

G1_Ara_21011910

by Hira Nur Morca

Submission date: 07-Jun-2024 09:28PM (UTC+0300)

Submission ID: 2397657050

File name: G1_Ara_21011910.pdf (2.5M)

Word count: 6087

Character count: 41740

TÜRKİYE CUMHURİYETİ
YILDIZ TEKNİK ÜNİVERSİTESİ
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ



HABER ÖZETLEYİCİ

21011910 – Hira Nur Morca

BİLGİSAYAR PROJESİ

Danışman
Dr. Öğr. Üyesi Göksel BİRİCİK

Haziran, 2024

© Bu projenin bütün hakları Yıldız Teknik Üniversitesi Bilgisayar Mühendisliği Bölümü'ne aittir.

TEŞEKKÜR

Bu proje boyunca öncesinde çok deneyime sahip olmadığım Doğal Dil İşleme ve Büyük Dil Modelleri konusunda bana literatür taraması yapma ve akabinde bir proje geliştirme adına bana fırsat tanıyan sayın danışman hocam Dr. Göksel BİRİCİK'e teşekkür ediyorum. Benim için oldukça eğitici ve öğretici geçen bu süreçte, biz öğrencilerini ortaya bu ve buna benzer projeler ortaya koyması için teşvik eden bölümümüz ve üyelerine de ayrıca teşekkür etmeliyim.

Hira Nur Morca

İÇİNDEKİLER

KISALTMA LİSTESİ	v
ŞEKİL LİSTESİ	vi
TABLO LİSTESİ	viii
ÖZET	ix
ABSTRACT	x
1 Giriş	1
2 Ön İnceleme	3
3 Fizibilite	5
3.1 Teknik Fizibilite	5
3.1.1 Donanım Fizibilitesi	5
3.1.2 Yazılım Fizibilitesi	5
3.2 Ekonomik Fizibilite	6
3.3 Zaman Fizibilitesi	6
3.4 Yasal Fizibilite	6
4 Sistem Analizi	7
4.1 Bileşenler	7
4.2 İşlevler	7
5 Sistem Tasarımı	9
5.1 Yazılım Tasarımı	9
5.1.1 T5, Mimarisi ve Eğitimi	9
5.1.2 PEGASUS, Mimarisi ve Eğitimi	11
5.1.3 mT5, Mimarisi ve Eğitimi	12
5.2 Veritabanı Tasarımı	12
5.3 Girdi - Çıktı Tasarımı	14
6 Uygulama	16

6.1	LLM Modelerin Fine-tuning İşleminden Geçirilmesi	16
6.1.1	Fine-tuning İşlemleri	16
6.2	Modelerin HuggingFace Platformunda Tutulması	20
6.3	Haber Sitelerinden Canlı Haber Çekimi	21
6.4	Arayüzün Gerçeklenmesi	22
7	Deneysel Sonuçlar	24
8	Performans Analizi	31
9	Sonuç	33
	Referanslar	34
	Özgeçmiş	34

KISALTMA LİSTESİ

LLM	Large Language Model
NLP	Natural Language Processing
T5	Text-to-Text Transfer Transformer
PEGASUS	Pre-Training with Extracted Gap-sentences for Abstractive Summarization
GSG	Gap Sentences Generation
MLM	Masked Language Modeling
ROUGE	Recall-Oriented Understudy for Gisting Evaluation

ŞEKİL LİSTESİ

Şekil 3.1 Gantt Şeması	6
Şekil 5.1 Örnek Transformer Mimarisi	10
Şekil 5.2 T5 Eğitiminde yapılan 'gap' çalışması	11
Şekil 5.3 PEGASUS Mimarisi ve Eğitim Tekniği	11
Şekil 5.4 CNN/DailyMail Veriseti	13
Şekil 5.5 'summary-of-news-articles' Veriseti	13
Şekil 5.6 TR-News Veriseti	13
Şekil 5.7 Haberlerin Tutulmasına İlişkin Temsili Görsel	14
Şekil 6.1 Veriseti ve Modelin Yüklenmesi	17
Şekil 6.2 Verisetinin Önislemeden Geçirilmesi	18
Şekil 6.3 Eğitimde elde edilen başarı skorları	18
Şekil 6.4 Pegasus ile elde edilen başarı skorları	19
Şekil 6.5 mT5 için İlk Başarım Sonuçları	20
Şekil 6.6 mT5 için son Başarım Sonuçları	20
Şekil 6.7 Modelin HuggingFace ortamına yüklenmesi	21
Şekil 6.8 Modele ait HuggingFace Kütüphanesi	21
Şekil 6.9 Arayüz	23
Şekil 7.1 İngilizce Opsiyonu için Arayüz	24
Şekil 7.2 Türkçe Opsiyonu için Arayüz	25
Şekil 7.3 1- Türkçe bir haberin tamamının görüntülenmesi	25
Şekil 7.4 1- mT5 sonucu elde edilen özet	25
Şekil 7.5 2- Türkçe bir haberin tamamının görüntülenmesi	26
Şekil 7.6 2- mT5 sonucu elde edilen özet	26
Şekil 7.7 3- Türkçe bir haberin tamamının görüntülenmesi	26
Şekil 7.8 3- mT5 sonucu elde edilen özet	26
Şekil 7.9 1- İngilizce bir haberin tamamının görüntülenmesi	27
Şekil 7.10 1- T5 sonucu elde edilen özet	27
Şekil 7.11 2- İngilizce bir haberin tamamının görüntülenmesi	27
Şekil 7.12 2- T5 sonucu elde edilen özet	28
Şekil 7.13 3- İngilizce bir haberin tamamının görüntülenmesi	28
Şekil 7.14 3- T5 sonucu elde edilen özet	28

Şekil 7.15 1- İngilizce bir haberin tamamının görüntülenmesi	29
Şekil 7.16 1- Pegasus sonucu elde edilen özet	29
Şekil 7.17 2- İngilizce bir haberin tamamının görüntülenmesi	29
Şekil 7.18 2- Pegasus sonucu elde edilen özet	30
Şekil 7.19 3- İngilizce bir haberin tamamının görüntülenmesi	30
Şekil 7.20 3- Pegasus sonucu elde edilen özet	30

TABLO LİSTESİ

Tablo 8.1 Modellerin Boyutları ve Parametre Sayıları	31
Tablo 8.2 Modellerin Eğitim ve Validasyon Loss Değerleri	31
Tablo 8.3 Modellerin Çeşitli Rouge Skor Değerleri	32

ÖZET

HABER ÖZETLEYİCİ

Hira Nur Morca

1
Bilgisayar Mühendisliği Bölümü
Bilgisayar Projesi

Danışman: Dr. Öğr. Üyesi Göksel BİRİCİK

Büyük Dil Modelleri, insanların doğal dil anlayışını taklit etmeyi amaçlayan kompleks yapay zeka modelleri olarak ortaya çıkmışlardır. Günümüzde, gelişen Büyük Dil Modelleri doğal dillerin yapı taşı olduğu kelime tahmini, soru cevaplama, duyu analizi gibi çeşitli görevlerde oldukça başarılı hale gelmiştir. Uzun ve detaylı metinlerden, önemli noktalar elde edilerek daha kısa bir metin oluşturulduğu özetleme işlemi de bu görevlerden birisidir. İşbu çalışmada, haber özetleme alanına yoğunlaşmış ve bunun için bir web uygulaması geliştirilmiştir. Bu süreçte, LLM'ler ve LLM'ler tarafından yapılan özetlemeler üzerine bir literatür taraması yapılmış ve önceden eğitilmiş büyük dil modelleri özetleme görevi için fine-tune edilmiştir. Haberlerin canlı olarak yayınlandıkları sitelerden çekilme işlemini yapacak uygulama yazılmış ve ardından fine-tune edilmiş modeller ve bu uygulama bir web uygulamasına implemente edilerek son kullanıcıya hazır bir hale getirilmiştir.

Anahtar Kelimeler: Doğal Dil İşleme, Büyük Dil Modelleri, Metin Özetleme

ABSTRACT

NEWS SUMMARIZER

Hira Nur Morca

1 Department of Computer Engineering
Computer Project

Advisor: Assist. Prof. Dr. Göksel BİRİCİK

Large Language Models are complex artificial intelligence models built to mimic the understanding of natural languages by humans. In today's world, LLMs have become very successful in performing tasks based on natural languages, such as word guessing, question answering and sentiment analysis. Summarization, creating a relatively short text by extracting important points from a long and detailed text, is one of these tasks. This project focuses on the area of news summarization and building a web application for it. During this period, a literature review of LLMs and Summarization have been made and pre-trained LLM's have been fine-tuned for news summarization. An application for scraping news from sites have been implemented. After that, a web application have been built which is ready for end-user, implementing fine-tuned LLMs and news scraping application.

Keywords: Natural Language Processing, Large Language Models, Text Summarization

1

Giriş

Büyük Dil Modelleri, doğal dillerle yazılmış metinleri anlamak ve benzer metinler üretmek üzere tasarlanmış gelişmiş yapay zeka sistemleridir. Kompleks mimarilere sahip bu sistemler büyük ölçekli veriler üzerinde eğitilerek soru cevaplamadan içerik üretmeye kadar birtakım doğal dile has işlevlerini gerçekleştirmeye kapasitesine sahiptirler. İşbu işlevlerden birisi olan metin özetleme, elde bulunan bir metnin önemli ve gerekli noktalarına yoğunlaşarak daha kısa hale getirmeyi amaçlar [1]. Büyük Dil Modelleri, bu işlemde neredeyse insan seviyesi başarıya ulaşmıştır [2].

Bu çalışmada, büyük dil modellerinden İngilizce haberler için fine-tune edilmiş T5 ve Pegasus; Türkçe haberler için ise fine-tune edilmiş mT5 modelleri entegre edilerek, canlı olarak çekilen haberleri özetleyebilecek bir sistem oluşturulmak hedeflenmiştir. Fine-tuning işlemi için literatürde sıklıkla kullanılan CNN/DailyMail ve HuggingFace ortamında açık kaynak verisetlerden yararlanılmıştır. CNN/DailyMail, CNN ve DailyMail haber sitelerinden çekilen haberlerin özetlerinin bulunduğu bir verisetidir [3].

Bu çalışmanın arkasında bulunan ana motivasyon, derin öğrenme, yapay sinir ağları ve büyük dil modelleri alanında literatür taraması yaparak mimariyi kavramak; edinilen bilgileri implemente ederek ortaya kullanılabılır ve ulaşılabilir bir ürün koymaktır. Böylelikle bu alanda yapılan çalışmalarla bir katkı sunulması, anlaşılabilmesi ve kullanımının yaygınlaşabilmesi beklenmektedir.

Çalışmada implemente edilecek T5, mT5 ve PEGASUS modelleri, doğal dil işleme alanında belli başlı 'state-of-art' başarıya ulaşmış modellerdir. Her model de Transformers mimarisine kurulmuş ve büyük ölçekli verilerle amaçlarına uygun şekilde eğitilmişlerdir. T5 modeli, özetleme dışında birçok doğal dil işleme görevlerini yerine getirir, eğitimi esnasına özetleme de yapabilmesi adına train edilmiştir. PEGASUS ise sadece metin özetleme üzerine eğitilmiş bir modeldir ve bu alanda en iyi model olarak nitelendirilmektedir [1][4]. mT5, T5 mimarisinin 101 farklı dil üzerine train edilmiş versiyonudur[5].

Sonuç olarak, işbu çalışmada LLM'lerin haber özetleme alanında haber özetleme verisetleri yardımıyla fine-tune edilmesi ve haber sitelerinde bulunan haberlerin özetini çıkarması üzerine yoğunlaşmıştır. Çalışmada ayrıca LLM'lerin belli bir 'görev' üzerine nasıl fine-tune edileceği ve bir uygulamaya nasıl entegre edileceği üzerine de bir anlayış sağlanmaya çalışılmıştır.

Raporun geri kalan kısımları olan 2. bölümde yapılan ön incelemelerden ve literatür taramasından bahsedilmiş, 3. bölümde projenin fizibilite çalışmasına yer verilmiş, 4. bölümde sistem analizi 5. bölümde sistemin tasarımı üzerinde durulmuş, 6. bölümde uygulama açıklanmış, 7. bölümde deneysel sonuçlar gösterilmiştir, 8. bölümde performans analizinden bahsedilmiş ve 9. bölümde açıklamalar sonuçlanmıştır. Son bölüm olan 10. bölümde referanslara yer verilmiştir.

2 Ön İnceleme

Proje için yapılan ön incelemede, öncelik verilen ilk nokta Derin Öğrenme ve NLP üzerine araştırma yapılması ve bu alanlarda yeterli ölçüde bilgi sahibi olunması olmuştur. Bunun ardından, LLM mimarileri ve çalışma prensipleri üstüne literatür taraması gerçekleştirilmiş ve alanda onde gelen makaleler incelenmiştir. Bu adımların gerçekleşmesindeki ana amaçlar LLM'lerin mimarilerini kavramak, farklı alanlarda farklı kullanım şekillerine hakim olarak, işbu projede uygulamaya implemente edilecek olanların belirlenmesidir. Diğer yandan, LLM'lerin eğitilme ve belli bir görev için 'ince ayar (fine-tuning)' aşamaları hakkında bilgi sahibi olmak hedeflenmiş, elde edilecek olan özet metinlerin ne kadar başarılı olduğunu ölçme konusunda izlenecek yolların kararlaştırılması beklenmiştir. Bunun dışında, uygulamada kullanılmak üzere elde edilecek haberlerin, yayıcı haber sitelerinden nasıl elde edileceği üzerine bilgi sahibi olmak için çeşitli kaynaklardan yararlanılmıştır.

Büyük dil modellerinin özetleme çalışmaları sadece haber özetleme kısıtına dayalı değildir. LLM'ler makaleler, diyaloglar, resmi belgeler gibi çeşitli metinlerin özetini de çıkarma konusunda eğitilmektedirler. Bunun dışında büyük dil modelleri metin özetlemelerinde iki farklı yol izlemektedir. Bunlar 'extractive' özetleme ve 'abstractive' özetleme olarak ikiye ayrılır. 'Extractive' özetleme, ilgili metin genelinde yaptığı hesaplamalar sonucu en önemli cümleleri bulur ve hiçbir değişiklik yapmadan özet metini bu cümlelerin birer kopyasıyla oluşturur. 'Abstractive' özetleme ise, metinin büyük dil modeli tarafından 'analiz' edilerek 'kendi cümleleriyle' özet çığırdığı tekniktir[6]. Projede, daha sağlıklı çıktı alınması açısından 'abstractive' özetleme yapılması hedeflenmiştir. Bu noktada, yapılan araştırmalar sonucu, projeye implemente etmek adına İngilizce arayüz için T5 ve PEGASUS, Türkçe arayüz için ise mT5 modelinde karar kılınmıştır. Google Research tarafından geliştirilmiş bu LLM'ler ana kullanım amaçları veya eğitim şekilleri gibi teknik konularda birbirlerinden ayrılmaktadır. T5, Text-to-Text Transfer Transformer, büyük bir metin verisi üzerinde, birden fazla NLP görevini yerine getirmek üzere eğitilmiş bir dil modelidir [4]. PEGASUS ise, direkt olarak özetleme işlemi alanında özelleştirilmek üzere eğitilmiş

bir dil modelidir [1]. Türkçe metinler için seçilen mT5 ise, T5 modelinin C4 veriseti üzerinde pretrain edilmiş versiyonudur. Bu model, 101 farklı dilde çıktı üretemesine rağmen herhangi bir NLP görevi için fine-tune edilmemişinden dolayı, kullanımından önce bu işlemen geçirilmesi gereklidir. Bu modellerin seçilmesindeki amaç, direkt olarak özetleme için geliştirilmiş bir dil modeliyle, daha farklı işler için de özelleştirilmiş dil modelinin performanslarını karşılaştırmakken; mT5 ise hem farklı dillerde çıktı üretme performansını değerlendirmek, hem de fine-tune edilmemiş LLM performansını değerlendirmek açısından projeye implemente edilmiştir.

Fine-tune'da kullanılmak üzere veriseti olarak CNN/DailyMail ve HuggingFace üzerinde açık kaynak olarak yayınlanan haber özetleme verisetlerinde karar verilmiştir.

Yapılan okumalarda, LLM'lerin özet çıkarma başarısının arkasındaki sebeplerin uzmanlar tarafından henüz tam anlamıyla anlaşılamadığı bilgisi edinilmiştir. Bu çalışmanın akademik ve endüstriyel anlamda bu soru işaretini ortadan kaldırırmaya katkı sunması hedeflenmiştir.

3

Fizibilite

İşbu bölümde projeye ait fizibilite araştırması incelenmektedir. Rapor kapsamında teknik, ekonomik, zaman ve yasal fizibileteye yer verilmektedir.

3.1 Teknik Fizibilite

Teknik fizibilite yazılım ve donanım olmak üzere ikiye ayrılarak incelenmektedir.

3.1.1 Donanım Fizibilitesi

Modellerin fine-tune işlemleri, lokal makinenin donanımsal eksiklerinden ötürü, Google Colaboratory üzerinden yapılmıştır. Bu noktada işlemci gücü olarak ilgili uygulamada bulunan 'A100' işlemcisi kullanılmış ve modellerin eğitimi için yaklaşık olarak 15.0 - 32.0 GB GPU kullanılmıştır.

3.1.2 Yazılım Fizibilitesi

Proje kapsamında uygulamanın tamamı Python yazılım dili kullanılarak gerçeklenmiştir. Python, yapay zeka ve LLM alanının omurgası olması, implementasyon kolaylığı sağlama ve gerekli platformlara erişebilme adına arayüz ve kütüphane sağlama konusunda avantajlı olduğundan dolayı doğal olarak tercih edilmiştir. Model eğitimlerinde PyTorch ve Transformers gibi kütüphanelerden yararlanılmış, haber çekiminde BeautifulSoup kullanılmış, arayüz gerçekleştirme esnasında ise streamlit kütüphanesi kullanılmıştır. Uygulama, streamlit arayüzü üzerinden canlıya alınmaya hazır şekilde tasarlanmış ve modellere erişim HuggingFace üzerinden bulutta gerçekleşmektedir.

3.2 Ekonomik Fizibilite

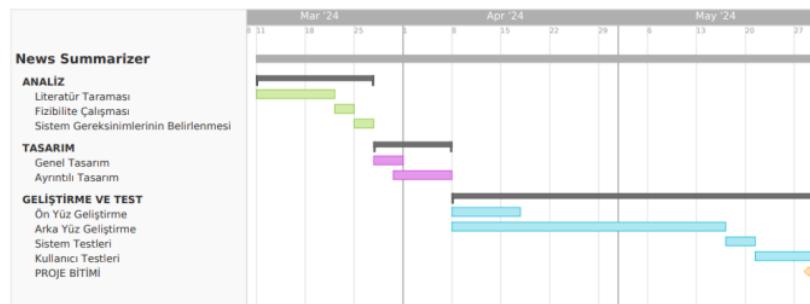
Projenin geliştirilme sürecinde gider olarak, lokal kısıtlardan ötürü kullanılan uygulamaların premium versiyonları kullanmak gereklidir. Bu noktada modellerin eğitimi için Google Colab Pro kullanılmış ve modellerin eğitilmesi için 'compute unit' için yaklaşık olarak 800 TL harcanmıştır. Bunun dışında elektrik ve internet erişimi için toplamda 100 TL ödendi.

İşbu proje sonucunda bir kar gözetilmemektedir, bir gelir beklenmemektedir. Proje tamamiyle gönüllü olarak kişisel araştırma ve gelişme amacıyla taşiyarak hazırlanmıştır.

3.3 Zaman Fizibilitesi

Projenin zaman çizelgesi, 10 Mart 2024 tarihinde başlamaktadır. Bu tarihte proje için gerekli araştırmalar yapmaya başlanmıştır. Gerekli araştırmanın yapılması ardından 15 Mart 2024 tarihinde yazılım süreci başlamıştır. 30 Mayıs 2024 tarihinde son halini almıştır.

Şekil 3.1'ta Gantt şemasına yer verilmiştir.



Şekil 3.1 Gantt Şeması

3.4 Yasal Fizibilite

Uygulamanın oluşturulması sürecinde yasalara aykırı olabilecek her türlü durumdan kaçınılmıştır. Proje kapsamındaki tüm parçalar orijinaldir. Projede yapılan çalışmalara tamamen gönüllülük esaslıdır ve bundan dolayı herhangi bir vergi durumu söz konusu değildir. Projede kullanılan verisetleri ve modeller açık kaynak olarak paylaşımı sunulduğu için üzerinde yapılacak işlemler herhangi yasal bir sıkıntıya yol açmamaktadır.

4 Sistem Analizi

Amaç: İşbu sistem analizi çalışması, bir haber özetleme uygulaması için gerekli olan bileşenleri ve bu bileşenlerin görevlerini tanımlamayı amaçlamaktadır.

4.1 Bileşenler

- Haber Çekici:** Özetleme işleminin üzerinde yapılabilmesi için uygulamaya girdi olarak verilebilecek, canlı olarak yayinallyıcı websitelerden elde edilebilecek bir haber havuzu gereklidir. Bu havuzun elde edilmesi noktasında girdi olarak kullanılması uygun metinlere sahip şekilde haberleri çekebilen bir uygulama gereklidir.
- Girdi Önleyici:** Elde bulunan haber metinleri, sisteme girdi olarak verilmeden önce işlenebilecek bir format haline gelmelidir. Böylelikle daha sağlıklı bir işleme yapılır.
- Özetleyici Model:** Haber özetleyici model, transformers mimarisi yardımıyla girilen haber metninin özettini tahmin eder.
- Arayüz:** Sistemin son bileşeni olarak, kullanıcı dostu olması ve kullanım kolaylığı sağlama açısından sistemin bir arayüzü olması gerekmektedir. Bu arayüz vasıtasıyla kullanıcıya haber ve özeti sunulur.

4.2 İşlevler

- Haber Çekici:** Gündelik haberlerin paylaşıldığı web sitelerine erişerek paylaşılmış haberlerin çekilmesini ve saklanması sağlar.
- Girdi Önleyici:** Elde bulunan haber metinlerini olmasının gereken format şeklinde hazırlar.

3. **Özetleyici Model:** Hazır veriseti üzerinde eğitilmiş olan model girilen girdinin özetini tahmin eder.
4. **Arayüz:** Sistemin kullanıma elverişli olması gerekliliğiyle kullanıcı dostu bir arayüze sahip olmalıdır. Bu arayüzde çekilen haberler kullanıcıya sunularak, özetlenmesi istenen haberi seçmeye olanak sağlama ve seçilen haberin özetini açıklayıcı şekilde sunması amaçlanmıştır.

Bu noktada haber özetleme sisteminin haber çekici, girdi önişleyici, model ve arayüz bileşenlerinden oluştuğunu söylemek mümkündür. Her bir bileşen uygulamanın sağlıklı çalışması için gereklidir. Bu sistem analizi çalışmasıyla gereken uygulama için bir çerçeve belirlenmiş ve sistemin oluşturulması aşamasında rehber görevi görmüştür.

5

Sistem Tasarımı

Uygulama kapsamında sistem yazılım, veritabanı tasarımları ve girdi-çıktı tasarımları yapılmış ve bu bölümde bu tasarımlar incelenmiştir.

5.1 Yazılım Tasarımı

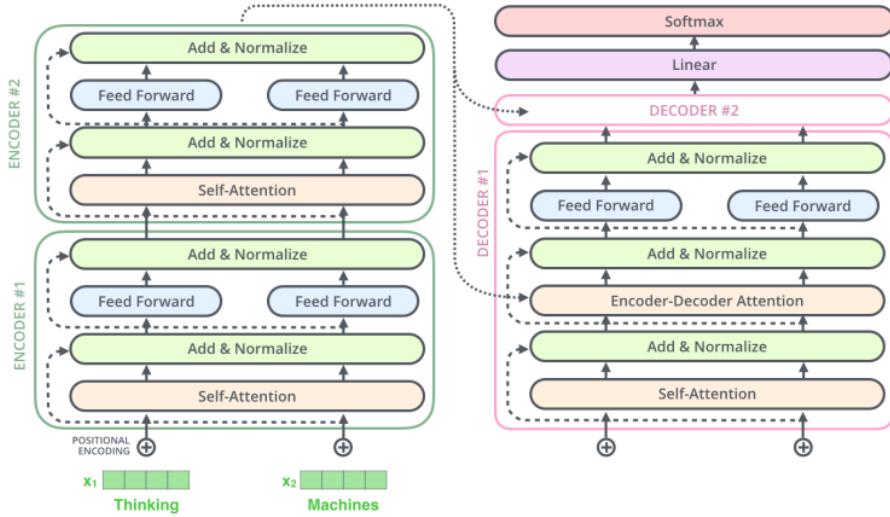
Uygulama kapsamında özetleme işlemi için İngilizce olarak T5 ve PEGASUS; Türkçe model olarak ise mT5 kullanılmasına karar verilmiştir. Aşağıda bu üç modelden detaylıca bahsedilmiştir.

5.1.1 T5, Mimarisi ve Eğitimi

Google Research tarafından geliştirilen T5, açılımı Text-to-Text Transfer Transformer olan, adında da geçtiği üzere Transformers mimarisi üzerine inşa edilmiş, doğal dil işlemeye özgü görevleri yerine getirebilen bir dil modelidir. Bu modelde doğal diller arası çeviri, metin üretme, cümlelerin birbirleri ile olan benzerliğinin tespiti gibi birçok görevi gerçekleştirmeye kapasitesine sahiptir. Yine adından da anlaşılabilcecgi üzere text-to-text özelliğine sahiptir, bu girdi olarak metin verisi alıp çıktı olarak da metin ürettiği anlamına gelir[4].

Yukarıda da belirtildiği üzere T5, Transformer mimarisi temellidir. Bu mimaride, girdinin farklı noktalarına 'dikkat' edilmesini sağlayan self-attention mekanizması kullanılmaktadır. Bu mekanizma ile yapay zeka modelleri metin gibi arka arkaya sıralı verileri sağlıklı bir şekilde işleyebilmektedir. T5 mimarisinin ana hatlarından bahsettiğimde bunlardan ilki Encoder-Decoder yapısıdır. Buradaki encoder, girdi metnini işleyerek 'gizli durum (hidden state)' dizisi haline getirir. Burada birden fazla Transformers bloğu bulunur ve her bloğun çok-başlı self-attention katmanı, ileri-besleme katmanı, fazladan bağlantı ve katman normalizasyonu bulunur. Decoder tarafı ise, bu 'gizli durum'lari kullanarak çıktı metnini oluşturur. Yine bu tarafta da birden fazla Transformers bloğu görev almaktadır, bunun dışında ayrıca çok-başlı

cross-attention mekanizmasıyla encoder'in oluşturduğu çıktıyı girdiye uygun şekilde analiz eder [4][7]. 5.1'de bu mimarinin görselleştirilmiş haline yer verilmiştir.

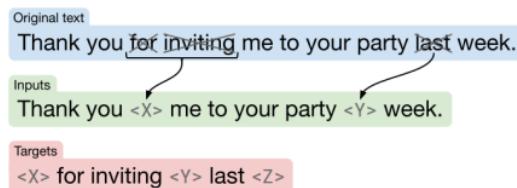


Şekil 5.1 Örnek Transformer Mimarisi

Her LLM gibi T5 de metin verisini direkt olarak işleyemez. Bu anlamda, metin verilerinin üstünde çeşitli işlemler yapılması gerekmektedir. Bu işlemlerden ilki 'tokenization' işlemidir. T5 için bu işlem denetimsiz metin tokenizer'ı olan SentencePiece kütüphanesi kullanılarak gerçekleştir. Genelgeçer olarak kullanılan, metnin boşluklar veya önceden belirlenen kurallara göre ayrıldığı tokenizer'lardan farklı olarak, SentencePiece metine karakter dizisi gibi davranışarak tamamıyla kelime veya kelime parçacıkları oluşacak şekilde ayırmaktadır [8].

T5 modelinin eğitiminde 'unsupervised objective' olarak 'Span Corruption' ve 'Text Infilling' kullanılmıştır. Span Corruption, arka arkaya gelen spesifik token'ların gizlenmesi akabinde modelin gizlenen kısımları tahmin etmesi veya oluşturması üzerine kurulu bir tekniktir. Bu tekninin uygulanabilmesi için eğitim sırasında önce girdi metinde bulunan veriler belli bir ölçüde özel tokenlar ile gizlenirler. Model, gizlenilmeyen kısımları inceleyerek gizlenen kısımları tahmin etmeye çalışır ve bu adım sayesinde bağlamı anlayarak ona uygun çıktıları üretmeyi öğrenir [4]. 5.3'de bu prosedürle alakalı örnek eklenmiştir.

Yukarda ilgili bilgilerin verildiği T5'in farklı model boyutları da mevcuttur. Small, Base, Large, 3B ve 11B olarak ayrılan bu modeller katman, attention-head sayısı ve gizli birimler konusunda birbirinden farklı konfigürasyonlara sahiptir [4].

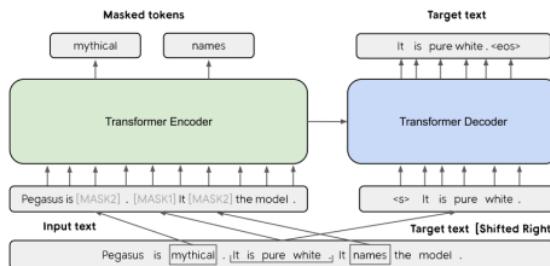


Şekil 5.2 T5 Eğitiminde yapılan 'gap' çalışması

5.1.2 PEGASUS, Mimarisi ve Eğitimi

Yine Google Research tarafından geliştirilen PEGASUS, Pre-training with Extracted Gap-sentences for Abstractive Summarization, özellikle metin özetleme için tasarlanmış bir büyük dil modelidir. PEGASUS da T5 modeli gibi standart Encoder-Decoder Transformers mimarisi üzerine kurulmuş bir modeldir. Özetteleme üzerine yoğunlaşan bir model olmasından dolayı doğal olarak text-to-text formatta çalışır. PEGASUS'un T5 ve diğer büyük dil modellerinden ayrıldığı nokta, ilk eğitim aşamasında direkt olarak özetteleme için eğitilmeye başlanmış olmasıdır. Bunu da kendisinin eğitimi için özel olarak geliştirilen GSG, Gap Sentences Generation ve MLM, Masked Language Modeling, tekniklerinin üzerinde uygulanmasıyla edilir [1].

Gap Sentences Generation, özetteleme işini 'taklit' eden bir konsepttir. Burada, girilen metin verisi içerisinde en önemli cümleleri gizleyerek (gap) yapar. 'Boşluk'larla doldurulmuş bu metin girdisi modele verilir ve model bu boşlukları tahmin etmek üzerine eğitilir. Burada amaç, modelin metinde bulunan bağlamı yakalaması ve buna uygun içerik üretmesini sağlayabilmektir. PEGASUS'un eğitilmesinde kullanılan bir diğer yöntem olan Masked Language Modeling oldukça yaygın kullanılan bir modeldir. GSG'ye benzer olarak, girdi olarak verilen metinlerdeki token'ların rastgele [MASK] token'ıyla değiştirilip, modele verilmesi şeklinde gerçekleştirilir. Bu yöntemle model kelimeler arasındaki ilişkileri kavrar ve daha geniş açıdan metini anlamlıdırabilir [1][9].



Şekil 5.3 PEGASUS Mimarisi ve Eğitim Tekniği

PEGASUS'un da T5'de olduğu gibi model boyutu olarak farklı türleri mevcuttur. Yayınlananlar arasında CNN/DailyMail verisetiyle fine-tune edilmiş Pegasus-CNN/DailyMail, XSUM verisetiyle eğitilmiş Pegasus-XSUM ve Pegasus-Large modeli mevcuttur. Bu modeller, parametre sayısı gibi birçok etkeden ötürü birbirlerinden farklıdır denebilir [1].

5.1.3 mT5, Mimarisi ve Eğitimi

Düzen modeller gibi yine Google Research tarafından geliştirilen mT5, Multilingual Text-to-Text Transfer Transformer, yukarıda detaylıca açıklanmış bulunan T5 modelinin Common Crawl tarafından sunulmuş, 101 farklı dilde web üzerinde bulunan metin içeriklerinin bulunduğu mC4 veriseti üzerinde eğitilmiştir [4][10]. T5.1.1'e ait tüm özelliklere sahip olan model, 101 farklı dil için de 'span corruption' teknğini kullanarak eğitilmiştir. Fakat model, bu pretraining işleminin ardından doğal dil işleme görevlerini yerine getirmek için fine-tune edilmemiştir, bu sebeple kullanıcıların modeli bir görev için implemente etmeden önce fine-tune etmesi gerekmektedir. [10]

mT5 de çeşitli boyutlarda kullanıma sunulmuştur. Bunlar küçükten büyüğe olmak üzere sırasıyla ⁶ mT5-Small, mT5-Base, mT5-Large, mT5-XL, mT5-XXL şeklindedir. 5 farklı şekilde bulunsa da, eğitildikleri verisetinin büyüklüğünden dolayı hepsinin 'normal' modellere kıyasla büyük olduğunu söylemek mümkündür.

5.2 Veritabanı Tasarımı

Projede T5 modeli için eğitim veriseti olarak CNN/DailyMail kullanılmıştır. CNN/DailyMail veriseti, soyut metin özetleme görevlerinde kullanılan 286.817 eğitim, 13.368 validasyon ve 11.487 test çiftinden oluşan, CNN ve DailyMail sitelerinden elde edilen haberlerle özetlerinin bulunduğu bir verisetidir [3]. Bu veri seti, içinde bulunan her veri için bir 'id', ana metinin tutulduğu bir 'article' ve son olarak da özeti bulunduğu 'highlight' değişkeni tutmaktadır. Özeti kısmı, direkt olarak verilen metnin özetini içermektedir. Şekil 5.4'de detaylı görsel eklenmiştir.

Pegasus modelinin eğitimi ise HuggingFace sitesinde 'therapara' kullanıcısı tarafından açık kaynaklı olarak sunulmuş haber özetleme veriseti olan 'summary-of-news-articles' kullanılmıştır. Pegasus modeli, LLM özetleme alanında oldukça yaygın kullanılan verisetleri üzerinde zaten zaten beslenmiş olduğundan dolayı bu veriseti seçilmiştir. Bu verisetinde 56.216 adet veri bulunmaktadır. Haberin ana metninin bulunduğu sütun 'document' ve bu haberlerin özetlerin bulunduğu sütün 'summary' olarak isimlendirilmiştir. Şekil 5.5'de detaylı görsel eklenmiştir. [11]

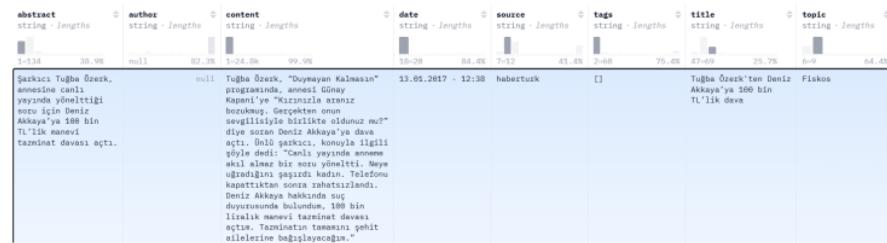


Şekil 5.4 CNN/DailyMail Veriseti



Şekil 5.5 'summary-of-news-articles' Veriseti

mT5 modeli, projede Türkçe olarak implemente edilen modeldir ve bu modelin eğitilmesinde buna bağlı olarak Türkçe haber özetleme veriseti kullanılmıştır. Bu veriseti de yine Huggingface sitesinde, 'batubayk' adlı kullanıcı tarafından açık kaynak olarak sunulan 'TR-News' verisetidir. Bu verisetinde toplamda 307.562 haber ve özetleri bulunmaktadır. Bu verisetinin bir diğer önemli özelliği ise diğer verisetlerine göre değişken yönünden daha zengin olmasıdır. Bu verisetinin 'abstract' kısmında haber özeti, 'author' kısmında metin yazarı, 'content' kısmında ana haber metini, 'date' kısmında haber tarihi, 'source' kısmında metinin çekildiği kaynak, 'tags' kısmında metine ait etiketler, 'title' kısmında haber başlığı, 'topic' kısmında haber konusu, 'url' kısmında habere ait link yer almaktadır. Şekil 5.6'de detaylı görsel eklenmiştir. [12]



Şekil 5.6 TR-News Veriseti

Projede, haber özetlemesinin yapılabilesi için internet üzerindeki haber sitelerinden veri kazıma işlemi yapılmaktadır. Bu işlem elde edilen haber linklerinin birer birer incelenip haberlerden uygulama için gerekli olan bilgilerin çekilmesiyle tamamlanır. Uygulama için haberlerin 'başlık' ve 'ana içerik' kısımlarına ihtiyaç vardır. Veri kazıma işlemiyle bu bilgiler elde edildikten sonra 'başlık', 'link' ve 'icerik' olmak üzere bir Python kütüphanesi olan Pandas'a ait 'dataframe' veri yapısında tutulur. Bu bilgileri tutma işlemi, o anki koşuma aittir; haberler bir kere tutulduktan sonra herhangi bir şekilde kaydedilip, saklanmazlar. Uygulamanın yenilendiği bir durumda, eski haberlere ait dataframe'ler yeniden haber çekimi yapıldığı için yeniden oluşturulur. Uygulama kapatıldığında çekilen haber verileri silinir. Haberlerin çalışma ortamında tutulmalarına dair temsili bir görsel Şekil 5.7'de verilmiştir.

```

headline
0 Famous scenic waterfall in China goes viral af...
1 World's best restaurant for 2024 revealed
2 Boeing Starliner's crew is now on the space sta...
3 Novak Djokovic wants to return to tennis 'as so...
4 North Korean trash balloons reach French coast
5 Biden heads to a poignant moment of American her...
6 Biden's D-Day visit may mark the end of an Ameri...
7 Biden defends democracy in Europe while Trump ...
8 World War II veteran Robert Persichitti dies a...
9 CNN asks Tom Hanks if he is worried about anot...
link
0 https://edition.cnn.com/2024/05/19/travel/china-waterfall-intl-hnk
1 https://edition.cnn.com/travel/article/worlds-best-restaurants-2024/index.html
2 https://edition.cnn.com/2024/06/06/us/boeing-starliner-space-station/index.html
3 https://edition.cnn.com/2024/06/06/europe/north-korea-balloon-fight/index.html
4 https://edition.cnn.com/2024/06/06/europe/north-korea-balloon-fight/index.html
5 https://edition.cnn.com/2024/06/07/politics/joe-bidens-d-day-visit/index.html
6 https://edition.cnn.com/2024/06/06/politics/bidens-d-day-visit/index.html
7 https://edition.cnn.com/2024/06/07/politics/bidens-d-day-visit/index.html
8 https://edition.cnn.com/2024/06/06/us/will-veteran-robert-persichitti-dies/index.html
9 https://edition.cnn.com/2024/06/06/world/video/cnn-asked-tom-hanks-if-he-is-worried-about-another-world-war/index.html

text
0 A famous waterfall in China has gone viral. The Yuntai...
1 And don't miss Spain's remains the finest in the world.
2 Boeing's Starliner mission has safely docked with the ISS.
3 Djokovic struggled with pain in his right knee during the tournament.
4 South Korean activists sent balloons carrying anti-nuclear messages across the border.
5 President Joe Biden is set to present a case for climate action at the G7 summit.
6 The new world for which the greatest generation will be remembered.
7 The former president is making his 2024 opposition campaign stops.
8 Robert Persichitti, a 102-year-old World War II veteran, has died.
9 CNN asked Tom Hanks if he is worried about another world war.

```

Şekil 5.7 Haberlerin Tutulmasına İlişkin Temsili Görsel

Çekilen haberler kullanıcıya gösterilebilir olacağı için, HTML kodundan kaynaklı metin karışıklığı yaşanmaması adına metinler çekim esnasında temizlenmektedir.

5.3 Girdi - Çıktı Tasarımı

İşbu tasarımda girdi olarak veri çekimi sonucu elde ettiğimiz haberler kullanılmıştır. Pandas Dataframe veriyapısında saklanan haberlerin 'icerik' kısmında bulunan metinler modele verilerek özetlenmesi sağlanır. Bu noktada kullanıcı herhangi bir haber metni giremeyeceği gibi, elde edilen haber metnini de düzenleyemez. Buna bağlı olarak kullanıcının girdi tarafında bir müdahalesinin olması, uygulama kapsamında, mümkün değildir. Kullanıcı arayüz yardımıyla haberlerin başlıklarını görür ve haber özetleme butonuna tıklayarak istediği haberin özetinin oluşturulmasını sağlayabilir. Bu aşamada model aracılığıyla ilgili haberin metnine ulaşılır ve model bu

haber metni ile beslenerek bir çıktı üretmesi beklenir. Elde edilen çıktı ekrana verilerek kullanıcıya sunulur.

Kullanıcının girdiye bir müdahalesi olmadığından dolayiuygun olmayan metin girilmesi gibi sorun yaratabilecek durumların kontrolü sağlanmamaktadır. Bunun dışında girdide olabilecek hatalar, modele verilmeden önce geçirildiği önişleme fonksiyonunda olabildiğince giderilmeye çalışılmıştır.

6

Uygulama

Bu bölümde projede istenen görevi yerine getirmesi için geliştirilmiş uygulama ve detayları, geliştirilme aşamasına da dephinerek açıklanmıştır.

6.1 LLM Modellerinin Fine-tuning İşleminden Geçirilmesi

Projeye entegre etmek için 3 adet model fine-tune edilmiştir. Bunlar sırasıyla, T5, Pegasus ve mT5 olmuştur. T5 ve Pegasus modelleri İngilizce veriyle çalışırken, mT5 Türkçe veri üzerinde çalışmaktadır.

6.1.1 Fine-tuning İşlemleri

Modellerin fine-tuning işlemleri lokal yetersizliklerden ötürü Google Colab üzerinden yapılmıştır. Google Colab ortamının sağladığı sanal işlemcilerden biri olan A100 çalışma ortamı proje için seçilmiştir. Google Colab Pro versiyonuyla kullanılabilen bu ortamda, 40 GB GPU, 200 GB disk alanı ve 85GB Sistem RAM'ile çalışma yapma fırsatı bulunmaktadır.

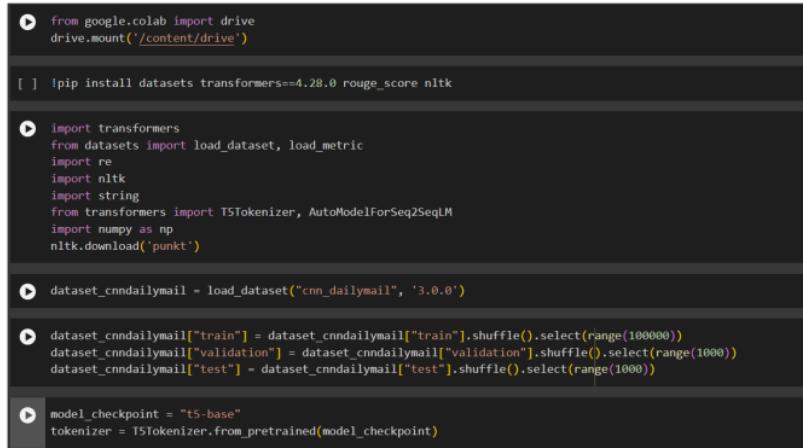
Veriseti olarak, haber özetleme fine-tuning işlemlerinde yaygın olarak CNN/Daily Mail [3] ve XSUM [2] kullanıldığı görülmüştür. XSUM verisetinde bulunan özetler, insanlar tarafından modeller tarafından oluşturulan özetlerden 'daha kötü' olarak nitelendirildiği için [2] eğitimde CNN/DailyMail veriseti kullanılmıştır. T5 verisetinde eğitim için bu verisetinin 100.000 adet verisi modele girdi olarak verilmiştir, Pegasus Large ise 'summary-of-news-articles' ile 50.000 adet örnek ile eğitilmiştir. Türkçe model mT5 için 'TRNews' veriseti kullanılmıştır.

Verisetleri Python'ın 'datasets' kütüphanesi üzerinden çekilmiştir, modeller ise 'transformers' kütüphanesi aracılığıyla HuggingFace'den runtime'a alınmıştır. Finetuning işlemleri için derin öğrenme kütüphanesi olarak PyTorch tercih edilmiştir.

6.1.1.1 T5 Modeli İçin Fine-tuning İşleminin Yapılması

Uygulama kapsamında özetleme için kullanılan LLM'lerden birisi T5-BASE olarak belirlenmiştir. T5-BASE, metin girdisi verilerek, bir metin çıktıları aldığı text-2-text büyük dil modelidir. Modelin eğitimi yüksek işlemci gücü isterken, aynı zamanda uzun süre almaktadır. Lokal kısıtlardan ötürü kısıtlı sayılabilen ölçüde veriseti ve koşullar üzerinde fine-tuning işlemi yapılmıştır.

Bu noktada bahsedilen işlemler aşağıda Şekil 6.1'deki gibi sağlanmıştır.



```
from google.colab import drive
drive.mount('/content/drive')

[ ] !pip install datasets transformers==4.28.0 rouge_score nltk

▶ import transformers
from datasets import load_dataset, load_metric
import re
import nltk
import string
from transformers import T5Tokenizer, AutoModelForSeq2SeqLM
import numpy as np
nltk.download('punkt')

▶ dataset_cnnndailymail = load_dataset("cnn_dailymail", '3.0.0')

▶ dataset_cnnndailymail["train"] = dataset_cnnndailymail["train"].shuffle().select(range(100000))
dataset_cnnndailymail["validation"] = dataset_cnnndailymail["validation"].shuffle().select(range(1000))
dataset_cnnndailymail["test"] = dataset_cnnndailymail["test"].shuffle().select(range(1000))

▶ model_checkpoint = "t5-base"
tokenizer = T5Tokenizer.from_pretrained(model_checkpoint)
```

Şekil 6.1 Veriseti ve Modelin Yüklenmesi

Fine-tuning işleminin sağılıklı bir şekilde yapılabilmesi için modele girdi olarak verilen metinlerin belli işlemlerden geçmesi gerekmektedir. Bu işlemler, önişleme ve metinleri tokenize etme olarak aşalamalıdır. Önişleme adımda, Python'ın 'nltk' ve RegEX kütüphanesi 're'den faydalanyılmıştır. Bu adımda metinlerde bulunan boşluklar ve boş cümleler çıkartılarak bir düzenleme işlemi yapılmaktadır. Tokenize etme adımda ise model ait tokenizör kullanılarak metinler vektöre edilmektedir. Burada 'max-input-length' değeri 1024 iken, "max-output-size" değeri 128 olarak belirlenmiştir. T5 modelinin özetleme işlemini gerçekleştirmesi için girdi olarak verilen metinlerin başında 'summarize:' ifadesinin bulunması gereklili olduğundan [4], bu bilgi de eklenerek metinler önişleme aşamasını tamamlamaktadır. İlgili adımlar Şekil 6.2'de olarak paylaşılmıştır.

Fine-tuning işlemi, modele verilen verisetinin yanında, büyük ölçüde hiperparametre olarak girilen değerleri de kapsamaktadır. Bu noktada, T5 modelinin eğitiminde çeşitli hiperparametrelere uygun görülen değerler atanarak fine-tuning sırasında kullanılmıştır. Eğitim boyunca her 100 adımda 250 validasyon verisi üzerinden eğitimin başarısı ve loss fonksiyonu hesaplanmıştır.

```

prefix = "summarize: "
max_input_length = 1024
max_target_length = 128

def clean_text(text):
    re.sub("http[s]?://[\S+]", "", text)
    sentences = nltk.sent_tokenize(text.strip())
    sentences_cleaned = [s for sent in sentences for s in sent.split('\n')]
    sentences_cleaned_ = [sent for sent in sentences_cleaned if len(sent) > 0 and sent[-1] in string.punctuation]
    text_cleaned = '\n'.join(sentences_cleaned_)
    return text_cleaned

def preprocess_data(examples):
    text_cleaned = [clean_text(text) for text in examples['article']]
    inputs = [prefix + text for text in text_cleaned]
    model_inputs = tokenizer(inputs, max_length = max_input_length)

    with tokenizer.as_target_tokenizer():
        labels = tokenizer(examples['highlights'], max_length=max_target_length)

    model_inputs['labels'] = labels['input_ids']
    return model_inputs

```

Şekil 6.2 Verisetinin Önislemeden Geçirilmesi

Modelin eğitimi sırasında elde edilen başarı skorlarına **Şekil 6.3**'de yer verilmiştir.

Step	Training Loss	Validation Loss	Rouge1	Rouge2	Rougel	Rougelsum
100	1.582800	1.509631	24.857900	11.768600	20.290100	23.272300
200	1.549300	1.504436	25.019300	12.099200	20.530100	23.448800
300	1.540200	1.501544	24.362500	11.594700	20.170300	22.906700
400	1.545700	1.504584	24.712400	11.879000	20.368600	23.103700
500	1.540000	1.502680	24.475500	11.548300	20.037300	22.934500
600	1.580600	1.497770	24.210600	11.584300	19.976000	22.769200
700	1.581200	1.501152	24.905700	12.078800	20.492300	23.377400
800	1.573200	1.498986	24.961200	11.994200	20.477100	23.372500
900	1.547700	1.492378	24.899400	12.144100	20.534500	23.445500
1000	1.592700	1.491565	25.036500	12.013300	20.549300	23.491300
1100	1.577100	1.496595	25.006900	12.012900	20.619400	23.528700

Şekil 6.3 Eğitimde elde edilen başarı skorları

Eğitim, 100 adımda başarı skorları da hesaplanarak toplamda 15700 adımda tamamlanmıştır. Bunlardan en başarılı sonuçları veren 10200. adımda bulunan model, projeye dahil edilecek model olarak seçilmiştir.

6.1.1.2 Pegasus Modeli İçin Fine-tuning İşleminin Yapılması

Uygulamada implemente edilecek ikinci model olarak PEGASUS-Large modeli seçilmiştir. Pegasus modeli de, T5 gibi, metin girdisi verildiğinde metin çıktısı vermek üzere tasarlanmış bir dil modelidir. T5-Base modeline göre daha büyük olan bu modelde kısıtlardan ötürü ve modelin özetleme için eğitilmesi de göz önüne alınarak daha küçük bir verisetiyle eğitilmiştir.

Pegasus-Large modeli, pretrain esnasında özetleme görevi üzerine eğitilmesinin

yanı sıra, alanında oldukça yaygın özetleme verisetleri üzerine fine-tune edilmiştir. T5'in eğitiminde kullanılan CNN/DailyMail veriseti de bunlara dahil olduğundan dolayı, Pegasus modeli için 'summary-of-news-articles' veriseti kullanılmıştır. Burada T5'de kullanılan aynı pipeline implemente edilerek, datasets kütüphanesi ile veriseti çekilmiş, HuggingFace üzerinden modele erişim sağlanmış ve eğitime hazır hale getirilmiştir. Modele beslenecek olan metin verileri, T5'de olduğu gibi, önişlemeden geçirilmiş ve temizlenmiştir.

Pegasus'un fine-tune'u 50.000 haber verisiyle sağlanmıştır. 'max-input-size' değeri 1024, 'max-output-size' değeri 128 token olarak belirlenmiştir. T5'de yapılan validasyon hesaplaması burada 100 adımda değil, 500 adımda yapılmaktadır. Burada fine-tune toplamda 8430 adımda tamamlanmıştır. Projeye entegre edilen adım, en yüksek ROUGE skoruyla 6500 olmuştur. Şekil 6.4'de Pegasus modelinin eğitim skorlarına yer verilmiştir.

3500	1.931700	1.990895	41.752600	16.070700	24.420600	36.546000
4000	1.944600	1.984212	42.239600	15.935300	24.458400	36.937800
4500	1.942600	1.975534	42.001800	15.844300	24.245000	36.645000
5000	1.935300	1.968861	42.327600	16.074200	24.398700	37.009700
5500	1.916900	1.964551	42.546300	16.127800	24.511400	37.309500
6000	1.781800	1.976481	42.747000	16.382700	24.687900	37.388800
6500	1.795200	1.973148	42.938800	16.573600	24.844800	37.588100

Şekil 6.4 Pegasus ile elde edilen başarı skorları

6.1.1.3 mT5 Modeli için Fine-tuning İşleminin Yapılması

Uygulamanın Türkçe kısmı için kullanılacak olan LLM, mT5'tir. mT5'in, T5 modelinin 101 farklı dil üzerinde eğitildiği farklı dillere hakim bir versiyonu olduğunu söylemek mümkündür. Proje kapsamındaibu modelin mT5-Small versiyonu implemente edilmesine karar verilmiştir. Seçilmesinin arkasındaki ana sebepler arasında, T5 mimarisine sahip olmasıyla beraber, modelin doğal dil işleme görevleri üzerine herhangi bir eğitimi olmamasından dolayı fine-tune sonucunda ne kadar iyileşebileceğini gözlemleyebilmek mevcuttur.

Seçilen model, Türkçe haber içerikleri ve özetlerinin bulunduğu bir veriseti olan 'TRNews' verisetinde 50.000 veri ile Türkçe haber özetlemesi için fine-tune edilmiştir. Verisetinde bulunan metinler, diğer modellerde olduğu gibi bu aşamada da önişleme aşamalarından geçirilmiş ve temizlenmiştir. Bu model için de max-input-size değeri 1024 olarak; max-output-size değeri 128 olarak belirlenmiştir. T5 modelinde olduğu gibi mT5 modeli için de metinlerin başında 'summarize' ifadesinin bulunması

gereklidir. Modele verilen bu komut, girdi ve çıktı dili üzerinde herhangi bir etki yapmamaktadır.

Modelin finetune işlemi sırasında kullanılması istenen hiperparametre değerleri girildikten sonra, modelin fine-tune işlemi yapılmaya başlanmıştır. Fine-tuning işleminin ilk adımlarında, modelin istenen görevde ne kadar yakınlaştığını takip edebilmek için kısa zaman aralıklarında validasyon hesaplaması yapılması sağlanmıştır. Şekil 6.5 bu işlemlere ait görsel verilmiştir.

Step	Training Loss	Validation Loss	Rouge1	Rouge2	Rougel	Rougelsum
50	No log	8.016172	3.909500	1.296900	3.640600	3.679900
100	No log	6.030367	6.255000	2.249700	5.687600	5.860900
150	No log	5.131608	9.222600	3.508500	8.235300	8.528900
200	No log	3.743767	12.114800	4.613100	10.674000	11.260200
250	No log	3.114571	13.881900	5.617700	12.022800	12.804100
300	12.001300	2.816550	15.791400	6.998000	13.483800	14.478200
350	12.001300	2.700572	14.871900	6.651100	13.137200	13.653500
400	12.001300	2.693258	14.392900	6.542100	12.893400	13.150600
450	12.001300	2.693939	15.387100	7.354400	13.793500	14.003300
500	12.001300	2.693285	15.571600	7.626700	13.966400	14.177400
550	12.001300	2.652911	15.957300	7.816600	14.250700	14.474000

Şekil 6.5 mT5 için İlk Başarım Sonuçları

Bu hesaplamalar arasında ROUGE-1 skoruyla görüldüğü üzere model her 50 batch işleyişinde istenen görevi daha başarılı şekilde yapmaktadır.

Modelin fine-tuning işlemi, önce 50 ve ardından 500 adımda validasyon değerleri hesaplanarak toplamda 9375 adımda tamamlanmıştır.

7500	2.819600	2.189778	24.080000	14.090200	21.585300	21.849700
8000	2.804300	2.182536	24.180600	14.211300	21.722800	22.006100
8500	2.810300	2.178463	24.491800	14.381500	21.978800	22.259800
9000	2.781100	2.181017	24.437200	14.314900	21.911100	22.203400

Şekil 6.6 mT5 için son Başarım Sonuçları

6.2 Modellerin HuggingFace Platformunda Tutulması

Elde edilen son modeller, erişilebilir olması ve arayüz noktasında kolaylık sağlama açısından HuggingFace platformuna kütüphane olarak yüklenmiştir. Burada model ile beraber elde edilen son tokenizer değerleri de kütüphaneye eklenmiştir. Bu işleme dair detaylar Şekil 6.7'te bulunmaktadır. Örnek olarak T5 modelinin HuggingFace ortamında tutulmasına şekil 6.8'de yer verilmiştir.

```

from transformers import T5ForConditionalGeneration, AutoTokenizer
import torch

model_name = "t5-base-ft/checkpoint-10200"
model_dir = f"drive/MyDrive/Models/{model_name}"

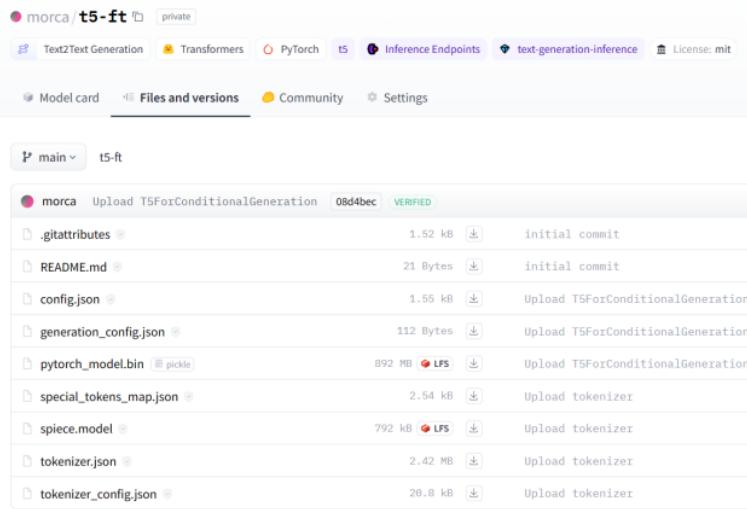
tokenizer = AutoTokenizer.from_pretrained(model_dir, legacy=False)
pt_model = T5ForConditionalGeneration.from_pretrained(model_dir)

model_name = 't5-ft'

tokenizer.push_to_hub(model_name)
pt_model.push_to_hub(model_name, safe_serialization=False)

```

Şekil 6.7 Modelin HuggingFace ortamına yüklenmesi



Şekil 6.8 Modele ait HuggingFace Kütüphanesi

T5 modeli 'morca/t5-ft', mT5 modeli 'morca/mt5-tr-ft' ve Pegasus modeli de 'morca/pegasus-l-ft' adresinde tutulmaktadır.

6.3 Haber Sitelerinden Canlı Haber Çekimi

Haber sitelerinden canlı haber çekimi işlemi 'web scraping' yoluya ile mümkündür. Bu işlem, çeşitli Python kütüphaneleri aracılığıyla haberlerin yayınlandığı sitelere erişerek bulunan içeriğin kod yardımıyla çekilmesini içermektedir. Bu işlem için projede kullanılmak üzere, kullanışlı ve entegre edilmesinin kolay olması açısından, 'BeautifulSoup' ve 'lxml' kütüphanesi kullanılmasına karar verilmiştir. Bu kararın alınmasındaki bir diğer etken ise, 'Selenium' kütüphanesi gibi web tarayıcılarına ihtiyaç duymadan, tamamen Python kütüphaneleri üzerinden bu işi gerçekleştirebilmesidir.

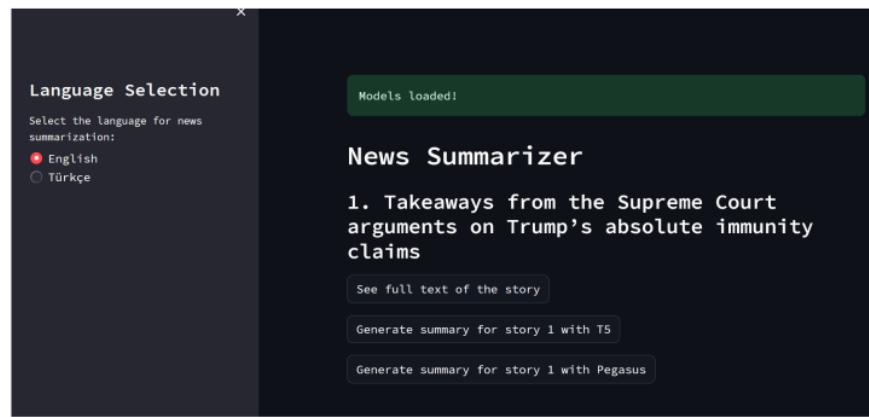
Böylelikle veri kazıma işlemi bulut ortamlarda da sorunsuz çalışabilir.

Proje kapsamında İngilizce haberler için CNN sitesi, Türkçe haberler için ise 'TRT Haber' sitesi veri kazıma işlemine tabii tutulmakta ve haberler çekilmektedir.

6.4 Arayüzün Gerçeklenmesi

Arayüzün gerçeklenmesi noktasında Streamlit kullanılması kararlaştırılmıştır. Streamlit, kullanıcı dostu, implemente edilmesi kolay ve ücretsiz canlıya alma gibi opsiyonlardan dolayı oldukça yaygın kullanılan bir arayüz geliştirme kütüphanesidir. Avantajları arasında HuggingFace üzerinde bulunan kütüphaneye erişerek modeli 'continuous deployment' mantığıyla arayüze direkt olarak entegre etmeye olanak sağlamaşı da vardır. Böylelikle, kullanıcı hiçbir zorluk yaşamadan sadece web uygulaması üzerinden modele erişerek haber özetleme işlemini sağlayabilir. Tüm bu sebepleri göz önünde bulundurarak, düşük maliyetli ve endüstri standartlarına uygun olarak Streamlit ile arayüz sağlanmıştır [13].

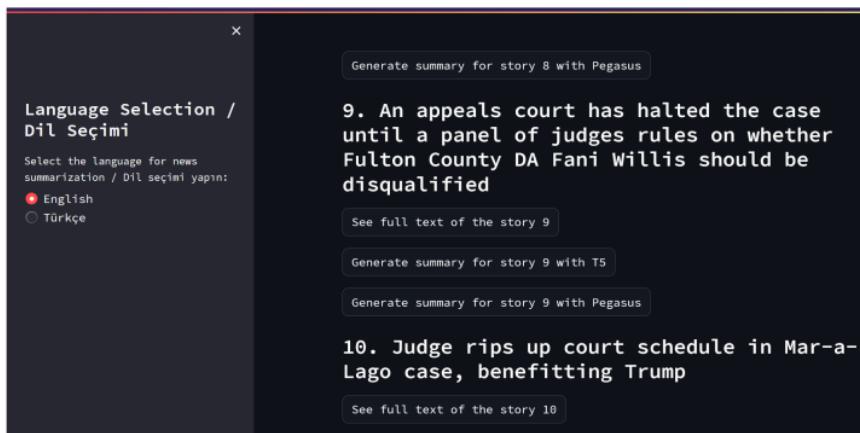
Arayüz ve modelin birbirine entegre şekilde doğru ve sağlıklı olarak çalışması uygulamanın işleyişi açısından oldukça önemli bir faktördür. Arayüzün gerçeklenmesi işlemi lokal ortamda Python ile sağlanmıştır. Burada, IDE üzerinden Streamlit kütüphanesi yardımıyla arayüz tasarlaması yapılmıştır. 'transformers' kütüphanesi ile modeller HuggingFace bulut ortamından çekilerek modelde kullanıma hazır hale getirilmiştir. Tasarımda kullanıcıya yan menü üzerinden dil seçebilme fırsatı sunulur, kullanıcı dili seçer. Seçtiği dile göre çekilen haberlere sırasıyla sayı atanmaktadır. Haber başlıklarını gösterilerek kullanıcılarla basmaları durumunda T5, PEGASUS veya Türkçe kısmında mT5 modeliyle özet çıkarma işlemini yapacak butonlar eklenmiştir. Kullanıcının butonlardan birisine basması durumunda ilgili modelin özetleme fonksiyonu çağrılarak, seçilen haberin ana metni fonksiyona gönderilir ve modele beslenerek bir özet oluşturulması sağlanır. Tüm metni görmek isteyebilecek kullanıcılar için böyle bir opsiyon da eklenmiştir. Şekil 6.9'da arayüz tasarımını görmek mümkündür.



Şekil 6.9 Arayüz

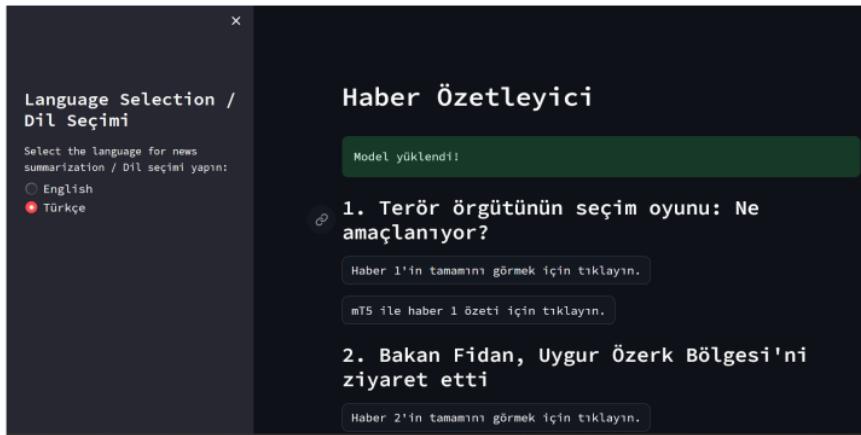
7 Deneysel Sonuçlar

Bu kısımda web arayüzü, haber çekimi ve modellerin özetleme yapması üzerinde çeşitli deneyler yapılmıştır. Aşağıda bulunan Şekil 7.1'de İngilizce taraf için, Şekil 7.2'de Türkçe taraf için arayüz örnekleri verilmiştir.



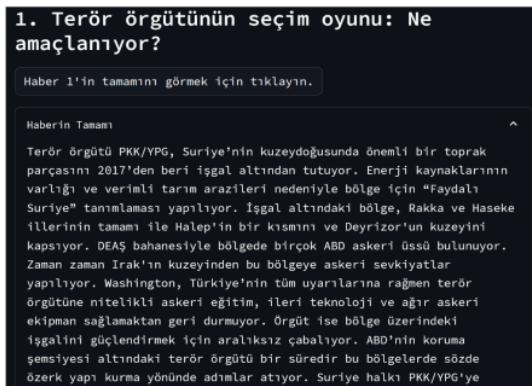
Şekil 7.1 İngilizce Opsiyonu için Arayüz

Web arayüzüne ilk girişte öncelikli olarak, yan menüde yapılmış dil tercihine göre, ilgili dilin özetini çıkaracak modellerin başarılı şekilde uygulamaya yüklenmesi beklenir. Bu işlem basit olarak, uygulamaların HuggingFace ortamından çekilmesi olarak açıklanabilir. Model yüklemesinin ardından haber çekimi işlemi başlar. Ekranda haber başlıklarının belirmesiyle haber çekme işlemi tamamlanmış olur. Bu noktada, kullanıcı dilediği haber başlığına ait haberin özeti veya tamamını görebilmek için özetleme veya haber metninin tamamını görme tuşlarını kullanabilir. Bu işlemler İngilizce opsiyonu için 'See full text of the story [habere ait numara]', 'Generate summary for the story [habere ait numara] with T5', 'Generate summary for the story [habere ait numara] with Pegasus' tuşlarıyla gerçekleştirilir. Türkçe opsiyonunda ise kullanıcının 'Haber [habere ait numara]'in tamamını görmek için tıklayın' veya 'mT5 ile haber özeti için tıklayın' tuşuna basması beklenir.

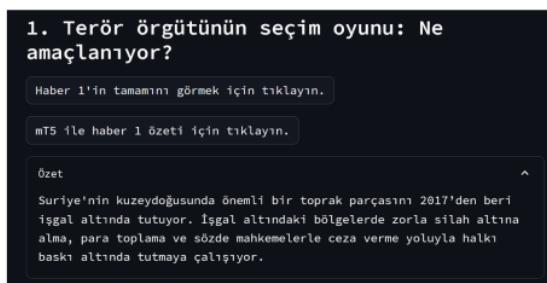


Şekil 7.2 Türkçe Opsiyonu için Arayüz

Örneklerde hem Türkçe, hem İngilizce için modeller denenmiş ve aşağıda görsel olarak verilmiştir. Haberlerin tamamının görüntülenmesi konusunda yapılan deneylere yer verilmiş ve hemen altlarında modellerin bu haberler için elde ettiği özetler eklenmiştir.



Şekil 7.3 1- Türkçe bir haberin tamamının görüntülenmesi



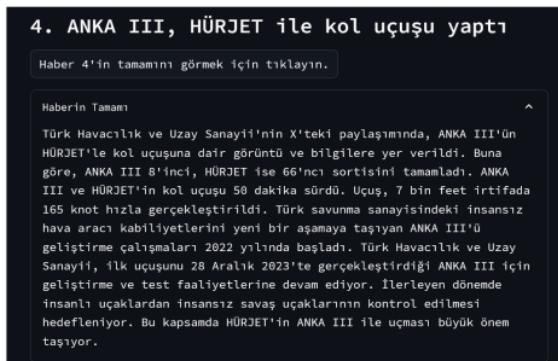
Şekil 7.4 1- mT5 sonucu elde edilen özet



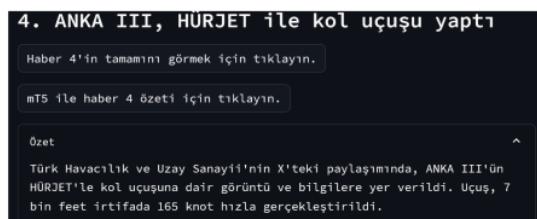
Şekil 7.5 2- Türkçe bir haberin tamamının görüntülenmesi



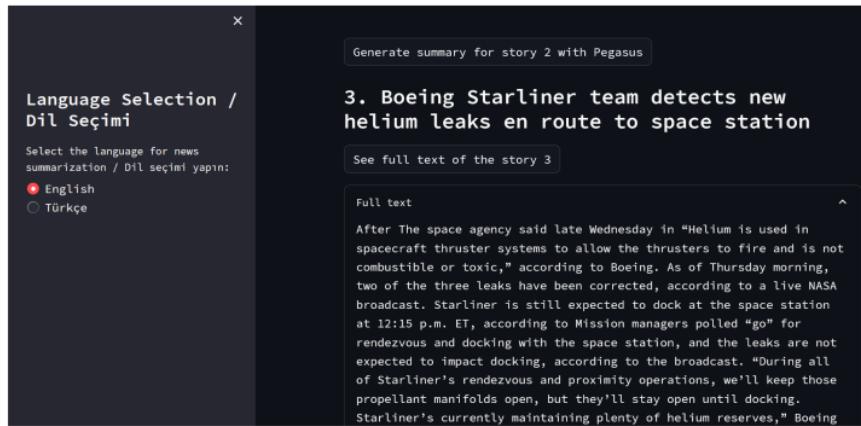
Şekil 7.6 2- mT5 sonucu elde edilen özet



Şekil 7.7 3- Türkçe bir haberin tamamının görüntülenmesi



Şekil 7.8 3- mT5 sonucu elde edilen özet



Şekil 7.9 1- İngilizce bir haberin tamamının görüntülenmesi



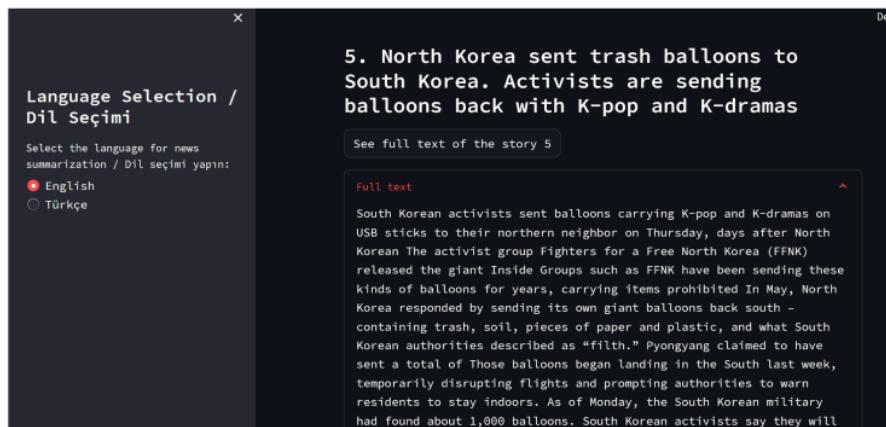
Şekil 7.10 1- T5 sonucu elde edilen özeti



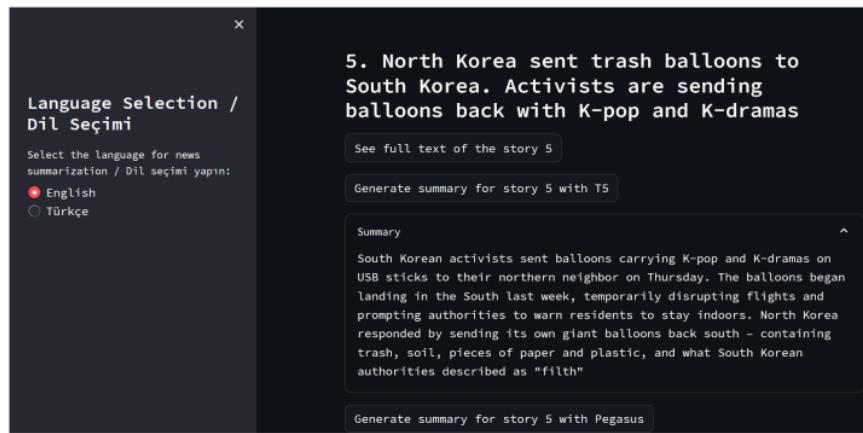
Şekil 7.11 2- İngilizce bir haberin tamamının görüntülenmesi



Şekil 7.12 2- T5 sonucu elde edilen özet



Şekil 7.13 3- İngilizce bir haberin tamamının görüntülenmesi



Şekil 7.14 3- T5 sonucu elde edilen özet

2. World's best restaurant for 2024 revealed

[See full text of the story 2](#)

[Full text](#)

Any doubts that Spain remains the fine dining center of the world may just have evaporated with the revealing of the Six eateries from the European nation made the list – three of them in the top five. The awards – considered the Oscars of global fine dining – were handed out at a ceremony at the Wynn in Las Vegas on Wednesday evening, with Barcelona's El Bulli should provide a few clues as to what diners can expect – imaginative and playful dishes executed with technical mastery, such as the caviar-filled Panchino doughnut, the frozen gazpacho sandwich and squab with kombu spaghetti, almond and grape. Coming in at number two was But you can never count Paris out. The French capital's Coming in at fifth was Lima's Securing six restaurants on the list was enough to make Spain the biggest winner.

Şekil 7.15 1- İngilizce bir haberin tamamının görüntülenmesi

2. World's best restaurant for 2024 revealed

[See full text of the story 2](#)

[Generate summary for story 2 with T5](#)

[Generate summary for story 2 with Pegasus](#)

[Summary](#)

Three of Spain's top five restaurants made the list. The awards were handed out at a ceremony at the Wynn in Las Vegas on Wednesday. In terms of cities, Paris and Bangkok tied for first. New York City's Three restaurants from Tokyo were awarded, with eatery Hong Kong saw two of its restaurants hit the list, with both making impressive moves.

Şekil 7.16 1- Pegasus sonucu elde edilen özet

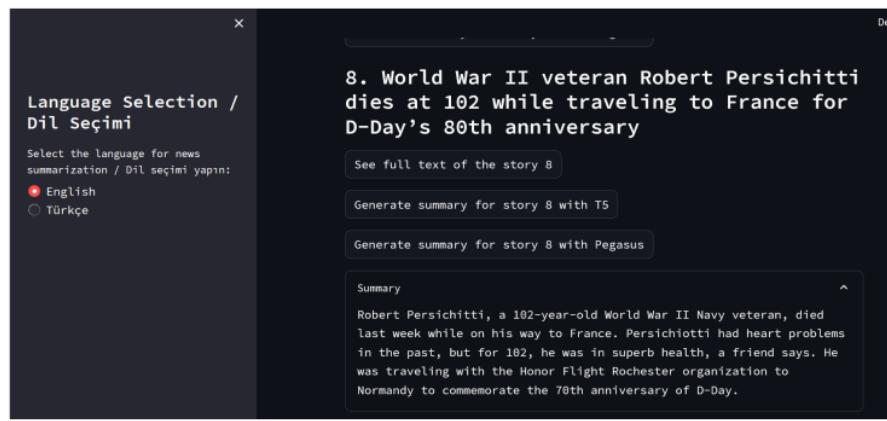
8. World War II veteran Robert Persichitti dies at 102 while traveling to France for D-Day's 80th anniversary

[See full text of the story 8](#)

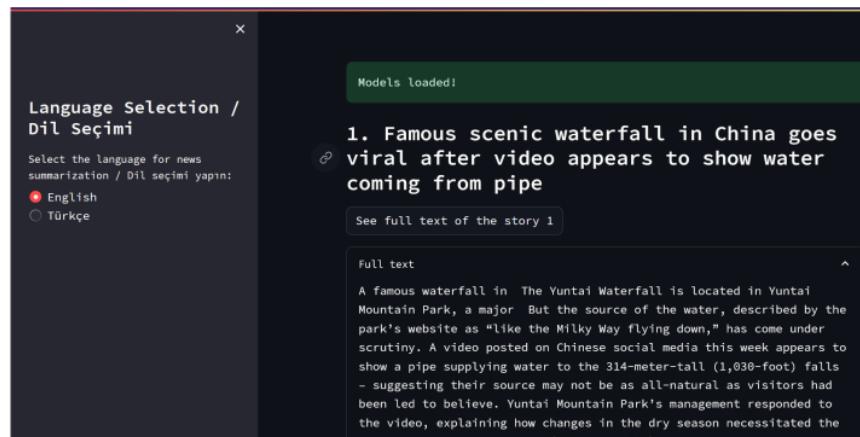
[Full text](#)

Robert Persichitti, a 102-year-old World War II US Navy veteran, died last week while on his way to France to commemorate Persichitti was a "wonderful, pleasant, humble guy," who was "easy know, easy to talk to," said Honor Flight Rochester President and CEO Richard Stewart, who told CNN he learned of his friend's death last Friday. "We miss him," said Stewart. While Persichitti passed away bound for Normandy – where the Allied forces' Persichitti fell ill last week during a stop in Germany while headed for Normandy, Al DeCarlo, a friend who was traveling with Persichitti, told "The doctor was with him. He was not alone, he was at peace and he was comfortable," DeCarlo said. "She put his favorite singer, Frank Sinatra, on her phone and he peacefully left us." Persichitti had heart problems in the past, "but

Şekil 7.17 2- İngilizce bir haberin tamamının görüntülenmesi



Şekil 7.18 2- Pegasus sonucu elde edilen özet



Şekil 7.19 3- İngilizce bir haberin tamamının görüntülenmesi



Şekil 7.20 3- Pegasus sonucu elde edilen özet

8

Performans Analizi

Performans analizi noktasında, proje için değerlendirmeye açık olan unsurlar fine-tune işleminden geçirilmiş modellerin başarısıdır. Burada modellerin başarısının değerlendirilmesi noktasında, LLM'lerin eğitilmek ve belli bir görev için fine-tune edilmek için oldukça yüksek GPU ve genellikle yüksek miktarda veriye sahip olması gerekmesi önemli bir husustur. Projede model eğitimi için gerekli güç oldukça kısıtlıdır. Daha yüksek veri sayısı ve işlemci gücüyle proje kapsamında elde edilen değerlerden daha 'iyi' sayılabilen sonuçlar elde etmek elbette ki mümkün değildir. Fakat projede implemente edilen modellerin zaten belli başlı NLP görevlerinde kabul görmüş modeller olması ve özellikle Pegasus'un özetleme yapmak için oluşturulmuş bir model olması yüksek skorlar alınması konusunda katkı veren detaylardır. Implemente edilen modellerin boyutları ve parametre sayılarına Tablo 8.1'de ve eğitim ve validasyon loss değerlerine ise Tablo 8.2'de yer verilmiştir.

Model İsmi	Boyut	Parametre Sayısı
T5 Base	892MB	222.903.552
Pegasus Large	2.28GB	570.797.056
mT5	1.2GB	300.176.768

Tablo 8.1 Modellerin Boyutları ve Parametre Sayıları

Model İsmi	Eğitim	Validasyon
T5 Base	1,401600	1,391697
Pegasus Large	1,795200	1,973148
mT5	2,8103002	2,178463

Tablo 8.2 Modellerin Eğitim ve Validasyon Loss Değerleri

Eğitim ve validasyon loss değerleri göz önüne alındığında, en iyi eğitim ve validasyon loss değerinin T5 modeli ile elde edildiği görülmüştür.

Literatürde, özetleme işlemi söz konusu olduğunda, modellerin başarısının ölçülmesi ROUGE skoruyla yapılır. Projede de başarı skorunun hesaplanması için ROUGE kullanılmıştır. ROUGE skoru, yapay zeka üretimi metinleri insan metinleriyle

karşılaştırma ve özetleme alanında kullanılır. ROUGE skorunun birden fazla çeşidi vardır ve eğitim sırasında 'ROUGE-1', 'ROUGE-2', 'ROUGE-L', ve 'ROUGE-LSUM' metrikleri kullanılmıştır. ROUGE-1 ve ROUGE-2, verilen iki metin arasında n-gram (n adet kelimenin bir arada olması) üzerinden karşılaştırma yapmaktadır. Bu metriklerin hesaplanmasında n-gram değerleri sırasıyla 1 (unigram) ve 2 (bigram) şeklindedir. ROUGE-L ise oluşturulan metin ve referans metin arasında 'longest common subsequence' hesaplaması yaparak sonuç bulur [14]. Yapılan değerlendirmelerde, ROUGE-L skorunun, ROUGE skoru dışındaki diğer metrikler de dahil olmak üzere, insanların özet değerlendirme kriterlerine daha yakın sonuç verdiği gözlemlenmiştir. [2] Fakat, burada ROUGE skorunun başarılı bir metrik olma durumunun verisetinin kalitesine göre değişen bir konu olduğunu belirtmek gereklidir. Verisetinde bulunan özetlerin, istenen çıktıda bulunması gereken özelliklere sahip olmaması halinde ROUGE skorunun istenen düzeyde olması yanıltıcı olabilir.

Aşağıda eğitilen modellerde elde edilen ROUGE skorlarına yer verilmiştir. ROUGE skoru ne kadar yüksekse, modelin özetleme görevinde o kadar başarılı olduğunu söylemek mümkündür.

Model İsmi	Rouge1	Rouge2	Rougel	Rougelsum
T5 Base	25,018700	11,829600	20,413300	23,447800
Pegasus Large	42,938800	16,573600	24,844800	37,588100
mT5	24,491800	14,381500	21,978800	22,259800

Tablo 8.3 Modellerin Çeşitli Rouge Skor Değerleri

Bununla beraber bugüne kadar elde edilmiş en yüksek ROUGE skoru değerinin 51.06[15] olduğu da değerlendirme sırasında göz önüne bulundurulmalıdır. Bu perspektifte bakıldığından entegre edilen Pegasus modelinin özetleme konusunda projede en başarılı model olduğunu söylemek mümkündür. Pegasus modelini T5 modeli takip etmekte ve 3. sırada mT5 modeli bulunmaktadır. Bu noktada projeye entegre edilen modellerin iyi/orta seviyede performansı olduğu söylenebilir.

Bu bölümün sonlandırmasından önce performans başarısının arttırılması noktasında, uygulanabilecek çözümlerin proje kapsamında kullanılan veriseti kalitesi ve işlemci gücüyle ilişkili olduğunu söylemek gereklidir.

9 Sonuç

Proje için yapılan performans analizi ve deneylerde elde edilen bulgular sonucuya, doğal dil işleme görevlerini yerine getirmeleri için tasarlanmış LLM'lerin özetleme görevinde oldukça başarılı sonuçlar verebileceği gözlemlenmiştir. Kullanılan LLM modellerinin önceden pre-train edilmiş modeller olması ve haber özetleme verisetlerinde fine-tune edilmesi başarılı sonuçlar elde edilmesinde önemli bir unsurdur. Modellerde implemente edilen attention (dikkat) mekanizmaları ve üzerine kuruldukları Transformers mimarisyle kendilerine beslenen metin verilerindeki ilişkileri kurabilmeleri ve önemli noktaları fark ederek bu noktalardan özet bilgi çıkarımı yapmaları mümkün hale gelmektedir.

LLM'lerin eğitilmesinin yanında ayrıca haber sitelerinden haber çekilme işlemi gerçekleştirilmiş ve bu sistem projeye entegre edilerek, kullanıcıların çekilen haberleri görüntüleyip, özetleme işlemi yapabilecekleri bir sistem kurulmuştur.

Çalışma sonucunda, LLM'ler ve özellikle kullanılan T5, Pegasus ve mT5 modelleri üzerine yapılan araştırma ve literatür taraması sonrasında bu modellerin özetleme işlemleri için kullanılabileceği ortaya konmuştur. Özellikle Pegasus modelinin özetleme işleminde oldukça başarılı olduğu, T5 ve mT5 modellerinin de fine-tuning ile oldukça iyi sonuçlar verebileceği kanıtlanmıştır.

Bu çalışmanın ilerleyen zamanlarda, artan haber kaynakları sonucu, kısa sürede haber almak isteyen kişilere özetleme fırsatı sunacak projelere ışık tutabilmesi mümkündür.

1. Özgeçmiş

BİRİNCİ ÜYE

İsim-Soyisim: Hira Nur Morca
Doğum Tarihi ve Yeri: 08.01.2002, Afyonkarahisar
E-mail: nur.morca@std.yildiz.edu.tr
Telefon: 05419149086
Staj Tecrübeleri:

1 Proje Sistem Bilgileri

Sistem ve Yazılım: Windows, Python
Gerekli RAM: 8GB
Gerekli Disk: 5GB



PRIMARY SOURCES

1	v1.overleaf.com Internet Source	1 %
2	www.mdpi.com Internet Source	<1 %
3	Çimen, Murat Erhan. "Ölü Zamanlı Sistemlerin Modellenmesi ve Denetlenmesi", Sakarya Üniversitesi (Turkey), 2022 Publication	<1 %
4	Submitted to The Scientific & Technological Research Council of Turkey (TUBITAK) Student Paper	<1 %
5	arxiv.org Internet Source	<1 %
6	eprints.sztaki.hu Internet Source	<1 %
7	acikbilim.yok.gov.tr Internet Source	<1 %
8	ia600303.us.archive.org Internet Source	<1 %

9

Milad Moradi, Maedeh Dashti, Matthias Samwald. "Summarization of biomedical articles using domain-specific word embeddings and graph ranking", Journal of Biomedical Informatics, 2020

<1 %

Publication

Exclude quotes Off

Exclude bibliography Off

Exclude matches Off