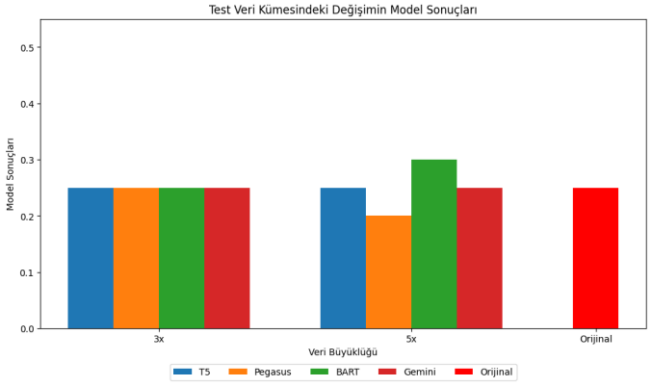
Şekil 13. Beş Katına Çıkarılmış Eğitim Kümesi İçin Karmaşıklık Matrisi

Tüm modeller için arttılışmış kümelerin sonuçları grafik olarak aşağıda verilmiştir.



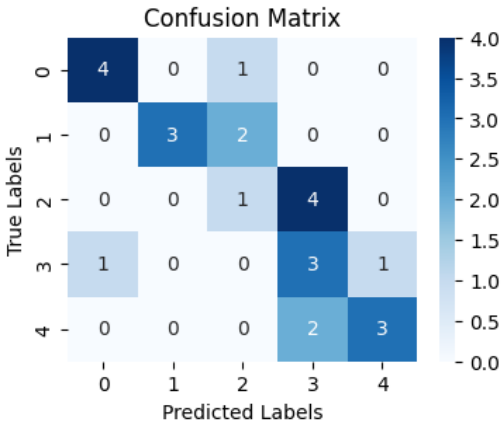
Şekil 14. Arttırılmış ve Orijinal Kümenin Eğitim Sonuçları

Proje kapsamında test kümesi üç ve beş kat arttırılmış ve tahmin aşamasında orijinal veri ile kararlar birleştirlmiştir. Uygulama kapsamında kararlar oylama yöntemiyle birleştirilmiştir. Bu uygulama için en yüksek accuracy sonucunu eugenesiow/bart-paraphrase modeli 0.30 olarak vermiştir.



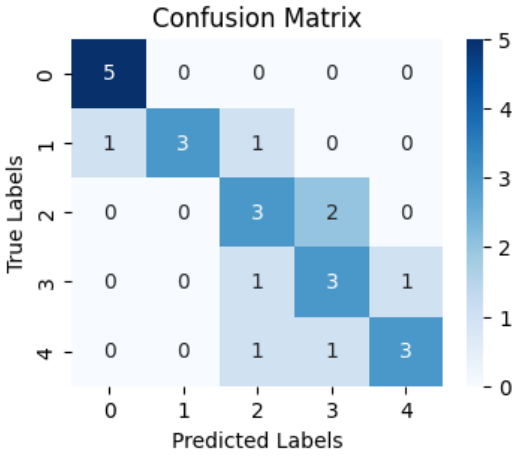
Şekil 15. Arttırılmış ve Orijinal Kümenin Test Kümesi İçin Birleştirilen Karar Sonuçları

Projenin ikinci veri kümesi için aynı işlemler tekrarlamıştır. Modelin orijinal başarısı 0.56 olarak elde edilmiştir.

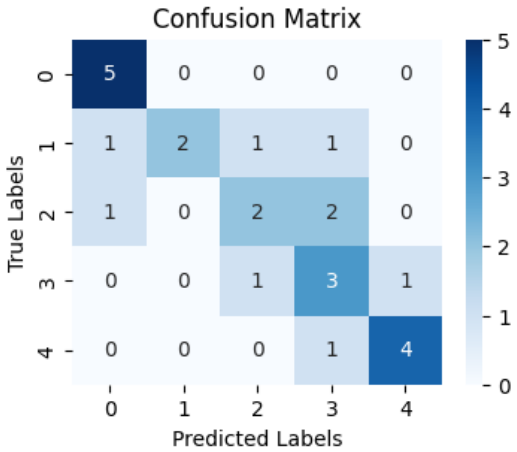


Şekil 16. Orijinal Eğitim Kümesi İçin Karmaşıklık Matrisi

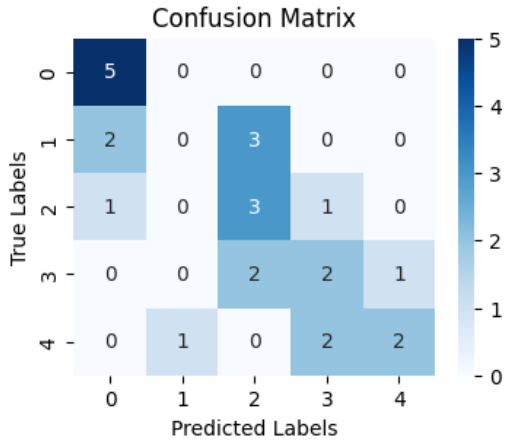
google/flan-t5-large modeli ile veri kümesi iki katına çıkarıldığında accuracy 0.68, üç katına çıkarıldığında 0.64 ve beş katına çıkarıldığında 0.48 olarak elde edilmiştir.



Şekil 17. İki Katına Çıkarılmış Eğitim Kümesi İçin Karmaşıklık Matrisi

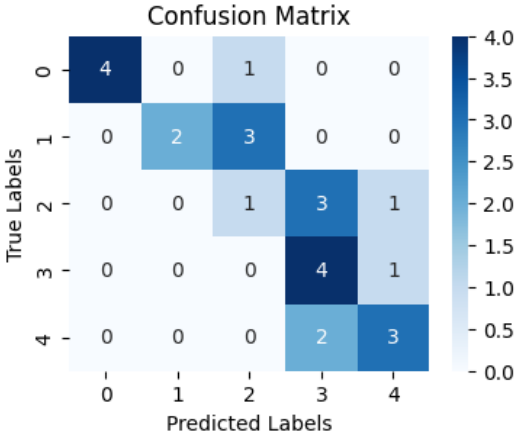


Şekil 18. Üç Katına Çıkarılmış Eğitim Kümesi İçin Karmaşıklık Matrisi

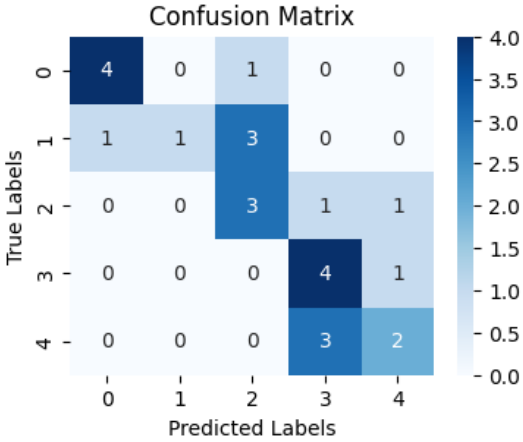


Şekil 19. Beş Katına Çıkarılmış Eğitim Kümesi İçin Karmaşıklık Matrisi

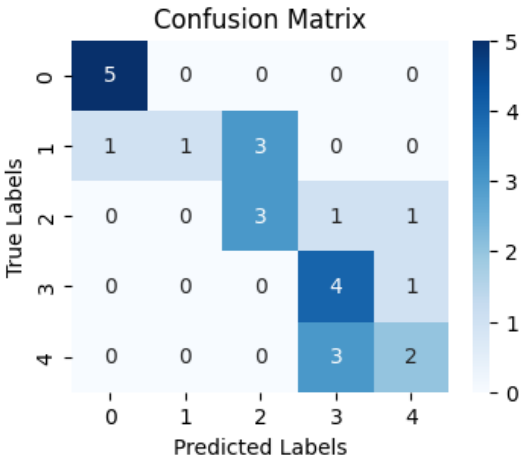
eugenesiow/bart-paraphrase modeli ile veri kümesi iki katına çıkarıldığında accuracy 0.56, üç katına çıkarıldığında 0.56 ve beş katına çıkarıldığında 0.60 olarak elde edilmiştir.



Şekil 20. İki Katına Çıkarılmış Eğitim Kümesi İçin Karmaşıklık Matrisi

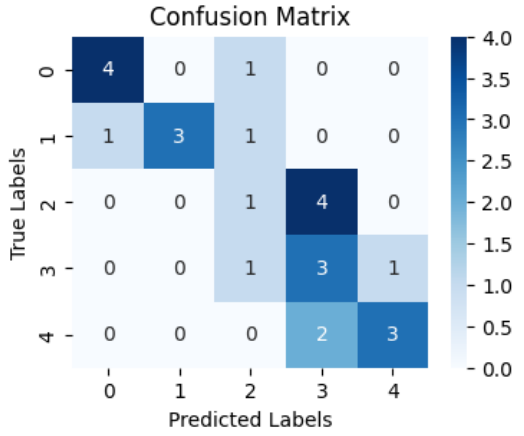


Şekil 21. Üç Katına Çıkarılmış Eğitim Kümesi İçin Karmaşıklık Matrisi

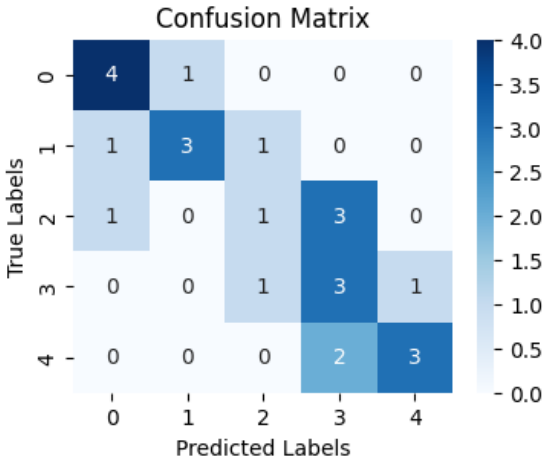


Şekil 22. Beş Katına Çıkarılmış Eğitim Kümesi İçin Karmaşıklık Matrisi

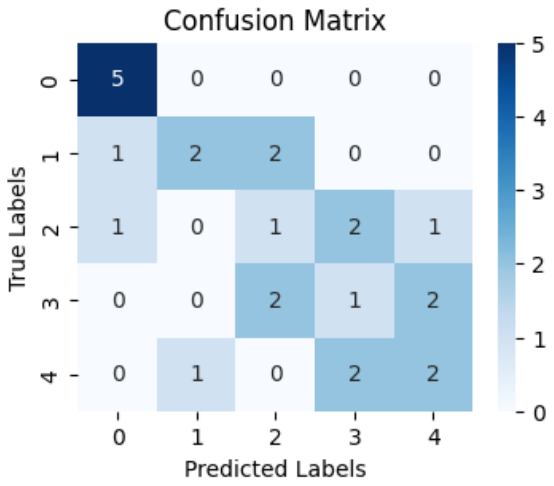
gemini 1.5 flash modeli ile veri kümesi iki katına çıkarıldığında accuracy 0.56, üç katına çıkarıldığında 0.56 ve beş katına çıkarıldığında 0.44 olarak elde edilmiştir.



Şekil 23. İki Katına Çıkarılmış Eğitim Kümesi İçin Karmaşıklık Matrisi

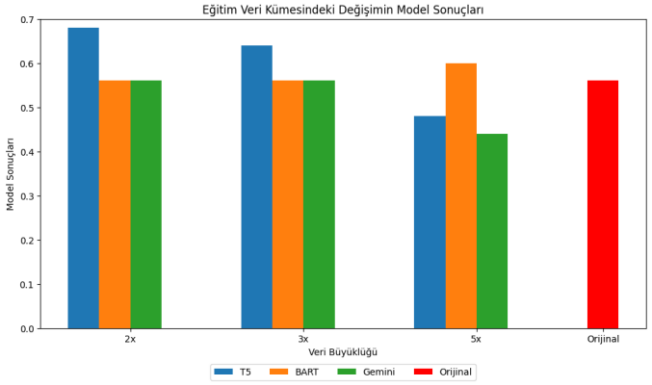


Şekil 23. Üç Katına Çıkarılmış Eğitim Kümesi İçin Karmaşıklık Matrisi



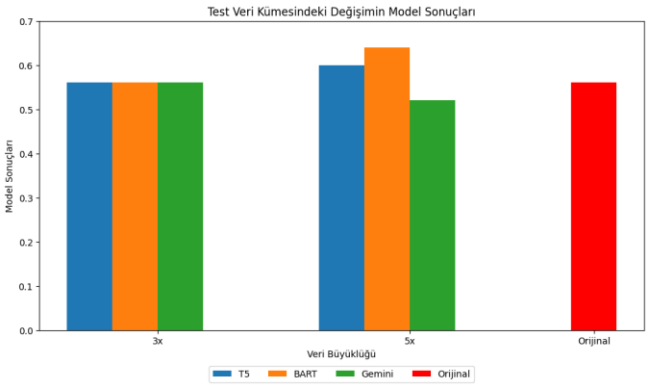
Şekil 24. Beş Katına Çıkarılmış Eğitim Kümesi İçin Karmaşıklık Matrisi

Tüm modeller için arttılışmış kümelerin sonuçları grafik olarak aşağıda verilmiştir.



Şekil 25. Arttırılmış ve Orijinal Kümenin Eğitim Sonuçları

Test kümesi için tahmin aşamasında orijinal veri ile üretilen verilerin kararları birleştirilmiştir. Kararlar oylama yöntemiyle birleştirilmiştir. Bu yöntem için en yüksek accuracy sonucunu eugenesiow/bart-paraphrase modeli kümenin beş karına çıkarılmasıyla 0.64 olarak vermiştir.



Şekil 25. Arttırılmış ve Orijinal Kümenin Test Kümesi İçin Birleştirilen Karar Sonuçları

IV. Sonuçlar

Ürün yorumları kümesi için uygulanan model ve yöntemler için eğitim kümesi için artışında en yüksek sonucu 0.50 ile google/flan-t5-large modeli vermiştir. Test kümesinin artışında ise 0.30 ile eugenesiow/bart-paraphrase modeli vermiştir. Sonuçta uygulamada orijinal kümenin 0.25 başarısı 0.50 olarak arttılmıştır.

Otel yorumları veri kümesi için en yüksek sonuçları eğitim kümesi için google/flan-t5-large modeli 0.68 olarak verirken test kümesi için eugenesiow/bart-paraphrase modeli 0.64 olarak vermiştir.

References

[1] Y. Li , K. Ding, J. Wang, K. Lee, “Empowering Large Language Models for Textual Data Augmentation”

[2] A. G. Chowdhury, A. Chadha, ” Generative Data Augmentation using LLMs improves Distributional Robustness in Question Answering”