



**TÜBİTAK-2209-A ÜNİVERSİTE ÖĞRENCİLERİ ARAŞTIRMA  
PROJELERİ DESTEĞİ PROGRAMI**

2024 Yılı

I. Dönem Başvurusu

**2209/A ÜNİVERSİTE ÖĞRENCİLERİ ARAŞTIRMA PROJELERİ DESTEĞİ PROGRAMI**  
**ARAŞTIRMA ÖNERİSİ FORMU**

**A. GENEL BİLGİLER**

<b>Başvuru Sahibinin Adı Soyadı:</b> Abdülsamet TÜRKMENOĞLU
<b>Araştırma Önerisinin Başlığı:</b> Türkçe Hukuk Metinleri analiz etmek için küçük dil modelleri(SMLs) ve Retrieval-Augmented Generation (RAG) teknolojisini entegre eden yenilikçi bir sistem geliştirme
<b>Danışmanın Adı Soyadı:</b> Dr. Öğr. Üyesi Tolga BERBER
<b>Araştırmanın Yürütüleceği Kurum/Kuruluş:</b> Karadeniz Teknik Üniversitesi

**ÖZET**

Türkçe özetin araştırma önerisinin (a) özgün değeri, (b) yöntemi, (c) yönetimi ve (d) yaygın etkisi hakkında bilgileri kapsamı beklenir. Bu bölümün en son yazılması önerilir.

**Özet**

Bu araştırma önerisi, Türkiye Cumhuriyeti'nin hukuki metinlerini (kanunlar, yönetmelikler, kararnameler) analiz etmek için küçük dil modelleri (SMLs) ve Retrieval-Augmented Generation (RAG) teknolojisini entegre eden yenilikçi bir sistem geliştirmeyi amaçlamaktadır.

(a) Özgün Değer: Projenin özgün değeri, büyük dil modelleri yerine küçük dil modellerini RAG mimarisiyle birleştirerek hesaplama verimliliğini sağlamak, bağlamsal doğruluk açısından üstün bir sistem geliştirmektir. **Bu yaklaşım, Türkçe hukuki metinlerin analizinde mevcut yöntemlerin ötesinde performans ve verimlilik sunma potansiyeline sahiptir.**

(b) Yöntem: Araştırma yöntemi dört ana aşamadan oluşmaktadır: (1) Veri toplama ve işleme: Türkiye'nin resmi mevzuat portalından alınan hukuki metinler toplanacak ve bu metinler, anlamlarını sayısal olarak temsil eden **gömme (embedding)** vektörlerine dönüştürülecektir. (2) Model geliştirme: İki ayrı küçük dil modeli (Retriever ve Generator) Türk hukuk diline uygun olacak şekilde özel olarak eğitilecektir. (3) RAG sistemi entegrasyonu: Eğitilen modeller RAG mimarisi içinde birleştirilecektir. (4) Performans değerlendirme: Sistemin doğruluğu, yanıt kalitesi ve hızı test edilecek, geleneksel büyük dil modelleriyle karşılaştırılacaktır.

(c) Yönetim: Proje, sistematik iş paketleri halinde yürütülecektir. Veri toplama, model eğitimi, sistem entegrasyonu ve optimizasyon aşamaları, belirlenen zaman çizelgesine göre yönetilecektir. Her aşama, proje hedeflerine uygunluk açısından düzenli olarak değerlendirilecektir.

(d) Yaygın Etki: Başarıyla tamamlandığında, bu proje Türk hukuk sistemine özgü, verimli ve doğru bir metin analiz sistemi sunacaktır. Bu sistem, hukuk profesyonellerine ve araştırmacılara destek sağlayacak, hukuki araştırma ve belge sınıflandırma süreçlerini hızlandıracaktır. **Ayrıca, geliştirilen metodoloji diğer özel alan metinlerinin (tıp, finans vb.) işlenmesine de uyarlanabilir, böylece NLP alanında geniş bir etki yaratabilir.**

**Bu proje, hukuk teknolojileri (LegalTech) alanında yeni bir paradigma sunarak, hukuki metin işleme ve analiz yöntemlerinde önemli bir ilerleme sağlama potansiyeline sahiptir.**

**Anahtar Kelimeler:** Türkçe hukuk, doğal dil işleme, küçük dil modelleri, bağlamsal analiz, hukuk metinleri.

**1. ÖZGÜN DEĞER**

**1.1. Konunun Önemi, Araştırma Önerisinin Özgün Değeri ve Araştırma Sorusu/Hipotezi**

Araştırma önerisinde ele alınan konunun kapsamı ve sınırları ile önemi literatürün eleştirel bir değerlendirmesinin yanı sıra nitel veya nicel verilerle açıklanır.

Özgün değer yazılırken araştırma önerisinin bilimsel değeri, farklılığı ve yeniliği, hangi eksikliği nasıl gidereceği veya hangi soruna nasıl bir çözüm geliştireceği ve/veya ilgili bilim veya teknoloji alan(lar)ına kavramsal, kuramsal ve/veya metodolojik olarak ne gibi özgün katkılarda bulunacağı literatüre atıf yapılarak açıklanır.

Önerilen çalışmanın araştırma sorusu ve varsa hipotezi veya ele aldığı problem(ler)i açık bir şekilde ortaya konulur.

**Bu araştırma önerisi, Türkiye Cumhuriyeti'nin hukuki metinlerini (kanunlar, yönetmelikler, kararnameler) küçük dil modelleri (Small Language Models - SMLs) ve Retrieval-Augmented Generation (RAG) teknolojisini kullanarak analiz eden yenilikçi bir sistem geliştirmeyi amaçlamaktadır. Hukuk dili, kendine has terminolojisi ve karmaşık yapısı nedeniyle doğal dil işleme (NLP) alanında önemli zorluklar sunmaktadır.**

## 2209/A ÜNİVERSİTE ÖĞRENCİLERİ ARAŞTIRMA PROJELERİ DESTEĞİ PROGRAMI ARAŞTIRMA ÖNERİSİ FORMU

Mevcut büyük dil modelleri, hukuki metinlerin analizinde genellikle yetersiz kalmakta ve aşırı kaynak tüketimi nedeniyle pratik uygulamalarda sınırlı kullanım alanı bulmaktadır .

Önerilen çalışmanın özgün değeri, büyük dil modelleri yerine küçük dil modellerini (SML) RAG mimarisiyle entegre ederek, hem hesaplama verimliliği hem de bağlamsal doğruluk açısından üstün bir sistem geliştirmeye odaklanmaktadır. Literatürde bu konuya yeterince eğilinmemiş olması, çalışmanın yenilikçi yönünü ortaya koymaktadır. Küçük dil modellerinin RAG ile entegrasyonu, çeşitli avantajlar sunmaktadır. Öncelikle, **SML'ler büyük modellere kıyasla çok daha az işlemci gücü ve bellek gerektirdiği için hesaplama verimliliği sağlar.** Ayrıca, bu modellerin daha küçük ve modüler yapıları, **Türk hukuk diline özgü uyarlamaları daha kolay hale getirir ve özelleştirme esnekliği sunar.** RAG mimarisinin sağladığı esneklik sayesinde bilgi tabanı, **yeni hukuki metinlerle hızlı bir şekilde güncellenebilir, böylece sistem her zaman güncel ve doğru bilgiyle çalışır.** Dahası, **RAG'ın yanıtların kaynağını izlenebilir kılması, hukuki analizlerde hayati bir öneme sahip olan kaynak gösterme işlevini de etkin bir şekilde destekler.** Bu özelliklerin birleşimi, çalışmayı hem teknik hem de pratik açıdan güçlü bir konuma getirmektedir.

Bu sistem, hukuki metin analizi, sınıflandırma ve bilgi çıkarımı gibi alanlarda mevcut yöntemlerin ötesinde bir performans ve verimlilik sunma potansiyeline sahiptir. Özellikle, Türkçe gibi kaynakları sınırlı dillerde hukuki NLP çalışmalarına önemli bir katkı sağlayacaktır .

Araştırma Sorusu: "Türkçe hukuki metinlerin doğru ve bağlamsal analizini gerçekleştirebilen, hesaplama açısından verimli bir sistem, küçük dil modellerinin RAG mimarisiyle entegrasyonu yoluyla nasıl geliştirilebilir?"

Hipotez: Küçük dil modellerinin RAG mimarisiyle entegrasyonu, Türkçe hukuki metinlerin analizinde, büyük dil modellerine kıyasla daha düşük hesaplama maliyetiyle eşdeğer veya daha yüksek doğruluk ve bağlamsal anlama performansı sağlayacaktır.

**Bu araştırma, hukuk teknolojileri (LegalTech) alanında yeni bir paradigma sunarak, hukuki metin işleme ve analiz yöntemlerinde önemli bir ilerleme sağlama potansiyeline sahiptir. Ayrıca, bu yaklaşım diğer özel alan metinlerinin (tıp, finans vb.) işlenmesine de uyarlanabilir, böylece NLP alanında geniş bir etki yaratabilir.**

### 1.2. Amaç ve Hedefler

Araştırma önerisinin amacı ve hedefleri açık, ölçülebilir, gerçekçi ve araştırma süresince ulaşılabilir nitelikte olacak şekilde yazılır.

**Bu araştırma önerisinin temel amacı,** Türkiye Cumhuriyeti'nin hukuki metinlerini (kanunlar, yönetmelikler, kararnameler) analiz etmek ve sınıflandırmak için küçük dil modelleri (Small Language Models - SMLs) ile Retrieval-Augmented Generation (RAG) teknolojisini entegre eden yenilikçi ve verimli bir sistem geliştirmektir. Bu amaç doğrultusunda, aşağıdaki spesifik, ölçülebilir, gerçekçi ve ulaşılabilir hedefler tanımlanmıştır. İlk olarak, veri toplama ve işleme aşamasında, Türkiye'nin resmi mevzuat portalından (mevzuat.gov.tr) en az **X (BİR SAYI YAZCAZ)** hukuki metin (kanun, yönetmelik, kararname) toplamak ve bu metinleri dijital formatta (JSON, XML veya düz metin) işlenebilir hale getirmektir. Toplanan metinleri, RAG mimarisi için uygun vektör temsillerine dönüştürecek ve indekslenecektir. Veri setini, eğitim (%X), doğrulama (%X) ve test (%X) olarak bölümlenecektir.

**Küçük dil modeli geliştirme aşamasında,** Türkçe hukuk diline özel en az iki farklı küçük dil modeli (Retriever ve Generator) geliştirmek. Bu modellerin parametreleri geleneksel büyük dil modellerinin en fazla %10'u kadar olacaktır. Retriever modelini, hukuki sorguları anlamak ve ilgili metinleri bilgi tabanından çekmek üzere optimize etmek olacaktır. Bu modelin, sorgu-döküman eşleştirme görevinde en az **%X** doğruluk oranına ulaşmasını sağlamak amaçlanmaktadır. Generator modelini, çekilen bilgileri kullanarak tutarlı ve doğru hukuki yanıtlar üretmek üzere eğitmektir. Bu modelin, üretilen yanıtlarda en az **%X** BLEU(YA DA FARKLI BİR BENCKMARK) skoru elde etmesini sağlayabilmektir.

**RAG sistemi entegrasyonu kapsamında,** Geliştirilen Retriever ve Generator modellerini, verimli bir RAG mimarisi içinde entegre edilecektir. Sistemin uçtan uca **latency'sini X milisaniyenin** altında tutulacaktır. RAG sisteminin bellek kullanımını, eşdeğer büyük dil modellerinin **en fazla %X'si ile sınırlanacaktır.**

**Uygulama geliştirme ve entegrasyon aşamasında,** Geliştirilen sistemin doğruluğunu, hız ve kaynak kullanımını kapsamlı testlerle değerlendirilecek. **En az 1000 farklı hukuki** sorgu üzerinde test yapılacaktır. Sistemin performansını, en az üç farklı büyük dil modeli (örneğin, BERT, GPT-3, LEGAL-BERT) ile karşılaştırılacaktır.

## 2209/A ÜNİVERSİTE ÖĞRENCİLERİ ARAŞTIRMA PROJELERİ DESTEĞİ PROGRAMI ARAŞTIRMA ÖNERİSİ FORMU

Hukuk profesyonellerinden (en az X avukat ve X hâkim(İNŞALLAH)) oluşan bir panel ile sistemin ürettiği yanıtların kalitesini değerlendirmek ve en az %X kullanıcı memnuniyeti skoru elde edilecektir. Sistem performansını artırmak için en az üç farklı optimizasyon tekniği (kuantizasyon, bilgi damıtma, model budama) uygulanacak ve her birinin etkisi ölçülecektir. Geliştirilen sistemi, kullanıcı dostu bir arayüz ile web tabanlı bir uygulamaya dönüştürülecektir. Sistemi, en az bir hukuk bürosu veya akademik kurumda pilot uygulama olarak entegre edilecek ve gerçek dünya performansını değerlendirilecektir. Sistemin API'sini geliştirerek, diğer hukuki yazılımlarla entegrasyon imkânı sunulacaktır. Araştırma sonuçlarını en az X uluslararası hakemli dergide yayınlanacaktır. Geliştirilen sistemi ve sonuçları en az bir uluslararası konferansta sunulacaktır. Projenin açık kaynak kodunu ve eğitilmiş modellerini, araştırma topluluğunun kullanımına sunulacaktır.

**Etik ve yasal uyum açısından,** geliştirilen sistem, kullanılan hukuki metinlerin yalnızca içerik analizi amacıyla işlendiğini ve herhangi bir şekilde dikte edilmediğini garanti altına alacaktır. Metinlerde yer alan isimler ve kişisel veriler gibi hassas bilgiler gizlilik ilkesine uygun şekilde maskelenecek veya anonim hale getirilecektir. Ayrıca, sistemin Türkiye'de yürürlükte olan Kişisel Verilerin Korunması Kanunu'na (KVKK) tam uyumluluğu sağlanacaktır. Bu kapsamda, tüm veri işleme süreçleri şeffaf bir şekilde yürütülecektir.

Bu hedefler, projenin 7 aylık süre içerisinde aşamalı olarak gerçekleştirilecektir. Her bir hedefin ilerleme durumu, belirli metrikler ve dönüm noktaları (kilometre taşları) ile düzenli olarak takip edilecektir. Kilometre taşları, projenin kritik aşamalarında ulaşılan önemli adımları ifade eder ve sürecin planlanan takvime uygun ilerleyip ilerlemediğini değerlendirmek için kullanılır. **Bu yaklaşım, projenin sistematik ilerlemesini ve nihai amaca ulaşmasını sağlayacaktır. Ayrıca, bu hedefler doğrultusunda geliştirilen sistem, Türk hukuk sisteminde metin analizi ve bilgi erişimi alanlarında önemli bir teknolojik ilerleme sağlayacak ve benzer sistemlerin diğer diller ve hukuk sistemleri için de geliştirilmesine öncülük edecektir.**

## 2. YÖNTEM

Araştırma önerisinde uygulanacak yöntem ve araştırma teknikleri (veri toplama araçları ve analiz yöntemleri dahil) ilgili literatüre atıf yapılarak açıklanır. Yöntem ve tekniklerin çalışmada öngörülen amaç ve hedeflere ulaşmaya elverişli olduğu ortaya konulur.

Yöntem bölümünün araştırmanın tasarımı, bağımlı ve bağımsız değişkenleri ve istatistiksel yöntemleri kapsamı gerekir. Araştırma önerisinde herhangi bir ön çalışma veya fizibilite yapıldıysa bunların sunulması beklenir. Araştırma önerisinde sunulan yöntemlerin iş paketleri ile ilişkilendirilmesi gerekir.

**Bu araştırma, Türkçe hukuki metinlerin analizi için küçük dil modelleri (SMLs) ve Retrieval-Augmented Generation (RAG) teknolojisini entegre eden yenilikçi bir sistem geliştirmeyi amaçlamaktadır.** Araştırma yöntemi, veri toplama, model geliştirme, sistem entegrasyonu ve performans değerlendirmesi aşamalarını içermektedir.

**Veri Toplama ve Ön İşleme aşamasında,** Veri seti, mevzuat.gov.tr üzerinden erişilen kanunlar, yönetmelikler ve kararnamelemler kullanılarak oluşturulacaktır. Bu veriler, JSON veya XML formatında indirilecek sırasıyla **Metin Temizleme**, özel karakterler gereksiz boşluklar ve formatlamalar temizlenecektir. **Tokenizasyon**, metinler kelime ve cümle düzeyinde token'lara ayrılacak. **Vektörleştirme**, metinler verimli arama için dense vektör temsillerine dönüştürülecek.

**Küçük Dil Modellerinin Geliştirilmesi aşamasında,** iki ayrı küçük dil modeli geliştirilecektir: **Retriever** ve **Generator**.

**Model Seçiminde,** DistilBERT veya ALBERT gibi küçük ön eğitilmiş modeller temel alınacak.

**Fine-tuning (İnce ayar),** Modeller, Türk hukuk metinleri üzerinde fine-tune edilecek. Fine-tuning, önceden eğitilmiş bir yapay zekâ modelini belirli bir görev veya veri kümesi için uyarlamak amacıyla kullanılan bir tekniktir. Bu süreçte, modelin ağırlıkları hedeflenen göreve özgü veriler kullanılarak güncellenir. Böylece model, genel bilgisini korurken spesifik alanlarda daha iyi performans gösterir. Fine-tuning, daha az veri ve hesaplama kaynağı gerektirdiği için, sıfırdan model eğitmeye kıyasla daha verimli bir yöntemdir.

**Hiperparametre Optimizasyonunda,** ADAM(ADEPTİVE MOMENTUM optimizasyon kullanılarak modeller optimize edilecek. Hiperparametre optimizasyonu, makine öğrenimi modellerinin performansını artırmak için kullanılan bir tekniktir. Bu süreçte, modelin eğitim algoritmasını kontrol eden hiperparametreler (örneğin öğrenme oranı, batch boyutu, ağaç derinliği) sistematik olarak ayarlanır. Amaç, modelin veriye en iyi şekilde uyum sağlamasını ve genelleme yeteneğini artırmasını sağlayan optimal hiperparametre kombinasyonunu bulmaktır.

## 2209/A ÜNİVERSİTE ÖĞRENCİLERİ ARAŞTIRMA PROJELERİ DESTEĞİ PROGRAMI ARAŞTIRMA ÖNERİSİ FORMU

Grid search, random search, Bayesian optimizasyon gibi çeşitli yöntemler kullanılarak gerçekleştirilen bu süreç, manuel deneme-yanılma yaklaşımına göre daha etkili ve verimlidir.

**Bağımsız Değişkenler:** Model mimarisi, eğitim veri seti boyutu, öğrenme oranı, batch size

**Bağımlı Değişkenler:** Modelin doğruluğu, F1 skoru, perplexity

**RAG Sistemi Entegrasyonu aşamasında,** RAG mimarisi, geliştirilen SML'ler ile şu şekilde entegre edilecektir. RAG (Retrieval-Augmented Generation) mimarisi, dil modellerinin performansını ve doğruluğunu artırmak için kullanılan bir yaklaşımdır. Bu mimari, geleneksel metin üretme modellerini bir bilgi geri çağırma sistemiyle birleştirir. İşleyişi şöyledir: Kullanıcının sorgusu alındığında, önce ilgili bilgiler geniş bir veritabanından çekilir. Ardından, bu bilgiler dil modeline girdi olarak verilir ve model, bu bağlamsal bilgileri kullanarak daha doğru ve güncel yanıtlar üretir. RAG, modelin halüsinasyon yapma olasılığını azaltır, bilgi tabanını güncel tutmayı kolaylaştırır ve özel alanlarda daha isabetli yanıtlar vermesini sağlar:

**Bilgi Tabanı Oluşturma:** Vektörleştirilmiş hukuki metinlerden oluşan bir bilgi tabanı oluşturulacak. Vektörleştirilmiş hukuki metinlerden oluşan bir bilgi tabanı oluşturma süreci, hukuki dokümanların dijital formatta işlenmesi ve anlamlı bir şekilde temsil edilmesini içerir. Bu süreçte, öncelikle geniş bir hukuki metin koleksiyonu (kanunlar, içtihatlar, akademik makaleler vb.) toplanır. Ardından, bu metinler doğal dil işleme teknikleri kullanılarak temizlenir ve ön işlemden geçirilir. Daha sonra, gelişmiş dil modelleri veya özelleştirilmiş gömme (embedding) teknikleri kullanılarak her bir metin veya metin parçası, anlamsal içeriğini yakalayan yüksek boyutlu vektörlere dönüştürülür. Bu vektörler, verimli arama ve geri çağırma işlemleri için optimize edilmiş bir veritabanında saklanır. Sonuç olarak, hukuki bilgilerin hızlı ve doğru bir şekilde sorgulanmasına, benzer davaların bulunmasına ve karmaşık hukuki analizlerin yapılmasına olanak tanıyan güçlü bir bilgi tabanı elde edilir.

**Retriever Entegrasyonu:** SML tabanlı retriever, kullanıcı sorgularını işleyecek ve ilgili bilgileri çekecek. Retriever entegrasyonu, RAG (Retrieval-Augmented Generation) mimarisinin önemli bir bileşeni olarak, kullanıcı sorgularını işleyerek ilgili bilgileri çekmeyi sağlar. SMLs (küçük dil modelleri) tabanlı retriever, kullanıcının sorgusunun anlamını analiz eder ve bilgi tabanında vektörleştirilmiş hukuki metinler arasından en alakalı olanları bulur. Bu yaklaşım, küçük dil modellerinin hızlı ve verimli çalışması sayesinde, büyük dil modellerine göre daha az kaynak kullanarak doğru ve spesifik bilgileri sunar. Böylece, hukuki bağlamda doğru yanıtlar elde etmek için gerekli bilgilerin hızlı bir şekilde erişilmesi sağlanır.

**Generator Entegrasyonu:** SML tabanlı generator, çekilen bilgileri kullanarak yanıtlar üretecek. SMLs (Small Language Models) tabanlı bir generator entegrasyonu, RAG mimarisinin çıktı üretme aşamasını optimize etmek için kullanılan bir yaklaşımdır. Bu süreçte, bilgi tabanından çekilen ilgili vektörleştirilmiş hukuki metinler, küçük ve özelleştirilmiş bir dil modeline girdi olarak verilir. SML, büyük dil modellerine kıyasla daha hızlı çalışır ve daha az kaynak tüketir. Generator, aldığı bağlamsal bilgileri kullanarak, kullanıcının sorgusuna özgü, tutarlı ve doğru yanıtlar üretir. Bu model, hukuki terminolojiye ve yapılar özel olarak eğitilmiş olduğundan, alan-spesifik bilgileri daha isabetli bir şekilde işler ve sunar. Sonuç olarak, hızlı, verimli ve hukuki bağlama uygun yanıtlar üreten bir sistem elde edilir.

**Bağlam Zenginleştirme:** Retriever'ın çektiği bilgiler, generator'a girdi olarak verilecek. Bağlam zenginleştirme, RAG mimarisinin kritik bir aşamasıdır ve retriever ile generator arasındaki köprüyü oluşturur. Bu süreçte, kullanıcının sorgusu temelinde retriever tarafından vektör veritabanından çekilen ilgili hukuki metinler, generator'a anlamlı ve işlenebilir bir formatta sunulur. Çekilen bilgiler, sorguyla ilişkilerine göre sıralanır, özetlenir veya birleştirilir. Bu işlem, gereksiz tekrarları önler ve en alakalı bilgilerin ön plana çıkmasını sağlar. Zenginleştirilmiş bağlam, generator'ın giriş formatına uygun şekilde yapılandırılır, böylece model bu bilgileri etkili bir şekilde kullanabilir. Bu aşama, generator'ın daha doğru, kapsamlı ve bağlama uygun yanıtlar üretmesini sağlayarak sistemin genel performansını artırır.

**Performans Değerlendirmesi ve Optimizasyon aşamasında,** sistemin performansı çeşitli metrikler ve yöntemlerle kapsamlı bir şekilde değerlendirilecektir. Doğruluk metrikleri arasında yer alan precision, recall ve F1 skoru, modelin tahmin gücünü analiz etmek için kritik öneme sahiptir. Precision (kesinlik), modelin pozitif tahminlerinin ne kadarının doğru olduğunu gösterir ve yanlış pozitiflerin (False Positive) etkisini azaltmayı amaçlar. Recall (duyarlılık) ise modelin, gerçek pozitif örneklerin ne kadarını doğru şekilde tespit edebildiğini ölçer ve yanlış negatiflerin (False Negative) sayısını azaltmaya odaklanır. F1 skoru, precision ve recall değerleri arasında bir denge kurarak dengesiz veri kümelerinde genel performansı daha doğru bir şekilde yansıtır.

Bağlamsal uygunluğu değerlendirmek için BLEU ve ROUGE skorları kullanılacaktır. BLEU (Bilingual Evaluation Understudy) skoru, özellikle makine çevirisi ve dil modelleme projelerinde, modelin ürettiği metnin referans metinle n-gram düzeyinde ne kadar örtüştüğünü analiz eder. ROUGE (Recall-Oriented Understudy for Gisting Evaluation) skoru ise metin özetleme gibi uygulamalarda öne çıkar ve modelin ürettiği metnin, referans metindeki n-gram'ları

## 2209/A ÜNİVERSİTE ÖĞRENCİLERİ ARAŞTIRMA PROJELERİ DESTEĞİ PROGRAMI ARAŞTIRMA ÖNERİSİ FORMU

ne oranda içerdiğini ölçer. ROUGE, özellikle recall ağırlıklı bir değerlendirme sunarken, her iki metrik de metinlerin dilsel doğruluğunu ve bağlamsal uygunluğunu ölçmekte önemli bir rol oynar.

Sistemin hesaplama verimliliği, işlem süresi ve bellek kullanımı gibi faktörlerle analiz edilecektir. İşlem süresi, bir işlemin tamamlanması için geçen süreyi ifade eder ve modelin veya algoritmanın hızını değerlendirmede kritik bir ölçüttür. Bellek kullanımı ise modelin çalışması sırasında ne kadar RAM (rastgele erişim belleği) tükettiğini gösterir ve sistem kaynaklarının verimli kullanımını belirler. İdeal bir sistem, hem işlem süresini hem de bellek kullanımını minimumda tutarak, büyük veri kümeleri veya karmaşık işlemlerle çalışırken yüksek performans sergilemelidir. Bu ölçütler, özellikle büyük ölçekli makine öğrenimi ve veri işleme projelerinde başarıyı belirlemek için önemli parametrelerdir.

Son olarak, sistem çıktılarının insan değerlendirmesi ile test edilmesi de büyük önem taşımaktadır. Hukuk uzmanlarından oluşan bir panel, sistemin ürettiği yanıtları değerlendirerek kalitesini ve pratikteki faydasını ölçecektir. Bu süreç, hem modelin hukuki doğruluğunu hem de kullanıcı memnuniyetini sağlamak açısından kritik bir rol oynayacaktır.

**Optimizasyon için kullanılacak teknikler arasında** erken durdurma (Early Stopping), öğrenme oranı ayarlama (Learning Rate Scheduling), model pruning ve quantization bulunmaktadır. Erken durdurma, makine öğrenimi modellerinin eğitim sürecinde aşırı öğrenmeyi önlemek amacıyla kullanılan bir tekniktir. Eğitim sırasında model, belirli bir süre boyunca doğrulama seti üzerinde performans artışı göstermezse, daha fazla eğitim yapmadan durdurulur. Bu sayede model, eğitim verilerine fazla uyum sağlamadan durdurularak, genel performansı düşürmeden en uygun noktada eğitim tamamlanır. Erken durdurma, eğitim süresini kısaltarak hesaplama maliyetini azaltırken, daha genelleşici ve dengeli modellerin elde edilmesine yardımcı olur.

Öğrenme oranı ayarlama, makine öğrenimi modelinin eğitim sürecinde kullanılan öğrenme oranını dinamik olarak değiştirme stratejisidir. Eğitim sürecinin başında yüksek bir öğrenme oranı kullanılarak modelin daha hızlı öğrenmesi sağlanırken, ilerleyen aşamalarda öğrenme oranı kademeli olarak azaltılır. Bu sayede model, geniş bir çözüm alanını hızlıca tararken, sonrasında küçük adımlarla ince ayar yaparak daha optimal sonuçlara ulaşır. Learning rate scheduling, eğitimin stabil hale gelmesini ve modelin minimum noktaya daha verimli bir şekilde yakınsamasını sağlar; bu teknik, hem daha hızlı eğitim süresi hem de daha düşük hata oranları elde etmek için kritik öneme sahiptir.

Model pruning ve quantization, makine öğrenimi modellerinin performansını artırmak ve kaynak kullanımını optimize etmek için kullanılan iki temel tekniktir. Model pruning, gereksiz veya düşük etkili bağlantıları veya ağırlıkları modelden çıkararak boyutunu küçültmeyi amaçlar. Bu işlem, modelin karmaşıklığını azaltırken daha hızlı işlem yapmasını sağlar. Quantization ise modeldeki ağırlıkların ve aktivasyonların daha düşük bit seviyelerinde temsil edilmesini sağlayarak hesaplama yükünü ve bellek kullanımını azaltır. Örneğin, 32-bit kayan noktalı sayılar yerine 8-bit tam sayılar kullanılabilir. Bu iki teknik, özellikle sınırlı kaynaklara sahip edge cihazlarda model performansını artırırken, doğruluğun makul seviyede tutulmasına katkı sağlar.

**“LEGALTURK OPTIMIZED BERT FOR MULTI-LABEL TEXT CLASSIFICATION AND NER”** başlıklı çalışma, Türk hukuki metinlerine yönelik BERT modelinin nasıl optimize edildiğini ele almaktadır. Bu çalışmada, BERT’in ön eğitim aşamasına çeşitli yenilikçi maskeleme stratejileri ve Sonraki Cümle Tahmini (NSP) yerine Cümle Sırası Tahmini (SOP) gibi yöntemler eklenmiş; bu sayede Türk hukuki metinlerinde daha iyi performans elde edilmesi amaçlanmıştır. Modeller, Adlandırılmış Varlık Tanıma (NER) ve çok etiketli sınıflandırma görevlerinde önemli iyileştirmeler göstermiştir. Bu çalışma, hukuki metinlerin işlenmesinde BERT tabanlı modellerin potansiyelini sergileyerek, projemizde kullanılacak dil modellerinin hukuki metinler için nasıl uyarlanabileceğine dair önemli bilgiler sunmaktadır.

**“Yeung Tai (2019) tarafından yapılan “Fine-tuning BERT for German Legal Language”** çalışması ise BERT modelinin Almanca hukuki metinler üzerinde nasıl ince ayarlandığını ve alan odaklı ince ayar (domain-adaptive fine-tuning) stratejilerinin etkilerini araştırmaktadır. Çalışmada, hukuki kelime dağarcığının modele entegrasyonu ile elde edilen performans iyileştirmeleri ele alınmış ve bu iyileştirmelerin marjinal olduğu belirtilmiştir. Bu bulgular, projemizde kullanılacak modellerin verimliliğini artırmak için doğru fine-tuning stratejilerinin belirlenmesine rehberlik etmektedir.

**“Wang ve arkadaşlarının (2024) “Exploring New Frontiers of Deep Learning in Legal Practice”** adlı çalışması, hukuki yapay zekâ platformlarının geliştirilmesi sırasında karşılaşılan zorlukları incelemektedir. ChatLaw gibi büyük dil modelleri, veri doğruluğu ve tutarlılığı açısından önemli zorluklarla karşılaşmış ve bu modellerde “konsültasyon”, “referans” ve “öneri” modülleri gibi yapılar geliştirilmiştir. Bu çalışma, projemizde RAG (Retrieval-Augmented Generation) mimarisini kullanırken benzer yaklaşımların benimsenmesine olanak tanıyacaktır.

**2209/A ÜNİVERSİTE ÖĞRENCİLERİ ARAŞTIRMA PROJELERİ DESTEĞİ PROGRAMI**  
**ARAŞTIRMA ÖNERİSİ FORMU**

---

Bu çalışmalar, projemizin temellerini oluşturarak hukuki metin işleme alanında kullanılan yöntemlerin belirlenmesine ve mevcut modellerin sınırlarının farkında olunmasına katkı sağlamaktadır. Projemiz, bu çalışmalar ışığında Türk hukuku alanında yenilikçi bir yaklaşım sunmayı hedeflemekte; Retrieval-Augmented Generation (RAG) mimarisinin küçük dil modelleri ile entegrasyonunu içermekte ve Türkçe hukuki metinlerin verimli ve doğru analizi için çözüm önermektedir.



## 3 PROJE YÖNETİMİ

## 3.1 İş- Zaman Çizelgesi

Araştırma önerisinde yer alacak başlıca iş paketleri ve hedefleri, her bir iş paketinin hangi sürede gerçekleştirileceği, başarı ölçütü ve araştırmanın başarısına katkısı “İş-Zaman Çizelgesi” doldurularak verilir. Literatür taraması, gelişme ve sonuç raporu hazırlama aşamaları, araştırma sonuçlarının paylaşımı, makale yazımı ve malzeme alımı ayrı birer iş paketi olarak gösterilmemelidir.

Başarı ölçütü olarak her bir iş paketinin hangi kriterleri sağladığında başarılı sayılacağı açıklanır. Başarı ölçütü, ölçülebilir ve izlenebilir nitelikte olacak şekilde nicel veya nitel ölçütlerle (ifade, sayı, yüzde, vb.) belirtilir.

## İŞ-ZAMAN ÇİZELGESİ (\*)

İP No	İş Paketlerinin Adı ve Hedefleri	Kim(ler) Tarafından Gerçekleştirileceği	Zaman Aralığı (.-.. Ay)	Başarı Ölçütü ve Projenin Başarısına Katkısı
1	Verilerin Elde Edilmesi	ABDÜLSAMET TÜRKMENOĞLU	1-KASIM İLE 15 KASIM	%33
2	Türkçe SLM'lerin Geliştirilmesi	ABDÜLSAMET TÜRKMENOĞLU	15 KASIM İLE 19 OCAK	%33
3	Başarım Testleri ve Değerlendirme	ABDÜLSAMET TÜRKMENOĞLU	19 OCAK İLE 1 MAYIS	%34

(\*) Çizelgedeki satırlar ve sütunlar gerektiği kadar genişletilebilir ve çoğaltılabilir.



**2209/A ÜNİVERSİTE ÖĞRENCİLERİ ARAŞTIRMA PROJELERİ DESTEĞİ PROGRAMI**  
**ARAŞTIRMA ÖNERİSİ FORMU**

### 3.2 Risk Yönetimi

Araştırmanın başarısını olumsuz yönde etkileyebilecek riskler ve bu risklerle karşılaşıldığında araştırmanın başarıyla yürütülmesini sağlamak için alınacak tedbirler (B Planı) ilgili iş paketleri belirtilerek ana hatlarıyla aşağıdaki Risk Yönetimi Tablosu'nda ifade edilir. B planlarının uygulanması araştırmanın temel hedeflerinden sapmaya yol açmamalıdır.

**RİSK YÖNETİMİ TABLOSU\***

İP No	En Önemli Riskler	Risk Yönetimi (B Planı)
1	Yağacağımız ADAM algoritması verimli sonuç vermesi	AdaGrad algoritması denenecektir
2		

(\*) Tablodaki satırlar gerektiği kadar genişletilebilir ve çoğaltılabilir.

### 3.3. Araştırma Olanakları

Bu bölümde projenin yürütüleceği kurum ve kuruluşlarda var olan ve projede kullanılacak olan altyapı/ekipman (laboratuvar, araç, makine-teçhizat, vb.) olanakları belirtilir.

**ARAŞTIRMA OLANAKLARI TABLOSU (\*)**

Kuruluşta Bulunan Altyapı/Ekipman Türü, Modeli (Laboratuvar, Araç, Makine-Teçhizat, vb.)	Projede Kullanım Amacı
Bilgisayar Bilimleri Laboratuvarı	

(\*) Tablodaki satırlar gerektiği kadar genişletilebilir ve çoğaltılabilir.

### 4. YAYGIN ETKİ

Önerilen çalışma başarıyla gerçekleştirildiği takdirde araştırmadan elde edilmesi öngörülen ve beklenen yaygın etkilerin neler olabileceği, diğer bir ifadeyle yapılan araştırmadan ne gibi çıktı, sonuç ve etkilerin elde edileceği aşağıdaki tabloda verilir.

**ARAŞTIRMA ÖNERİSİNDEN BEKLENEN YAYGIN ETKİ TABLOSU**

Yaygın Etki Türleri	Önerilen Araştırmadan Beklenen Çıktı, Sonuç ve Etkiler
<b>Bilimsel/Akademik</b> (Makale, Bildiri, Kitap Bölümü, Kitap)	
<b>Ekonomik/Ticari/Sosyal</b> (Ürün, Prototip, Patent, Faydalı Model, Üretim İzni, Çeşit Tescili, Spin-off/Start-up Şirket, Görsel/İşitsel Arşiv, Envanter/Veri Tabanı/Belgeleme Üretimi, Telif Konu Olan Eser, Medyada Yer Alma, Fuar, Proje Pazarı, Çalıştay, Eğitim vb. Bilimsel Etkinlik, Proje Sonuçlarını Kullanacak Kurum/Kuruluş, vb. diğer yaygın etkiler)	

**2209/A ÜNİVERSİTE ÖĞRENCİLERİ ARAŞTIRMA PROJELERİ DESTEĞİ PROGRAMI**  
**ARAŞTIRMA ÖNERİSİ FORMU**

**Araştırmacı Yetiştirilmesi ve Yeni Proje(ler) Oluşturma**  
(Yüksek Lisans/Doktora Tezi, Ulusal/Uluslararası Yeni Proje)

Tübitak yada ardeb projesi geliştirilecektir.

**5. BÜTÇE TALEP ÇİZELGESİ**

Bütçe Türü	Talep Edilen Bütçe Miktarı (TL)	Talep Gerekçesi
Sarf Malzeme		
Makina/Teçhizat (Demirbaş)		
Hizmet Alımı		
Ulaşım		
TOPLAM		

**NOT:** Bütçe talebiniz olması halinde hem bu tablonun hem de TÜBİTAK Yönetim Bilgi Sistemi (TYBS) başvuru ekranında karşınıza gelecek olan bütçe alanlarının doldurulması gerekmektedir. Yukardaki tabloda girilen bütçe kalemlerindeki rakamlar ile, TYBS başvuru ekranındaki rakamlar arasında farklılık olması halinde TYBS ekranındaki veriler dikkate alınır ve başvuru sonrasında değiştirilemez.

**6. BELİRTMEK İSTEDİĞİNİZ DİĞER KONULAR**

Sadece araştırma önerisinin değerlendirilmesine katkı sağlayabilecek bilgi/veri (grafik, tablo, vb.) eklenebilir.

**7. EKLER**

**EK-1: KAYNAKLAR**

Yapay Zekâ Okuryazarlığı Ölçeğinin Türkçeye Uyarlanması(<https://doi.org/10.53694/bited.1376831>)

Aydın ve Mehmet, 2024(<https://doi.org/10.48550/arXiv.2403.01309>)

LegalTurk Optimized BERT for Multi-Label Text Classification and NER(<https://doi.org/10.48550/arXiv.2407.00648>)

Chatlaw: A Multi-Agent Collaborative Legal Assistant with Knowledge Graph Enhanced Mixture-of-Experts Large Language Model(<https://doi.org/10.48550/arXiv.2306.16092>)

LawBench: Benchmarking Legal Knowledge of Large Language Models(<https://doi.org/10.48550/arXiv.2309.16289>)

Lawformer: A pre-trained language model for Chinese legal long documents(<https://doi.org/10.1016/j.aiopen.2021.06.003>)

Retrieval-Augmented Generation for Knowledge-Intensive NLP

**2209/A ÜNİVERSİTE ÖĞRENCİLERİ ARAŞTIRMA PROJELERİ DESTEĞİ PROGRAMI**  
**ARAŞTIRMA ÖNERİSİ FORMU**

---

Tasks(<https://doi.org/10.48550/arXiv.2005.11401>)