Kaynak	Büyüklük	Türkçe içerik	Açıklama
Common Crawl[1]	yaklaşık 3.15 milyar web sayfası (380 TiB) [8]	%0.7897 [9]	Kâr amacı gütmeyen kuruluş.
BookCorpus [10]	11,000 kitap, 985 milyon kelime.	*bilinmiyor*	OpenAI'nın ilk GPT modeli için kullanıldı.
C4 ve T5 [11]	745 GB.	Sadece İngilizce veri.	Common Crawl verisinin temizlenmiş hali. Google tarafından yayımlanmıştır.
Openwebtext	8 milyon doküman, 38GB veri.	Sadece İngilizce veri.	GPT-2'in eğitildiği Webtext'e alternatif olarak hazırlanmıştır.
RedPajama [12]	1.2 Trilyon belirtke.	*bilinmiyor*	Diğer büyük kaynaklardan veriseti oluşturmayı sağlayan proje.
Vikipedi	731 MB.	Sadece Türkçe veri.	Bu çalışmada kullanılmıştır.

Tablo 1: Açık erişimli metin kaynakları.

tilmiş T5, BERT, BART, GPT-2 ve BLOOM gibi önemli modeller yer almaktadır.

Türkçe metinlerden veriseti oluşturma. Hali hazırda açık erişimle sunulan verisetlerinde Türkçe içeriğin hiç olmaması ya da çok az yer alması nedeniyle büyük dil ağları araştırmalarında kullanmak üzere sıfırdan bir veriseti oluşturmak elzem olmaktadır. Bu çalışmada yer alan deneyleri yürütülmek için böyle bir veriseti sadece Vikipedi (*Wikipedia* Türkçe sürümü) makaleleri kullanılarak gerçekleştirilmiştir. Bunun için, güncel ve sıkıştırılmış `trwiki`[2] arşivi indirilmiş ardından bir Python betiği yardımıyla `json` formatında veriseti oluşturulmuştur. 731 MB büyüklüğündeki bu veride ön işleme ve temizlik yapıldıktan sonra farklı uzunluklarda toplam 818.454 adet metin elde edilmiştir. Bu metinler, `tiktoken` modülü [16] ile belirtkeleştirilince 296.1 milyon adet belirtke oluşmuştur. Bu sonuç GPT-2 modelinin de belirtkeleştirildiği 50 bin ögeden oluşan `r50k_base` dil kodlaması kullanıldığında elde edilen sayıdır. GPT-3.5 ve GPT-4'te kullanılan 100 bin ögelik `cl100k_base` kullanılacak olursa oluşan sayı 242.6 milyon olmaktadır. Bunun Türkçe için daha uygun olduğu düşünülebilir.

Belirtkeleştirme (*Tokenization*) Sözcüksel analizde bir girdi metni oluşturan parçaların sınıflandırılması ve ayırt edilmesi işlemidir. Oluşturulan belirtkeler takip eden bir başka işlemde kullanılırlar. Girdi verisetinden yer alan bütün veri belirtkelere ayrılarak bir sözvarlığı seti oluşturulur. Büyük dil modellerinin eğitilmesinde kullanılan verisetleri üzerinde çoğunlukla *Byte-Pair Encoding* (BPE) belirtkeleştirme algoritması uygulanmaktadır.

3 Modellerinin eğitilmeleri ve ince ayarlanmaları

Açık erişimli büyük dil modelleri. Ticari büyük dil modelleri dışında bazı şahıs ve kurumlar tarafından kaynağı paylaşılan büyük dil modelleri mevcuttur. Bunlar arasında, Meta şirketi tarafından yayımlanan LLaMa [17] modelinin 7, 13, 33 ve 65 milyar parametre içeren varyantları bulunmaktadır. Bu modeller 1 ve 1.4 trilyon belirtke (*token*) ile eğitilmişlerdir. Malesef eğitim verisetinde yer alan 20

[2] `https://dumps.wikimedia.org/trwiki/20230520/`