**PDF PROCESSING SYSTEM**

Barış Yasin ŞAHİN

Yusuf YENİGÜN

Beyzanur KAYA

[barısyasinsahin@stu.topkapi.edu.tr](mailto:barısyasinsahin@stu.topkapi.edu.tr)

[yusufyenigun@stu.topkapi.edu.tr](mailto:yusufyenigun@stu.topkapi.edu.tr)

[beyzanurkaya1@stu.topkapi.edu.tr](mailto:beyzanurkaya1@stu.topkapi.edu.tr)

ÖZET

Bu projede, birden fazla PDF dosyasından metinlerin çıkarılması, metinlerin özetlenmesi ve tek bir dosyada birleştirilerek kullanıcının kolay erişebileceği bir PDF dosyası oluşturulması hedeflenmiştir. Bunun için **transformer** mimarisine dayanan ve derin öğrenme tabanlı bir model olan **BART modeli** kullanılmıştır. Geliştirilen model, verilen PDF dosyalarını başarıyla okur ve özetler.

*Keywords: deeplearning, pdf summary, transformer, attention mask, BART*

1. GİRİŞ

Uzun PDF dosyaları özellikle öğrenciler için çalışmayı zorlaştıran bir faktördür. Günümüzde yapay zeka ve derin öğrenme hayatın her alanında etkin bir yardımcı olarak kullanılır. Bu bağlamda, derin öğrenme yöntemlerinin eğitim sektöründe de yaygınca kullanıldığını söyleyebiliriz.

Bu proje kapsamında transformer kütüphanesi kullanılarak uzun metinlerin özetlenmesi amaçlanmıştır. Projenin temel amacı birden fazla PDF dosyasından metinlerin çıkarılması, özetlenmesi ve tek bir pdf halinde kullanıcıya sunulmasıdır. Bu raporda kütüphane seçiminden, modelin oluşturulmasına kadar tüm aşamalar ele alınmıştır.

2. PDF PROCESSING SYSTEM

Bu bölümde PDF dosyalarından metin çıkarma, metni özetleme, yeni PDF oluşturma ve oluşturulan PDF dosyalarının birleştirilmesi ele alınmıştır.

2.1. METİN ÇIKARMA

PDF dosyalarındaki metinler **PymuPDF (fitz)** kütüphanesi kullanılarak çıkarılmıştır. Bu işlem her sayfayı ayrı ayrı işler ve tekrar birleştirir.

2.2 METİN ÖZETLEME

Transformer kütüphanesinin bir modeli olan BART**,** metinlerin özetlenmesi için kullanılmıştır. Transformer mimarisi dikkat mekanizmaları (attention mask) ve çok katmanlı sinir ağlarına dayanmaktadır. **Attention mask**, derin öğrenme modellerinde padding gibi gereksiz bilgilerin dikkate alınmamasını sağlar. Modelin verimliliğini arttırır.

2.3 PDF OLUŞTURMA

FPDF kütüphanesi pdf dosyaları oluşturmaya yarar. Türkçe karakterleri de desteklediği için tercih edilmiştir.

2.4 PDF DOSYALARININ BİRLEŞTİRİLMESİ

Modelde her PDF dosyası ayrı ayrı okunur ve özetlenir. PyPDF2 kütüphanesi oluşturulan PDF dosyalarını birleştirmek için kullanılır. Birleştirilen PDF kullanıcıya sunulur.

3. METOD VE MATERYAL

Transformers kütüphanesinden, çok büyük bir metin verisi üzerinde önceden eğitilmiş olan ve yüksek doğruluk yüzdesine sahip BARTmodeli tercih edilmiştir.

BART, hem bir kodlayıcı (encoder) hem de bir kod çözücü (decoder) içeren bir Transformer modelidir. Bu, modeli hem giriş metinlerini anlamakta (kodlayıcı) hem de özet veya çıktı metni üretmekte (kod çözücü) çok yönlü kılar.

Projemizde kullandığımız **facebook/bart-large-cnn**, metin özetleme görevlerine özel olarak ince ayar yapılmış bir BART modelidir. facebook/bart-large-cnn çok büyük veri kümelerinde eğitildiği için insan benzeri özetler üretebilir ve CNN/DailyMail **veri seti** üzerinde eğitilmiştir.

4. TEST

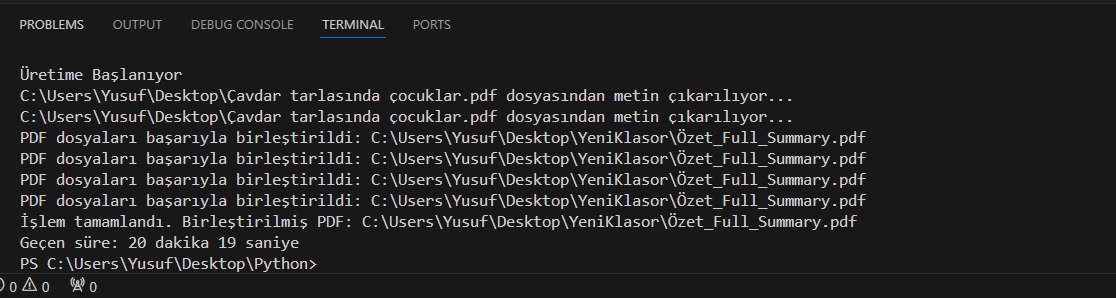
Modelin test aşamasında Türkçe ve İngilizce PDF dosyaları kullanılmıştır.

Yüklenen PDF dosyalarının özeti elde edilmiştir.

Model her iki dili de algılar, okur ve özetler. PDF dosyalarını başarılı bir şekilde birleştirir ve yolu verilen dosyaya kaydeder.

5. SONUÇ

Bu çalışmada uzun PDF dosyalarının işlenmesi için facebook/bart-large-cnn modeli kullanılmıştır. Transformer kütüphanesi üzerinden geliştirilen proje birden çok PDF dosyasından metni ayrı ayrı çıkarır, özetler ve birleştirir. Kullanıcıya özet PDF dosyasını sağlar.

 **ŞEKİL 1: PROJE SONUCU**  
Şekil 1’de projenin sonuç görseli verilmiştir. Projenin ortalama çalışma süresi 20.19 dakikadır.

**CV**

**------------------------------------**

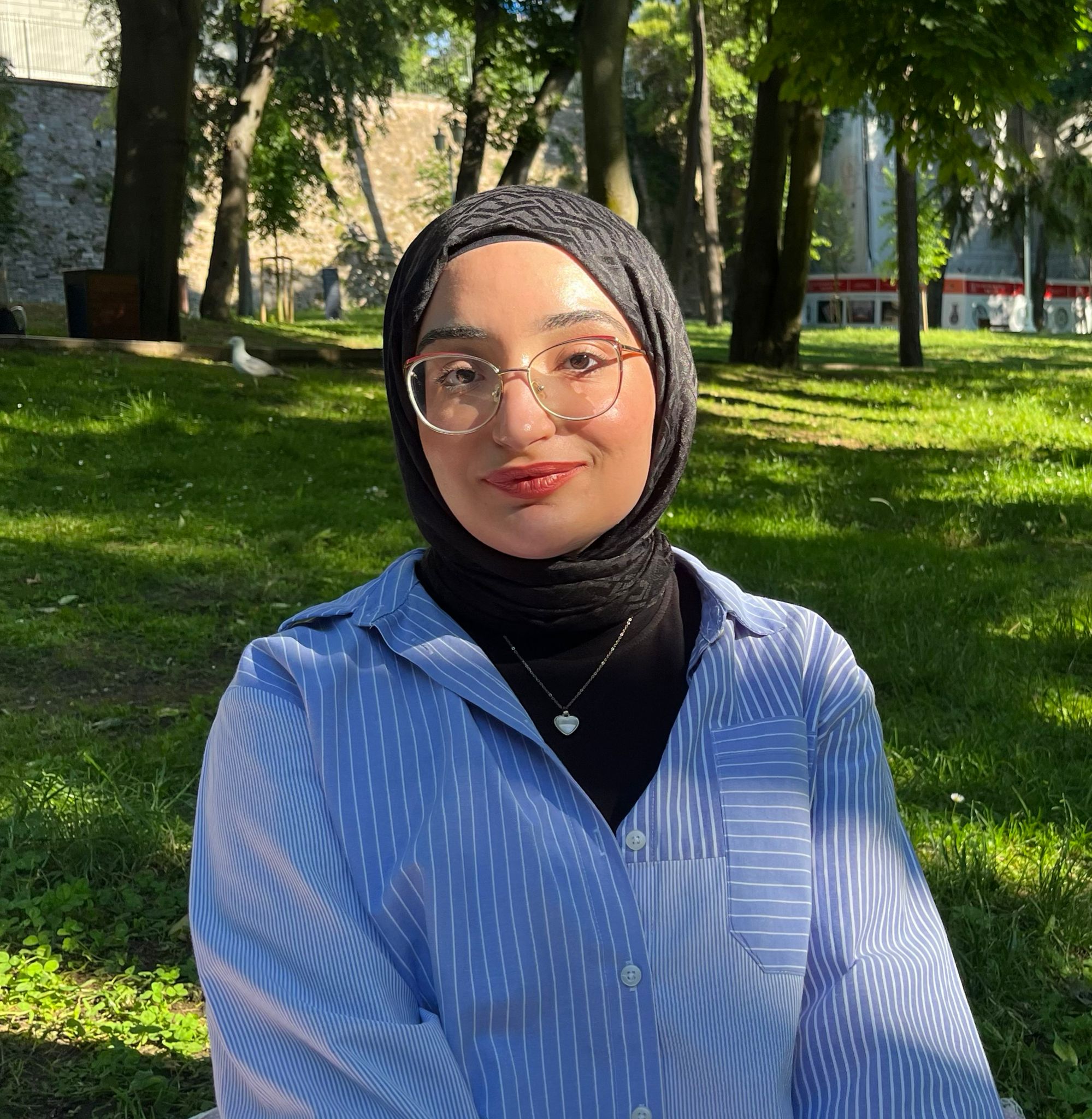
**Barış Yasin ŞAHİN**

İstanbul Topkapı Üniversitesi Yazılım Mühendisliği 3. sınıf öğrencisiyim ve Python, Java, C, C++ ile HTML gibi programlama dillerinde güçlü bir teknik altyapıya sahibim. Yapay zeka alanında cilt kanserlerinin teşhisini hedefleyen yenilikçi bir proje geliştirerek, makine öğrenimi modelleri ve veri analizi konularında derinlemesine deneyim kazandım. Problem çözme yeteneğim, hızlı öğrenme kabiliyetim ve takım çalışmasına yatkınlığım sayesinde projelerde etkin bir katkı sağlıyorum. Teknolojiye olan tutkum, yapay zekanın sınırlarını zorlayan ve sektörde iz bırakacak projeler geliştirme kararlılığımla birleşiyor. Hedefim, bilgi birikimimi sürekli genişleterek yenilikçi çözümler sunmak ve yapay zeka alanında unutulmaz projelere imza atmaktır.

**Yusuf YENİGÜN**

İstanbul Topkapı Üniversitesi Yazılım Mühendisliği 3. sınıf öğrencisiyim. Python, Java, C, C++ ve C# gibi programlama dillerinde bilgiliyim. Yapay zeka alanında belirli cilt kanserlerinin teşhisi üzerine bir program yaptım. Bir projede çalışırken istenilen görevleri yerine en düzgün haliyle getirmeye çabalayan biriyim. Yapay zeka alanından ilerleyip projelerimi buna göre şekillendirmek istiyorum.

**Beyzanur KAYA**

İstanbul Topkapı Üniversitesi Yazılım Mühendisliği 3. sınıf öğrencisiyim. Python, Java, C#, C++, HTML, PHP gibi programlama dillerine hakimim. Makine öğrenimi modelleri veri analizi ve frontendde deneyim kazandım. Yaratıcılığım ve sorun çözme kabiliyetim ile grup çalışmalarına katkı sağlıyorum. Kendimi her gün geliştirmeye çalışıyorum. Hedefim yaratıcılığımı bilgilerimle birleştirerek fark yaratmak.