EDUBOT:
KİŞİSELLEŞTİRİLMİŞ DERS YARDIMCISI

2024-2025 GÜZ DÖNEMİ
VERİ BİLİMİ DERSİ
PROJE SUNUMU

245112011
Begüm Erva Şahin
245112044
Murat Görkem ÇOBAN

PROJE TANITIMI

- Amacı:

  Ders notları üzerinden özelleştirilmiş ve doğru cevaplar üretebilen bir sistem geliştirmek.

- Çözülen problem

  Öğrencilerin ders notlarından hızlı ve doğru cevaplar almasını sağlamak.

TEKNİK YAKLAŞIM VE KAPSAM

- **Kullanılan yöntem:** Fine-tuning tabanlı soru-cevap sistemi.
- **Kullanılan teknoloji:** Gemini modeli ve API, AI Studio, Google Colab.
- **Denenen modeller:** Gemini base modeli, Hugging Face(google/mt5-small,ozcangundes/mt5-small-turkish-summarization, savasy/bert-base-turkish-squad)
- **Geliştirme araçları:** Python, Hugging Face.
- **Çalışma ortamı:** AI Studio, Colab.

LİTERATÜR TARAMASI

Yıl	Makale Adı	Veri Kümesi	Yazılım Dili	Modele Verilen Girdi	Kullanılan Modeller	Fine-Tuning / RAG	Kaynakça
2024	"RichRAG: Crafting Rich Responses for Multi-faceted Queries in Retrieval-Augmented Generation"	Çeşitli veri kümeleri	Python	Metin ve Soru-Cevap Çiftleri	BERT, T5	RAG	RichRAG Makalesi[1]
2024	"Retrieval-Augmented Generation: Keeping LLMs Relevant and Current"	OpenAI API, Custom (9 farklı küme)	Python	Metin ve Soru-Cevap Çiftleri	GPT-3, BERT	RAG	RAG Makalesi[2]
2023	"Evaluation of Retrieval-Augmented Generationi"	Custom	Python	Metin ve Soru-Cevap Çiftleri	BERT, T5	RAG	Türkçe RAG Uygulamaları[3]
2022	"Empirical Insights on Fine-Tuning Large Language Models for Question-Answering"	SQuAD, TriviaQA	Python	Metin ve Soru-Cevap Çiftleri	BERT, GPT-3	Fine-Tuning	Fine-Tuning QA Makalesi[4]

VERİ KÜMESİ

Veri Setinin Oluşturulması

- **Kaynak:** Veri Bilimi dersinde kullanılan pptx sunumları.
- **Yöntem:**

  PPTX dosyalatı PDF dosyalarına dönüştürüldü.Ve csv formatı için Gemini modeline yüklendi.

  Gemini modeli sayesinde bu verilerden input-response (soru-cevap) çiftleri içeren bir CSV dosyası oluşturuldu.
- **Verinin modellenmesi:** CSV dosyasındaki "input" sütunu soruları, "response" sütunu ise cevapları temsil ediyor.
- **Ön işleme:**

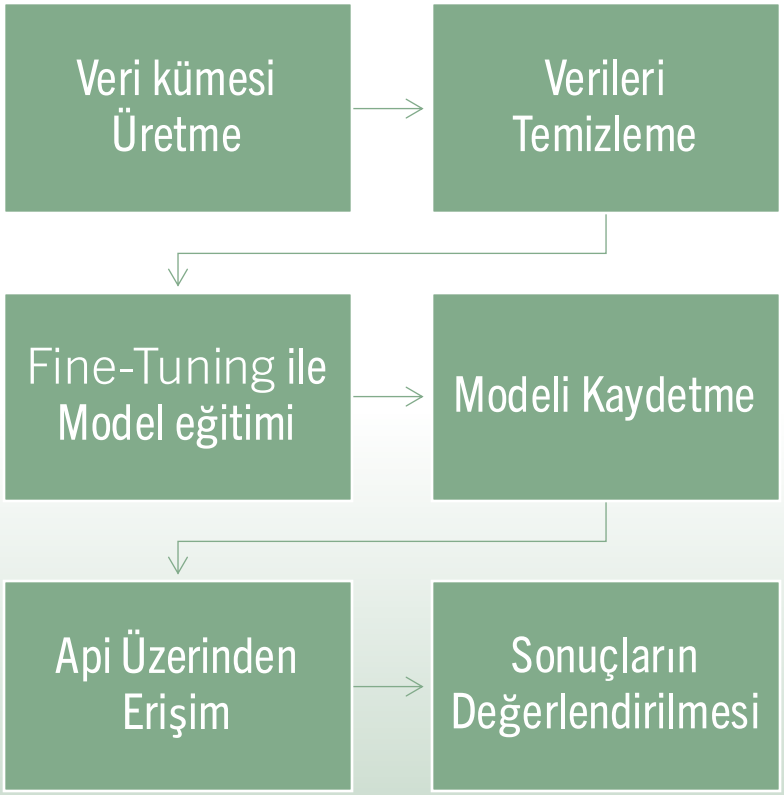
  Gereksiz bilgilerin ayıklanması (örneğin, format bozuklukları).

  Sorular ve cevapların tutarlı ve temiz bir şekilde düzenlenmesi.

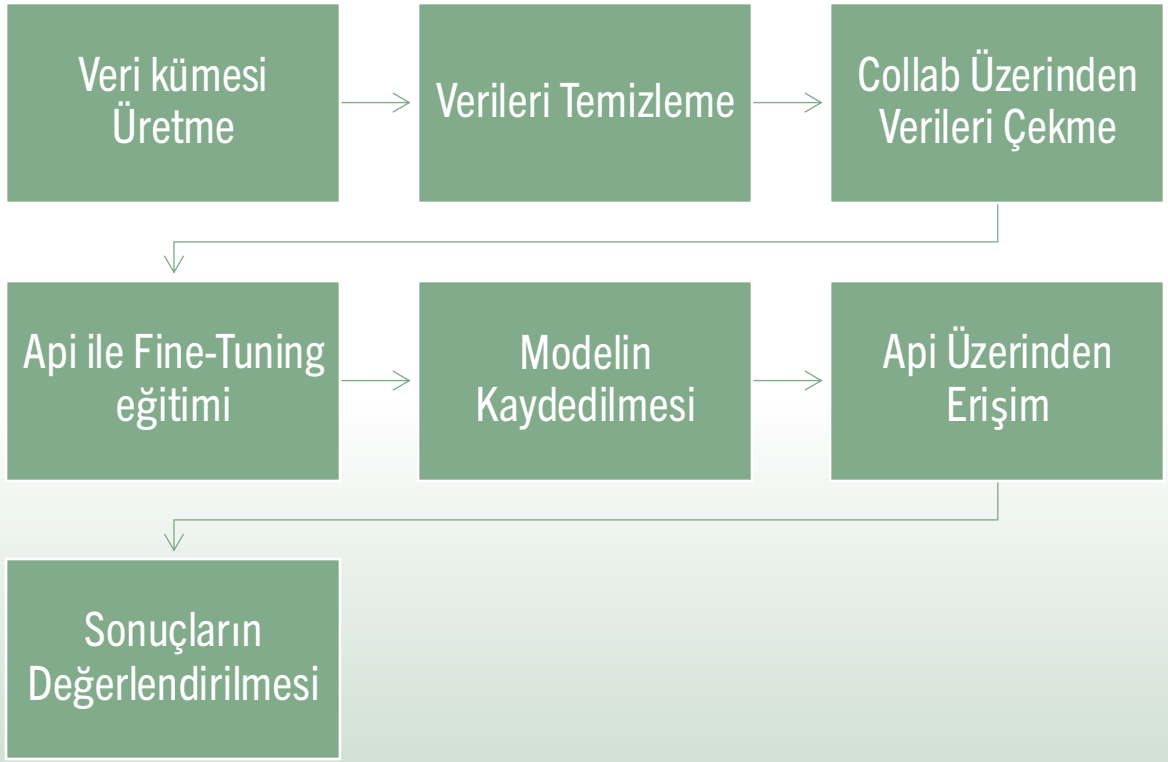
  500 satır ile sınırlandırıldı.

Data Science nedir?	Data Science, veri odaklı hesaplama ve çıkarım odaklı düşüncenin dünyayı anlamak ve sorunları çözmek için kullanılmasıdır.
Data Science'ın temel elementleri nelerdir?	Data Science, programlama becerileri, matematik/istatistik bilgisi ve konu uzmanlığını kapsar.
Python'da NLP için kullanılan kütüphaneler hangileridir?	Python'da NLP için kullanılan bazı kütüphaneler NLTK ve spaCy'dir.
Scikit-learn'ün makine öğrenmesindeki rolü nedir?	Scikit-learn, sınıflandırma, regresyon, kümeleme gibi makine öğrenmesi algoritmalarını içeren bir Python kütüphanesidir.
Outlier nedir? Data setlerinde nasıl bir etkisi vardır?	Outlier, diğer verilerden çok farklı olan değerlerdir. Data setlerinde analiz sonuçlarını çarpıtabilir.
Data augmentation neden yapılır?	Data augmentation, makine öğrenimi modellerinin daha iyi genelleme yapması ve aşırı öğrenmeyi engellemesi için veri setini artırmaya yönelik tekniklerdir.
R'da text mining için hangi kütüphaneler kullanılır?	R'da text mining için 'tm' ve 'quanteda' kütüphaneleri kullanılabilir.
Python'da makine öğrenimi için hangi kütüphaneler kullanılıyor?	Python'da makine öğrenimi için Scikit-learn, TensorFlow ve PyTorch gibi kütüphaneler yaygın olarak kullanılır.
Python ve R arasındaki farklar nelerdir?	Python genellikle web uygulamaları ve veritabanı entegrasyonları için kullanılırken, R genellikle akademik analizler için kullanılır.
Temel matematik ve istatistik bilgisi neden önemlidir?	Data science'ta matematik ve istatistik, veriyi anlamak, model kurmak ve sonuçları yorumlamak için temeldir.
Subject matter expertise nedir? Data science'taki önemi nedir?	Subject matter expertise, belirli bir alanda uzmanlık bilgisidir ve data scientistlerin sonuçları daha iyi yorumlamalarını sağlar.
DataOps nedir?	DataOps, veri analitiğini iyileştirmeyi amaçlayan uygulamalar, süreçler ve araçlar setidir.
Data integration tools nelerdir?	Data integration tools arasında Apache NiFi, Talend ve Fivetran bulunur.
Python'da data analizi için kullanılan temel kütüphaneler nelerdir?	Pandas, NumPy, Matplotlib, Seaborn gibi kütüphaneler Python'da veri analizi için kullanılır.

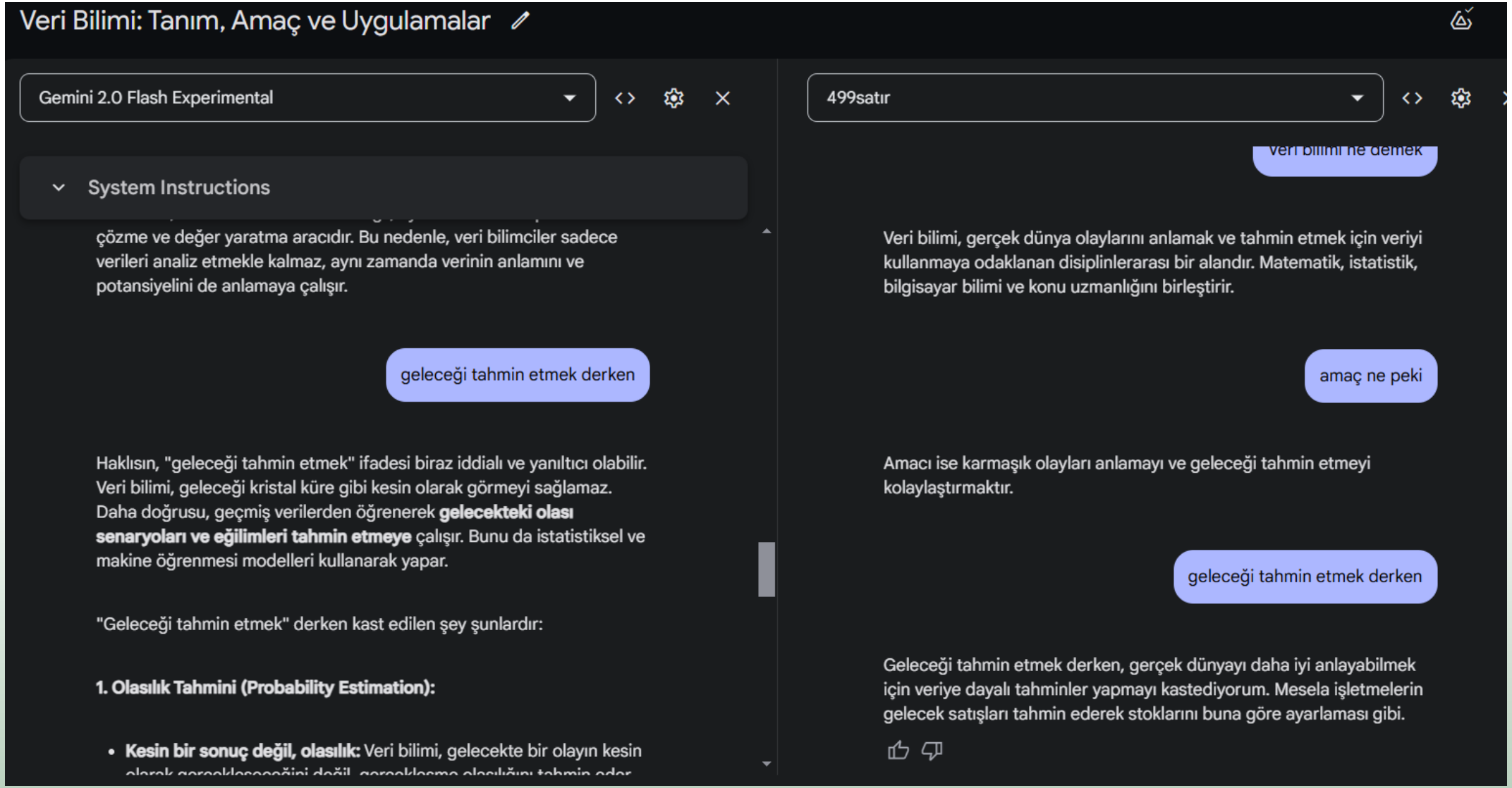
YÖNTEM 1



YÖNTEM 2



DEMO


Veri Bilimi: Tanım, Amaç ve Uygulamalar
Gemini 2.0 Flash Experimental
System Instructions
çözme ve değer yaratma aracıdır. Bu nedenle, veri bilimciler sadece verileri analiz etmekle kalmaz, aynı zamanda verinin anlamını ve potansiyelini de anlamaya çalışır.
geleceği tahmin etmek derken
Haklısın, "geleceği tahmin etmek" ifadesi biraz iddialı ve yanıltıcı olabilir. Veri bilimi, geleceği kristal küre gibi kesin olarak görmeyi sağlamaz. Daha doğrusu, geçmiş verilerden öğrenerek gelecekteki olası senaryoları ve eğilimleri tahmin etmeye çalışır. Bunu da istatistiksel ve makine öğrenmesi modelleri kullanarak yapar.
"Geleceği tahmin etmek" derken kast edilen şey şunlardır:
1. Olasılık Tahmini (Probability Estimation):
Kesin bir sonuç değil, olasılık: Veri bilimi, gelecekte bir olayın kesin
499satır
veri bilimi ne demek
Veri bilimi, gerçek dünya olaylarını anlamak ve tahmin etmek için veriyi kullanmaya odaklanan disiplinlerarası bir alandır. Matematik, istatistik, bilgisayar bilimi ve konu uzmanlığını birleştirir.
amaç ne peki
Amacı ise karmaşık olayları anlamayı ve geleceği tahmin etmeyi kolaylaştırmaktır.
geleceği tahmin etmek derken
Geleceği tahmin etmek derken, gerçek dünyayı daha iyi anlayabilmek için veriye dayalı tahminler yapmayı kastediyorum. Mesela işletmelerin gelecek satışları tahmin ederek stoklarını buna göre ayarlaması gibi.

GELECEK ÇALIŞMALAR

- Veri kümesi arttırılması
- Model eğitim hızı optimizasyonu
- En doğru karar veren model seçimi

KAYNAKLAR

- Yazarlar, "RichRAG: Crafting Rich Responses for Multi-faceted Queries in Retrieval-Augmented Generation," *arXiv:*2406.12566 , 2024. [Online]. Available:https://arxiv.org/pdf/2406.12566
- Yazarlar, "Retrieval-Augmented Generation: Keeping LLMs Relevant and Current," *arXiv:*2407.16833v1, 2023. [Online]. Available: hhttps://arxiv.org/html/2407.16833v1.
- Yazarlar, "Gelişmiş RAG Uygulamalarının Oluşturulması ve Değerlendirilmesi,2023. [Online]. Available: "https://arxiv.org/pdf/2405.07437
- Yazarlar, "Fine-Tuning Language Models for Question Answering Tasks," *arXiv:*2409.15825, 2023. [Online]. Available: https://arxiv.org/abs/2409.15825

TEŞEKKÜRLER

Begüm Erva Şahin & Murat Görkem ÇOBAN