çekleştirebilirler. Bu görevler arasında, *duygu analizi*, *makine tercümesi*, *soru cevaplama*, *eksik kelime tamamlama* bulunmaktadır. İstem teknikleri, modele çözülmek istenen problemi bir metin istemi olarak sunar. Bu yapılırken bu probleme benzeyen bir ya da daha fazlası çözümleriyle birlikte istemin içerisinde yer alabilir. Böylece, dil modeli neyi çözmesi gerektiğini kestirebilir. GPT-3 gibi çok daha güçlü modeller buna gereksinim duymadan bu görevleri yapabilmektedirler.

Son yıllarda geliştirilen önemli büyük dil modellerinin tamamına yakını *transformer* derin öğrenme modelini esas almıştır. Bu modelde en önemli unsur, öz-dikkat denilebilecek (*self-attention*) girdi verinin her bir parçasının önem derecesini ayırt edici biçimde ağırlıklandırma tekniğidir [2]. Transformer'ların uygulama alanları arasında makine tercümesi, doküman özetleme, doküman oluşturma, biyolojik dizi analizi ve video anlamlandırma yer almaktadır. Transformer'lar genelde önce gözetimsiz öneğitim ve ardından gözetimli ince ayarlama içeren bir kendi gözetimli eğitimden geçirilmektedir. GPT-2, GPT-3, GPT-4, BERT, XLNet, RoBERTa ve ChatGPT gibi büyük dil modelleri transformer yapısındadırlar.

Transformer mimarisinin Google tarafından 2017 yılında oluşturulmasının ardından [3], OpenAI 2018 yılında önceden eğitilmiş üreteç transformer yapısını yayınlamış ve ilk örneği olarak GPT-1 modelini geliştirmiştir [4]. Bu yapı 12 öz-dikkat içeren 12 katmandan ve her birinde 64 olmak üzere toplamda 768 adet boyutsal durum içermektedir. Daha sonra geliştirilen GPT-2 modeli, GPT-1'dne göre hem parametre sayısı hem de veriseti bakımından 10 kat büyüktü [5]. GPT-2 kamunun kullanımı için açık erişim olarak yayımlandır. 2020 senesinde sunulan GPT-3 ise 175 milyar parametre ile çok daha büyük bir yapıya sahipti. GPT-3'ün kaynak kodu hiçbir zaman açıklanmadı. OpenAI, GPT-3.5 adı verilen modelin ince ayarlanması ile geliştirilen ve adına ChatGPT dediği ürününü Kasım 2022'de sundu. InstructGPT adı verilen bu ince ayarlamada, eğitilen büyük dil modellerine istemlere yanıt dönme ve istem takip etme özellikleri ekleme amacı güdülmektedir [6, 7].

Bu çalışmada, hali hazırda açık erişime sunulmuş bulunan büyük dil modelleri ve bu modellerin eğitilmelerinde ve ince ayarlanmalarında kullanılan gene açık erişimli verisetleri incelenmiştir. Ayrıca, bir büyük dil modelinin Türkçe içerik ile eğitilmesi denenmiş, bir başka önceden eğitilmiş ağın Türkçe istem girdileri ile ince ayarlanması ve bu istemlere yanıt dönmesi incelenmiştir. Bu deneyler ile ilgili yürütülen hazırlıklar ve aşamalar Bölüm 4'te, sonuçlar ise Bölüm 5'te sunulmuştur.

2 Verisetleri

Başarım oranı yüksek bir dil modelinin eğitilebilmesi için gerekli olan en önemli aşamalardan birisi çok büyük ve ön işlemden geçmiş bir metin verisetinin hazırlanmasıdır. Bu aşama, hali hazırda sunulan açık erişimli verisetleri indirilerek yapılabileceği gibi kaynaklar indirilerek sıfırdan da gerçekleştirilebilir. İngilizce ve diğer yaygın diller için bu hazır verisetlerinin kolaylıkla bulunabilmesine karşın malesef Türkçe için hazır verisetleri yeterli ve kolay ulaşılabilir değiller. Bu konudaki bir diğer önemli husus da telif hakları meselesidir. Tablo-1'de açık erişimli veri kaynakları hakkında bilgiler özetlenmiştir.

Hugging Face, Inc. firmasının sağladığı altyapı ile önceden hazırlanmış verisetleri ve derin öğrenme ağ modelleri herkese açık bir şekilde paylaşılmaktadır [13]. Bu verisetleri arasında farklı görevler için oluşturulmuş Openwebtext [14], C4 ve PIQA [15] gösterilebilir. Huggin Face tarafından sunulmakta olan modeller arasında ise farklı görevler için eği-

[1]Common Crawl, belli aralıklarda bu kayıtları almakta ve sunmaktadır. Bu çalışmanın yapıldığı andaki en son arşiv kaydı Mart/Nisan 2023 tarihlidir ve kayıt adı CC-MAIN-2023-14'tür.