M. Ali Bayram - Doktora Araştırma Özeti

Doktora çalışması, düşük kaynaklı diller ve özel alanlara yönelik büyük dil modellerinin (LLM) geliştirilmesi ve/veya düşük kaynaklı ortamlara adaptasyonu üzerine yoğunlaşmaktadır. Araştırma sürecinde, Türkçe doğal dil işleme için kapsamlı benchmark sistemleri, alan-özel veri setleri, tokenizer yapıları ve öğrenme teknikleri geliştirilmiştir.

Türkçe eğitim sistemi temelli 6.200 sorudan oluşan TR-MMLU adlı çoktan seçmeli değerlendirme seti hazırlanmış ve 47 büyük dil modeli bu veri seti üzerinde test edilmiştir. Değerlendirme sistemi Hugging Face üzerinde açık kaynak olarak sunulmuş, ayrıca her modelin ayrıntılı sonuçları ve karşılaştırmaları yayımlanmıştır. Bu çalışma, **Türkçe için ilk defa bu düzeyde bir kıyaslama altyapısı** sunarak alandaki standartları belirlemiştir.

Sağlık alanına yönelik 167.000'den fazla **gerçek hasta-doktor etkileşiminden oluşan bir veri seti derlenmiş**, bu veri setiyle Türkçe medikal LLM'ler eğitilmiştir. LLAMA ve Gemma modelleri üzerine kurulan DoctorLlama ve DoctorGemma serileri, LoRA ve SLerp gibi parametrik adaptasyon teknikleri kullanılarak modeller hem verimli hem de modüler biçimde uyarlanmıştır.

Tokenizasyon alanında Türkçe'nin morfolojik yapısına duyarlı yeni bir tokenizer tasarlanmış, bu yapı 15'ten fazla tokenizer ile karşılaştırılmıştır. %TR (Dile özgü token oranı) ve %Pure (anlamsal bütünlük) gibi metrikler önerilmiş ve ölçülen performanslar, model doğruluğu ile doğrudan ilişkilendirilmiştir. Bu sayede yalnızca model değil, kullanılan tokenizasyon yaklaşımının da dil başarısına etkisi **ilk defa sistemli** olarak gösterilmiştir.

Geliştirilen veya katkı sağlanan açık kaynak projeler arasında tokenizer ve model mimarilerine dair kapsamlı kütüphaneler, Apple Silicon için optimizasyon içeren MLX örnekleri ve verimli LLM eğitimi sağlayan Unsloth araçları yer almaktadır. Tüm projeler GitHub ve Hugging Face profillerinde açık kaynak olarak paylaşılmıştır.

Araştırmanın ileri aşamalarında, küçük ölçekli Türkçe modellerin eğitimi, adapter tabanlı dinamik LLM servis sistemleri ve dilden bağımsız anlam uzaylarında model eğitimi gibi konular yer almaktadır.

Bu doktora süreci boyunca SIU 2025 konferansında iki bildiri kabul edilmiş, dört makale uluslararası dergilere gönderilmiştir. Ayrıca, Kasım 2024'te İzmir Bakırçay Üniversitesi ev sahipliğinde düzenlenen IV. Uluslararası Sağlıkta Yapay Zeka Kongresi'nde iki bildiri sunulmuştur. Bu sunumda ilk defa, **veri kalitesine dayalı öğrenme oranı** yaklaşımı önerilmiş ve Türkçe sağlık modellerinin sınıflandırma başarısı detaylı bir şekilde paylaşılmıştır.

Ayrıca, **tamamen özgün Türkçe kaynaklardan oluşturulmuş 293.468 soruluk** dev bir soru havuzu hazırlanmıştır. Bu veri seti, TUS, KPSS gibi çok sayıda önemli sınavı kapsamaktadır. Veri seti Hugging Face ve Zenodo üzerinden açık kaynak olarak yayımlanmış, Türkiye'nin en kapsamlı NLP kaynaklarından biri haline gelmiştir.

Beyazperde, Hepsiburada, Kitapyurdu ve Yorumbudur platformlarından toplanan toplam **5.817.233 etiketli yorum toplanarak açık olarak yayınlanmıştır.** Bu çalışma, Türkçedeki duygusal ifade zenginliğini ve söylem çeşitliliğini veri temelli olarak ortaya koymakta, duygu analizi, kişiselleştirilmiş öneri sistemleri ve metin sınıflandırma gibi uygulamalara katkı sağlamaktadır.