EDUBOT: İyileştirilmiş Öğrenme İçin Kişiselleştirilmiş Veri Bilimi Ders Yardımcısı

Prof. Dr. AHMET SAYAR
Begüm Erva Şahin Murat Görkem ÇOBAN
Kocaeli Üniversitesi, Türkiye

11 Ocak 2025

Özet

Bu çalışma, veri bilimi ders materyallerine dayalı olarak özelleştirilmiş ve doğru cevaplar sunan kişiselleştirilmiş bir ders yardımcısı olan EDUBOT'u sunmaktadır. İnce ayar tabanlı bir soru-cevap sistemi ve Gemini modeli kullanılarak geliştirilen EDUBOT, öğrencilere ders notlarından hızlı ve kesin yanıtlar sağlamayı amaçlamaktadır. Bu makale, kapsamlı bir literatür taraması, metodoloji ve sonuçlar sunarak, özellikle veri bilimi gibi karmaşık alanlarda kişiselleştirilmiş öğrenme sistemlerinin potansiyelini ve önemini vurgulamaktadır. **Bu çalışma, eğitim teknolojileri alanında özgün bir yaklaşım sunarak, doğal dil işleme ve ince ayar tekniklerini kişiselleştirilmiş öğrenme sistemlerine entegre etme konusundaki potansiyeli göstermektedir. EDUBOT, öğrenme süreçlerini daha etkili ve kişisel hale getirme yolunda önemli bir adımdır.**

Anahtar Kelimeler: Eğitim Teknolojisi, Doğal Dil İşleme, İnce Ayar, Soru Cevaplama Sistemleri, Kişiselleştirilmiş Öğrenme

1 Giriş

Veri bilimi alanı, giderek artan önemiyle birlikte, karmaşık kavramları ve uygulamaları içeren zorlu bir disiplin haline gelmiştir. Bu artan karmaşıklık, öğrencilerin öğrenme süreçlerini destekleyecek yenilikçi araçlara olan ihtiyacı ortaya koymaktadır. EduBot, özellikle veri bilimi gibi karmaşık alanlarda, öğrencilerin öğrenme deneyimlerini iyileştirmek için tasarlanmıştır.

EduBot'un temel amacı, ders notlarından hızlı, doğru ve kişiselleştirilmiş cevaplar sağlayarak öğrenme sürecini kolaylaştırmak ve öğrenci performansını artırmaktır. Sistem, sıradan bir soru-cevap yapısının ötesine geçerek, bireysel öğrenme ihtiyaçlarına uyarlanmış çözümler sunmaktadır. İnce ayar tabanlı mekanizmasıyla, öğrencilerin ders materyallerini daha etkin kullanmalarını sağlamaktadır.

Bu çalışmanın amacı, eğitimde yapay zeka tabanlı sistemlerin potansiyelini ortaya koyarak, öğrenme süreçlerini geliştiren yenilikçi çözümler sunmaktır.

2 Literatür Taraması

Bu çalışma, öğrenme sistemlerinde Retrieval-Augmented Generation (RAG) ve ince ayar (Fine Tuning)metodolojilerinin kullanımını incelemektedir. Bu yaklaşımlar, özellikle büyük dil modellerinin (LLM'ler) eğitiminde ve bilgiye erişim süreçlerinde önemli avantajlar sunmaktadır.

RAG, öncelikle bir bilgi kaynağından (bu çalışmada, veri bilimi ders notları) ilgili bilgileri çekerek dil modeline sunar. Bu sayede, modelin sadece eğitim verileriyle sınırlı kalmaması ve güncel bilgilere erişebilmesi sağlanır. [1] çalışması, RAG sistemlerinin çok yönlü sorulara zengin ve bağlamsal cevaplar üretme potansiyelini vurgulamaktadır. Bu çalışma, RAG'nin özellikle karmaşık ve çok boyutlu sorgulara verdiği yanıtların kalitesini artırma yöntemleri sunmaktadır. [2] makalesi ise, büyük dil modellerini

Tablo 1: İlgili Literatür Çalışmaları

Yıl	Makale Adı	Veri Kümesi	Yazılım Dili	Modele Verilen Girdi	Kullanılan Modeller	Fine-Tuning/RAG
2024	"RichRAG: Crafting Rich Responses for Multi-faceted Queries in Retrieval-Augmented Generation"	Özel veri kümesi	Python	Metin ve Soru-Cevap Çiftleri	BERT, T5	RAG
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
2023	"An information fusion based approach to context-based fine-tuning of GPT models"	Özel veri kümesi	Python	Metin ve Soru-Cevap Çiftleri	GPT-2,GPT-3.5	Fine-Tuning
2023	"Evaluation of Retrieval-Augmented Generation"	Özel veri kümesi	Python	Metin ve Soru-Cevap Çiftleri	BERT, T5	RAG
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
2021	"Short Answer Questions Generation by Fine-Tuning BERT and GPT-2"	Özel veri kümesi	Python	Metin ve Soru-Cevap Çiftleri	BERT, GPT-2	Fine-Tuning

desteklemek için RAG'nin güncel ve ilgili bilgileri sağlamadaki kritik rolünü vurgular. Bu, modelin 'halüsinasyon' olarak bilinen yanıltıcı bilgileri üretme olasılığını azaltır ve öğrenme sistemlerinin daha güvenilir olmasını sağlar.

İnce ayar (fine-tuning) yöntemi, önceden eğitilmiş bir dil modelini belirli bir göreve (örneğin, soru-cevap) uyarlamayı içerir. Bu yöntem, modelin belirli bir veri kümesinde ve görev tipine daha iyi performans göstermesini sağlar. [4] çalışması, büyük dil modellerinin soru-cevap görevleri için ince ayar yapmanın etkinliğini gösterir. Bu çalışma, farklı ince ayar stratejilerini ve bunların soru-cevap performansı üzerindeki etkilerini değerlendirir. [3] ise, RAG'in Türkçe dilindeki uygulamalarının zorluklarını ve fırsatlarını tartışır. Bu makale, Türkçe gibi kaynakların ve modellerin sınırlı olduğu bir dilde RAG uygulamasının ne gibi zorlukları olduğunu gösterir.[5] çalışması öğretmenlerin soru üretme yükünü hafifletmek için yapılmış olup, alan bilgisine özel eğitilmiş(fine tuning) bir soru cevap sistemi sunmaktadır özgün bir veri kümesi kullanılmış olup Bert ve GPT-2 modelleri kullanılmıştır.[6] Toan ve arkadaşları yaptıkları çalışmada GPT modellerini bağlamdan kopuk olduklarını tespit edmiş olup Dempster–Shafer kanıt teorisi ile yeni bir bilgi füzyonu tabanlı yaklaşım önermişlerdir.

Bu çalışmada, ince ayar (Fine Tuning) yaklaşımlarından yararlanarak EDUBOT'u geliştirerek literatürdeki eksikliği gidermeyi amaçladık. Özgün değerimiz kullandığımız modelden gelmektedir.Gemini modelini hem veri üretme hemde fine tuning işlemi içik geliştirmiş bulunmaktayız. Ayrıca, ders notları gibi belirli bir bilgi kaynağına özel bir soru-cevap sistemi geliştirerek, öğrenme süreçlerini daha özelleştirilmiş

ve etkili hale getirmeyi hedefledik. Bu şekilde, literatürde var olan yöntemlerin pratik ve uygulamalı bir örneğini sunarak alana katkı sağlamayı amaçlıyoruz.

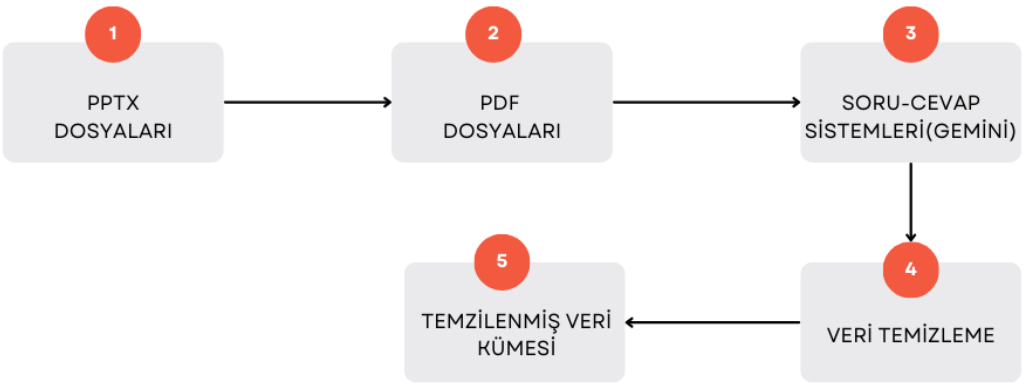
3 Veri Kümesi Oluşturulması

Veri kümemiz, veri bilimi dersinde kullanılan PPTX formatındaki ders notlarından derlenmiştir. Bu notların, EDUBOT sisteminde kullanılabilir hale gelmesi için aşağıdaki adımlar izlenmiştir:

1. **PPTX'ten PDF'e Dönüşüm:** Öncelikle, PPTX formatındaki ders notları PDF formatına dönüştürülmüştür. Bu dönüşüm, Microsoft PowerPoint araçları kullanılarak yapılmıştır. PDF formatı, metin ve görsel içeriği koruyarak, içeriğin doğru bir şekilde elde edilmesini sağlamaktadır. Bu adım, PDF'ten daha kolay metin çıkarımı ve işleme için gereklidir.
2. **PDF'ten Soru-Cevap Çiftlerinin Oluşturulması:** PDF formatındaki ders notları, daha sonra soru-cevap (Input-Response) formatına dönüştürülmüştür. Bu süreçte, Google'ın Gemini 1.5 Flash modeli kullanılmıştır. Modele, ders notlarından belirli bölümler metin komutları (prompt) halinde verilmiş ve modelden bu bölümlere ilişkin soru ve cevaplar üretmesi istenmiştir. Özellikle metnin anlamını koruyan ve kapsamlı cevaplar veren promptlar tasarlanmıştır. Modelin ürettiği soru ve cevaplar, daha sonra manuel olarak gözden geçirilmiş ve uygun görüldüğü şekilde veri kümesine dahil edilmiştir. Modelin verdiği cevapların anlam bütünlüğü, doğruluğu, ve ders materyali ile olan uyumu değerlendirilerek daha kaliteli ve tutarlı bir veri kümesi oluşturulması sağlanmıştır. Bu aşamada, modelin ürettiği soru ve cevaplar, ders notlarındaki bilgileri kapsayacak şekilde optimize edilmiştir.
3. **Csv Formatına Dönüştürme:** CSV formatı, verilerin tablo şeklinde düzenlenmesi ve makine öğrenimi modellerine kolaylıkla aktarılması için seçilmiştir.
4. **Veri Ön İşleme:** Veri kümesinin kalitesini artırmak için ön işleme adımları uygulanmıştır. Bu aşamada, gereksiz başlık ve dipnot bilgileri, paragraf aralarındaki fazla boşluklar, metin içindeki HTML etiketleri veya özel karakterler gibi veriler temizlenmiştir. Metinlerin anlamını bozabilecek her türlü format hatası düzeltilmiştir. Verilerin daha tutarlı ve doğru olmasını sağlamak için, büyük-küçük harf düzeltmeleri, yazım hatalarının düzeltilmesi, gereksiz boşlukların temizlenmesi gibi adımlar da uygulanmıştır.

CSV dosyasının 'input' sütunu, öğrenci sorularını temsil ederken, 'response' sütunu ise bu sorulara verilen cevapları içermektedir. Veri seti, toplamda 500 soru-cevap çifti içermektedir.

Veri Seti Oluşturma Süreci



Şekil 1: Veri Seti Oluşturma Akış Şeması

Bu akış şeması, veri kümesinin oluşturulma sürecini basit ve anlaşılır bir şekilde özetlemektedir.

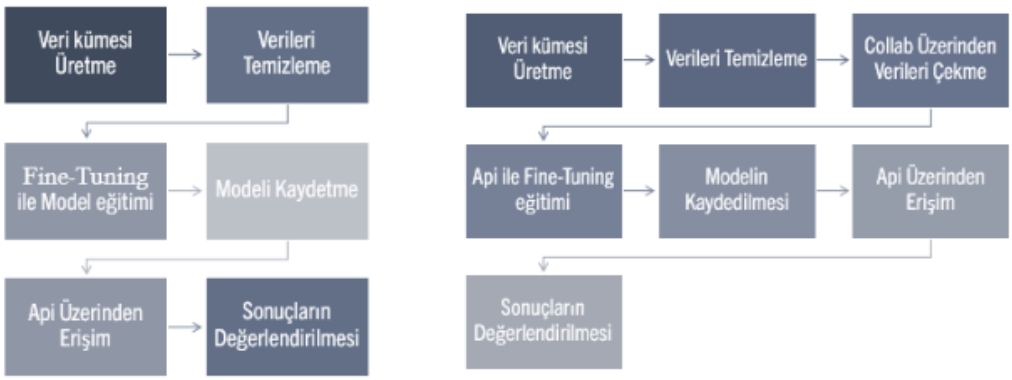
4 Yöntem

EDUBOT sistemi, iki farklı yöntem kullanılarak uygulanmıştır:

- **Yöntem 1:** Google AI Studio kullanılarak gerçekleştirilen ince ayar yaklaşımıdır. Bu süreç, öncelikle kapsamlı bir veri temizleme aşamasıyla başlamıştır. Veri setindeki tutarsızlıklar giderilmiş, eksik veriler tamamlanmış ve eğitim için optimal format elde edilmiştir. Sonrasında, Google AI Studio platformunda gemini-1.5-flash-001-tuning modeli seçilerek ince ayar işlemine geçilmiştir. Bu aşamada, modelin performansını optimize etmek için çeşitli hiperparametreler üzerinde çalışılmıştır. Tuning epochs değeri modelin veri seti üzerinden kaç kez geçeceğini belirlerken, learning rate multiplier öğrenme hızını kontrol etmiş, batch size ise her iterasyonda işlenecek veri miktarını düzenlemiştir. İnce ayar işlemi tamamlandıktan sonra, geliştirilen model kaydedilmiş ve API aracılığıyla erişime açılmıştır.
- **Yöntem 2:** Colab ortamında gerçekleştirilen bir ince ayar sürecini içermektedir. Bu yaklaşımda, hazırlanan veri kümesi öncelikle bulut ortamına aktarılmış ve Colab'da kullanıma uygun hale getirilmiştir. Colab ortamında, API key kullanılarak Gemini base modele erişim sağlanmış ve bu model üzerinde ince ayar işlemleri gerçekleştirilmiştir. Eğitim süreci tamamlandıktan sonra, elde edilen model yine API key aracılığıyla erişilebilir hale getirilmiştir.

Her iki yöntem de sistemin farklı ortamlardaki performansının kapsamlı bir şekilde değerlendirilmesine olanak sağlamıştır. Performans metrikleri, doğruluk oranları, işlem hızı ve kaynak kullanımı gibi çeşitli parametreler üzerinden analiz edilmiştir. Bu analizler, sistemin güçlü ve zayıf yönlerinin belirlenmesinde ve optimizasyon çalışmalarının yönlendirilmesinde önemli rol oynamıştır.

Uygulama sürecinde, her iki yöntemin de kendine özgü avantajları ve zorlukları olduğu gözlemlenmiştir. Google AI Studio yaklaşımı, daha kullanıcı dostu bir arayüz ve entegre araçlar sunarken, Colab ortamı daha fazla özelleştirme imkanı ve esneklik sağlamıştır. Bu farklılıklar, projenin gereksinimlerine ve kullanım senaryolarına göre en uygun yaklaşımın seçilmesinde belirleyici olmuştur.



Şekil 2: Sistem Mimarisi

5 Sonuçlar

EDUBOT sistemi üzerinde yapılan çalışmalar ve testler, sistemin veri bilimi alanındaki eğitim desteği konusunda oldukça başarılı sonuçlar ortaya koymuştur. Sistem, özellikle öğrencilerin veri bilimi sorularına hızlı ve doğru yanıtlar sağlama konusunda etkin bir performans sergilemiştir. Ders içeriklerine ve bilgilere erişim konusunda sağladığı kolaylık, sistemin en önemli başarılarından biri olarak öne çıkmaktadır.

Yapılan demonstrasyonlar, EDUBOT'un karmaşık kavramları anlama ve açıklama yeteneğini net bir şekilde ortaya koymuştur. Sistem, öğrencilerin sorduğu zorlu soruları anlayabilmekte ve bu sorulara kapsamlı ve anlaşılır yanıtlar üretebilmektedir. Bu özellik, öğrencilerin öğrenme sürecine değerli bir destek

sağlamaktadır. Özellikle karmaşık fikirlerin ifade edilmesi ve öğrencilerin sorularının etkili bir şekilde ele alınması konusundaki başarısı, sistemin eğitim alanındaki potansiyelini göstermektedir.

Ayrıca, geliştirilen modelin bağlamdan kopukluk sorununa çözüm sunduğu gözlemlenmiştir. Yapılan ince ayar işlemleri ve hiperparametre optimizasyonları sayesinde modelin bağlamı daha iyi kavradığı ve kullanıcı sorularını daha tutarlı bir şekilde yanıtladığı tespit edilmiştir. Bu iyileştirme, sistemin eğitim süreçlerine sağladığı katkıyı artırmış ve EDUBOT'un daha güvenilir bir öğrenim desteği aracı olmasını sağlamıştır.

6 Gelecek Çalışmalar

Gelecek çalışmalar, EDUBOT sisteminin daha gelişmiş ve verimli bir hale getirilmesini hedeflemektedir. Bu kapsamda, öncelikli olarak veri setinin genişletilmesi ve modelin doğruluk ile verimliliğinin artırılması için optimize edilmiş eğitim süreçlerine odaklanılacaktır.

Veri setinin genişletilmesi kapsamında, daha fazla çeşitlilik ve kapsam sağlayan veri örnekleri toplanarak, modelin farklı konulara ve bağlamlara yönelik performansı artırılacaktır. Özellikle, karmaşık ve çok anlamlı soruların ele alınması için yeni veri kategorileri eklenmesi planlanmaktadır. Bunun yanı sıra, etiketleme süreçlerinin iyileştirilmesi ve veri kalitesinin artırılması için otomatik veri işleme ve temizleme teknikleri uygulanacaktır.

Model eğitimini optimize etmek için ise farklı model mimarileri ve hiperparametre ayarları test edilecektir. Bu süreçte, eğitim süresini kısaltırken doğruluk oranını artırmayı hedefleyen yöntemler üzerinde durulacaktır.

Gelecek çalışmalar aynı zamanda kullanıcı deneyimini geliştirmeye yönelik yenilikleri de içermektedir. Kullanıcı geri bildirimlerini toplamak ve analiz etmek için bir mekanizma geliştirilecek ve bu veriler doğrultusunda modelin yanıt kalitesinin iyileştirilmesi sağlanacaktır. Bununla birlikte, modelin bağlam anlama kapasitesini daha da artırmak için dikkat mekanizmaları üzerinde derinlemesine araştırmalar yapılacaktır.

Tüm bu çalışmaların, EDUBOT'un veri bilimi ve eğitim alanındaki etkinliğini artırarak daha geniş bir kullanıcı kitlesine hitap eden, yenilikçi bir öğrenim destek sistemi olmasına katkı sağlayacağı öngörülmektedir.

Kaynaklar

[1] 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

[2] Yazarlar, "Retrieval-Augmented Generation: Keeping LLMs Relevant and Current," arXiv:2407.16833v1, 2023. [Online]. Available: https://arxiv.org/html/2407.16833v1

[3] Yazarlar, "Gelişmiş RAG Uygulamalarının Oluşturulması ve Değerlendirilmesi, 2023. [Online]. Available: https://arxiv.org/pdf/2405.07437

[4] Yazarlar, "Fine-Tuning Language Models for Question Answering Tasks," arXiv:2409.15825, 2023. [Online]. Available: https://arxiv.org/abs/2409.15825

[5] Yazarlar,"Short Answer Questions Generation by Fine-Tuning BERT and GPT-2", 2021. [Online]. Available: https://library.apsce.net/index.php/ICCE/article/view/4285

[6] Yazarlar,"An information fusion based approach to context-based fine-tuning of GPT models", 2023. [Online]. Available: https://l24.im/uDhmFe5