****

YILDIZ TEKNİK ÜNİVERSİTESİ

ELEKTRİK-ELEKTRONİK FAKÜLTESİ

BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

BLM5105 Doğal Dil İşlemeye Kavramsal Bir Bakış

PROJE

Yıldız CEth

CHATBOT

Muhammet Ali ŞEN- 23501088

Muhammed Kayra BULUT – 23501059

# CHATBOT NEDİR?

Chatbot, yazılı bir sohbeti insanlarla gerçekleştirebilen bir yazılım programıdır. Bu botlar, genellikle insan diliyle iletişim kurabilme, anlama ve buna yanıt verme özellikleri ile ön plana çıkarlar. Chatbotların kullanım amaçları çeşitlilik gösterse de, en sık müşteri hizmetleri, rezervasyon işlemleri ve bilgi sağlama gibi konularda tercih edilirler. Genel olarak, bir internet sitesi veya mesajlaşma uygulaması aracılığıyla faaliyet gösteren chatbotlar, kullanıcıların sorularına yanıt verir veya istedikleri işlemleri yerine getirirler.

* 1. Chatbotlar nasıl çalışır?

Chatbotlar, insanlarla yazışarak etkileşime geçebilmek adına çeşitli tekniklerle geliştirilir ve işletilir. İletişim kurabilmek için özel bir dil öğrenme ve anlama yetisine sahip olurlar. Bu dil, genelde insanların kullandığı dile benzetilmiş ancak chatbotların işleyişi için biraz daha yapay bir yapıda olabilir. İnsanların kullandığı dili anlama kabiliyetine eriştiklerinde, bu dile dayalı sorulara yanıt verebilirler. Chatbotların verdiği cevaplar, çoğunlukla önceden hazırlanmış yanıtlar ya da bir veritabanından alınan bilgilerle şekillenir.

* 1. Chatbotların Doğal Dil İşleme İle İlgisi Nedir?

Doğal Dil İşleme (NLP), bilgisayarların doğal dili anlama ve işleme yeteneği sağlayan bir teknolojidir. Bu teknoloji sayesinde, chatbotlar gibi araçlar, insan dilinde sorulan sorulara cevap verebilir ve çeşitli görevleri yerine getirebilir.

NLP, dilin yapısal, dilbilimsel ve anlamsal özelliklerini analiz eden yöntemlerden faydalanır. Bu yöntemler, chatbotların insanların kullandığı kelimeleri ve cümle yapılarını anlamasını ve buna uygun yanıtlar geliştirmesini sağlar. Ayrıca, chatbotların dil öğrenimini ve kullanımını geliştirerek, insanlarla daha akıcı bir iletişim kurmalarına olanak tanır.

Doğal dil işleme teknolojisi, sadece chatbotlarla sınırlı kalmayıp, metin çevirisi, metin analizi ve sınıflandırma gibi birçok farklı alanda da kullanılır. Bu teknoloji aracılığıyla, insanlar arama motorlarında doğal dillerini kullanarak arama yapabilir veya mobil cihazlarında dil öğrenme uygulamalarından faydalanabilirler.

# CHATBOT ÇEŞİTLERİ

Retrieval Based Chatbot

Retrieval Based Chatbotlar, kullanıcıların sorularına önceden hazırlanmış ve bir veritabanında saklanan yanıtlarla cevap veren chatbot çeşididir. Bu tür botlar, doğal dili algılayarak kullanıcıların sorularını anlar ve veritabanında uygun yanıtları arar. Eğer uygun bir yanıt bulunamazsa, genellikle standart bir cevap verirler.

Bu chatbotlar, doğal dilde yapılan soruları anlayıp cevaplamak için programlanmış yanıtların saklandığı bir veritabanı kullanır. Özellikle kullanıcılar belirli bir konu hakkında bilgi almak istediklerinde bu tür chatbotlar oldukça yararlı olabilir. Mesela, bir turizmle ilgili chatbot, turistik yerler hakkında soruları yanıtlamak üzere Retrieval Based Chatbot yapısını kullanabilir.

Retrieval Based Chatbotların sınırlı veritabanına bağlı olmalarından dolayı, daha özel ve karmaşık doğal dil sorularına yanıt vermede güçlük yaşayabilirler. Bu yüzden, sadece bu tip chatbotların yanı sıra, dil modelleme üzerine kurulu Rule Based Chatbotlar ve öğrenme yeteneğine sahip chatbotlar da geliştirilmiştir.

Generative Based Chatbot

Generative based Chatbotlar, kullanıcıların doğal dillerde sorduğu sorulara kendi başlarına doğal dilde yanıt verebilen chatbot çeşididir. Bu tür botlar, dil modelleme tekniklerini kullanarak dilin yapı ve anlamını öğrenir ve bu bilgilere dayanarak yanıtlar oluştururlar. Bu chatbotlar, özellikle karmaşık ve spesifik sorulara yanıt vermede daha etkilidirler ve insanlarla daha doğal bir iletişim kurma kapasitesine sahiptirler.

Öğrenmeye dayalı diğer chatbotlar gibi, Generative based Chatbotlar da eğitim verileri üzerinden geliştirilirler. Bu botlar, aldıkları eğitim verisini analiz eder ve dilin yapısını ve anlamını kavrayarak dil modelleme teknikleriyle öğrenir. Böylece, insanların doğal dilde yönelttikleri soruları anlayıp uygun yanıtları üretebilirler. İnsan dilini taklit ederek daha doğal iletişim sağlarlar.

Bu chatbotlar, doğal dillerde sorulan sorulara daha insani yanıtlar verme yetenekleri sayesinde çok çeşitli alanlarda tercih edilirler. Örnek olarak, bir müşteri hizmetleri chatbotu, doğal dilde yapılan sorulara daha doğal ve insani yanıtlar verebilir, bu da onların insanlarla daha doğal bir diyalog kurmasını sağlar.

# LLM Nedir?

Büyük Dil Modelleri (LLM), milyonlarca hatta milyarlarca kelimelik kapsamlı veri setlerinden öğrenme yeteneğine sahip yapay zeka sistemleridir ve insan dilinde anlama ve oluşturma kapasitesine sahiptirler. Bu sistemler, geniş metin koleksiyonları üzerinde eğitilerek dilin yapısını, anlamını ve nasıl kullanıldığını özümsüyorlar. LLM'ler, metin üretme, özetleme, soru-cevap sistemleri oluşturma ve dil çevirisi gibi bir dizi işlevde kullanıma uygun hale gelirler.

* 1. **LLM'lerin Önemi**

Yapay zeka ve doğal dil işlemesi (NLP) disiplinlerinde LLM'ler sayesinde son yıllarda dikkate değer gelişmeler yaşanmıştır. İnsan dilini kavrama ve taklit etmede gösterdikleri üstün yeteneklerle, makinelerin insanlarla daha akıcı ve etkin bir biçimde iletişim kurmalarını mümkün kılmışlardır. LLM'ler, kullanıcıların sorularına yanıtlama, yazılı içerik oluşturma ve duygu analizi gibi konularda önemli ilerlemeler sağlamıştır.

* 1. **LLM'lerin Kullanım Alanları**

LLM'ler, çok çeşitli uygulama alanlarına sahiptir:

* **Metin Üretimi ve Yazarlık Yardımı**: Makaleler, hikayeler, kod ve e-postalar gibi belirli formatlardaki metinlerin otomatik olarak üretilmesi.
* **Otomatik Çeviri**: Farklı diller arasında yüksek kaliteli çevirilerin yapılması.
* **Soru-Cevap Sistemleri**: Kullanıcıların sorularına doğrudan ve anlamlı cevaplar üreterek, bilgiye erişimi kolaylaştırma.
* **Duygu Analizi**: Müşteri yorumları, sosyal medya gönderileri gibi metinlerden duygusal ton ve tutumların analizi.
* **Eğitim ve Öğretim**: Öğrencilere özel ders materyalleri üretme, ödev yardımı ve dil öğreniminde destek sağlama.
* **Yasal ve Tıbbi Asistanlık**: Hukuki belgelerin hazırlanması, tıbbi raporların analizi gibi uzmanlık gerektiren alanlarda yardımcı olma.
  1. **LLM Çeşitleri**

Büyük Dil Modelleri (LLM), kullanılan eğitim verileri, modelin yapısı ve amaç bakımından birbirlerinden ayrılır. Bu modellere örnek olarak GPT (Generative Pre-trained Transformer), BERT (Bidirectional Encoder Representations from Transformers) ve T5 (Text-to-Text Transfer Transformer) sayılabilir. Her bir model, kendine has özellikleri ve avantajlarıyla dikkat çeker:

* **GPT Serisi:** Açık uçlu metin oluşturmada oldukça başarılı olan GPT, çeşitli metin tabanlı işlevlerde üstün performans sergiler.
* **BERT:** Metni anlama ve dilin bağlamını derinlemesine çözme yeteneğiyle ön plana çıkan BERT, metin sınıflandırma ve soru-cevap uygulamalarında özellikle güçlüdür.
* **T5:** "Metinden metne çeviri" prensibini kullanarak hemen hemen her türlü metin tabanlı işi modelleyebilen T5, geniş bir kullanım alanına hitap eder.
  1. **GPT-2'nin Tanımı ve Özellikleri**

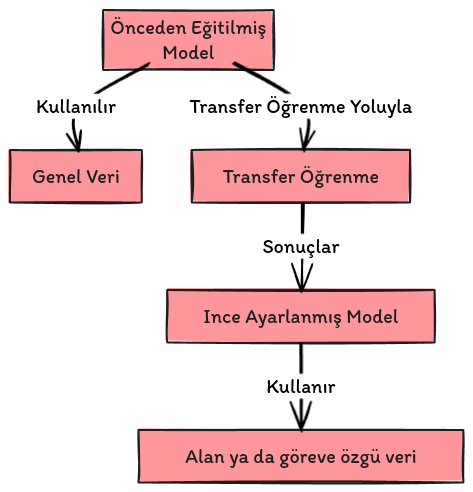
GPT-2, OpenAI tarafından geliştirilmiş ve GPT serisinin bir sonraki evresi olarak ortaya çıkan bir Büyük Dil Modelidir (LLM). Bu model, geniş bir metin havuzundan öğrenerek çeşitli doğal dil işleme (NLP) görevlerinde metin oluşturabilme yeteneğine sahiptir.

Transformer tabanlı bir mimariye sahip olan GPT-2, internet üzerinden elde edilen milyonlarca sayfa bilgisini içeren 1,5 milyar parametre ile donatılmıştır. Model, verilen metnin devamını tahmin edebilme yeteneği sayesinde metin üretimi, dil çevirisi ve özetleme gibi alanlarda etkin bir şekilde görev yapar.

Metin Oluşturma: GPT-2, haber yazıları, öyküler ve diyaloglar üretmek amacıyla kullanılabilir. Dil Dönüştürme: Çeşitli diller arasında yüksek kalitede çeviri yapma kapasitesine sahiptir. Metin Özeti Çıkarma: Geniş metinleri kısaltarak özetler oluşturabilir. Soru ve Cevap: Sorulan sorulara doğru ve mantıklı yanıtlar verebilir.

# Fine-Tune Nedir?

Fine-Tune, önceden eğitilmiş bir yapay zeka modelinin, belirli bir görev, veri kümesi veya uygulama alanı için performansını optimize etmek üzere ek bir eğitim sürecinden geçirilmesidir. Bu süreç, modelin genel bilgi ve yeteneklerini koruyarak daha dar bir alanda daha yüksek doğruluk ve etkinlik sağlamasına olanak tanır. Fine-tuning'in önemi, modelin genel amaçlı bilgisini, spesifik bir görevde uygulanabilir, daha hassas ve etkili hale getirme yeteneğinden kaynaklanır. Bu, özellikle büyük veri setlerinin işlenmesi ve yüksek maliyetli hesaplama kaynaklarının gerektiği durumlarda, zaman ve kaynaklardan tasarruf sağlar.



# CHATBOT PROJESİ

Bu projede derin öğrenme tekniklerini kullanarak bir chatbot oluşturacağım. Chatbot, **“Soru: xxx Cevap: xxx”** formatındaki bir veri kümesi üzerinde eğitilecektir. Kullanıcının mesajının cevabını doğrudan model verecektir.

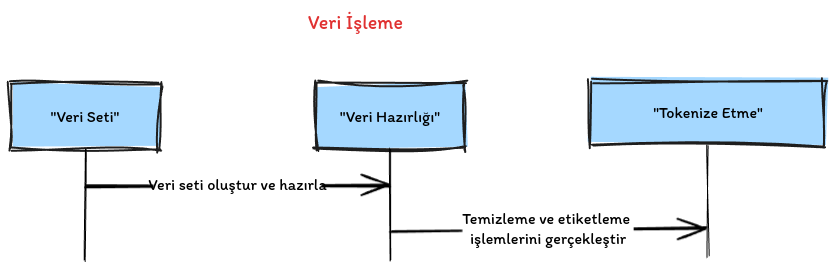
Chatbot projesinde **transformers, torch, pandas,** **PyQt5** gibi Python kütüphanelerini Python 3.9 versiyonunu kullandık.

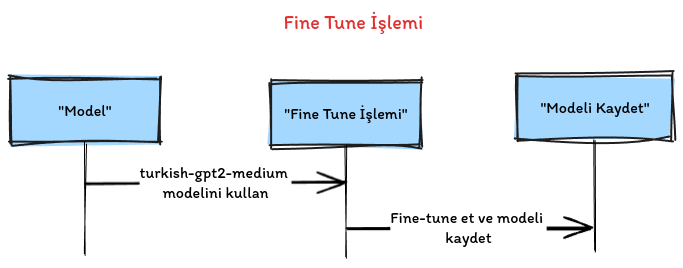
Doysamız ve açıklamalarına kısaca değinirsek;

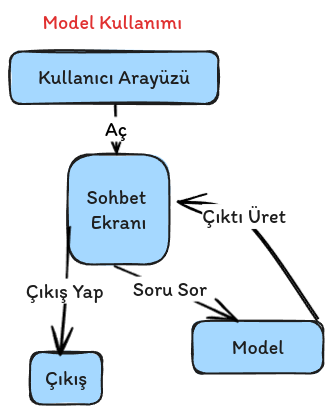
* **ana\_veri\_seti.csv** – Soru ve cevapları içeren veri seti.
* **model\_egit.ipnyb** – Modeli eğitmek için yazılmış python dosyası.
* **BestModelParameters** – model.save fonksiyonu sonucunda oluşan modeli kullanmak için gerekli bilgileri tutan dosya.
* **gui.py** – Arayüz tasarımını PyQt5ile yaptığımız pyhon dosyası.

Projeyi kabaca özetlersek

1. Veri setini oluşturduk ve sorulara göre cevapları hazırladık.
2. Veriler üzerinde bir dizi temizleme, etiketleme (tokenization) işlemlerini gerçekleştirerek veri setini eğitime hazırladık.
3. Eğitimi modelledik. Bunun için **‘turkish-gpt2-medium’** modelini kullandık. Modeli fine tune ettik.
4. Modeli kaydederek yeni sorgulara hazırladık.
5. Arayüz üzerinden yeni sorguları alarak fine tune edilen modeldeki oluşan tahmini cevapları ekranda gösterdik.







Veri Setini Oluşturma ve Yükleme

Öncelikle, içeriği oluşturmak için ders notları sayfasından hoca ve ders yorumlarını elde ettik.(<https://github.com/baselkelziye/YTU_Bilgisayar_Muhendisligi_Arsiv>) Sonra, Bu verilerle ChatGPT4 yardımıyla 651 adet soru ve cevap çifti elde ettik. Bu soru-cevap çiftlerini **‘ana\_veri\_seti.csv’** dosyasına kaydettik. Veri dosyası CSV biçimindedir, bu nedenle CSV dosyasını Python'a ayrıştırmak için **'pandas’** kütüphanesini kullandık.

**‘ana\_veri\_seti.csv’** dosyası aşağıdaki şekilde görünmektedir.



Veri Setini Temizleme ve Önişlemden Geçirme

Makine öğrenimi ya da derin öğrenme modelleri kullanarak metin verileri üzerinde çalışmaya başlamadan önce, verilerin üzerinde bir dizi ön işleme adımı uygulanması zorunludur. Bu işlemler, ihtiyaçlara göre değişiklik gösterir ve verileri modellemeye hazırlar.

Metin verilerini işlemenin ilk ve en temel adımı tokenizasyondur. Tokenizasyon, metni kelime gibi daha küçük birimlere bölmeyi içerir.

Bu aşamada, **turkish-gpt2-medium** modelinin tokenizer'ını kullanarak metni tokenlere ayırıyoruz. Bu model, Türkçe üzerine özelleştirildiği için tokenizer'ı güncellememize gerek kalmıyor. Ama **‘tokenizer.add\_special\_tokens’** metoduya **<|soru|>** ve **<|cevap|>** tokenlarını ekliyoruz. Bunun sebebi modelin soru ve cevapları metinde ayırt edebilmesi.

Modeli Oluşturma

Eğitim verilerim tamamlandıktan sonra, **'YTU COSMOS'** araştırma ekibinin Türkçe üzerine özel olarak ayarladığı 'turkish-gpt2-medium' modelini indirdik. Bu işlem için **transformers** kütüphanesinde bulunan **AutoModelForCausalLM** fonksiyonundan yararlandık. Modeli 15 tur (epoch) eğitimden geçirdikten sonra, model yüksek seviyede bir doğruluk performansı sergiledi. Son olarak modeli **'BestModelParameters'** ismiyle kaydetmeye karar verdik.

Cevap Tahmini

Eğitilmiş modelimi yükleyip, kullanıcının girdiğine göre modelin vereceği yanıtı tahmin edebilmek için grafiksel bir kullanıcı arayüzü kullanmayı planlıyoruz. Model tarafından üretilen yanıtları işleyerek bu arayüz üzerinde göstereceğim.

İhtiyaç duyduğum paketleri içe aktardık ve eğitim sürecinde oluşturduğum **'BestModelParameters'** dosyasını kullanarak hem modeli hem de tokenizer'ı yükledik.

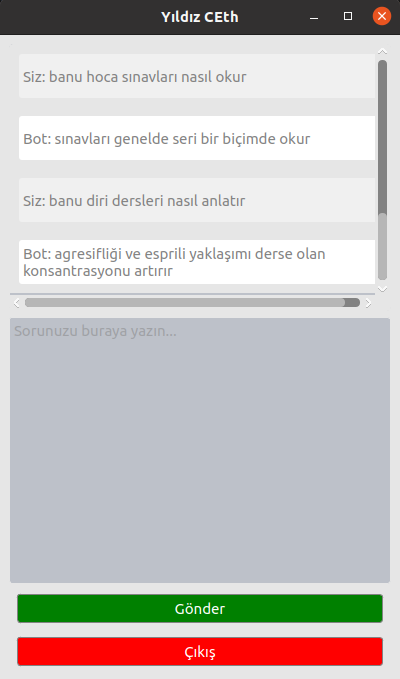
Kullanıcı arayüzü oluşturmak için **PyQt5** kütüphanesinden faydalandık. Kullanıcıdan bir giriş almak, modelden yanıtı çekmek ve bunu GUI üzerinde göstermek için yazdığım yardımcı fonksiyonları kullandık.

# PROGRAMI ÇALIŞTIRMA

Chatbotu kullanabilmek için **’gui.py’** dosyasını çalıştırmak yeterli. Tabi ki eğitilmiş modelin de mevcut olması gerekiyor. Eğer eğitilmiş bir model yoksa önce eğitim işlemini halletmemiz lazım. Bunun için de **Colab’e** para ödemek gerekiyor. Çalıştırılması gereken dosyaysa **‘model\_egit.ipnyb’.**

Eğitim sırasında herhangi bir hata görülmezse modeli başarılı bir şekilde oluşturulmuş demektir.

Grafik arayüzü aşağıda gösterilmektedir.



# SONUÇ

Yürüttüğümüz doğal dil işleme ve veri bilimi projesinde, sohbet robotları aracılığıyla doğal dil işleme dinamiklerini derinlemesine inceledik. Bu süreçte, Python kullanarak LLM tabanlı bir sohbet robotu geliştirerek Large Language Models üzerine kapsamlı bir araştırma gerçekleştirdik. Proje kapsamında, iş gereksinimlerine göre özelleştirilebilen verilerle çalıştık. Veri setimizi, Yıldız Teknik Üniversitesi ve Bilgisayar Bilimleri temalarıyla özdeşleşecek şekilde tasarladık. Verileri manuel olarak sisteme girme süreci zaman alsa da, elde ettiğimiz sonuçlar emeğimizin karşılığını verdi. Günümüzde chatbot'lar her alanda kullanılıyor ve işletmeler, iş akışlarını iyileştirmek için bot teknolojisini dört gözle bekliyor.