Tarihi El Yazması Belgelerde OCR Hatalarına Dayanıklı Karakter Tabanlı Metin Benzerliği Tespiti

Ahmet Demir[1], Seda Kırtay Demir[2]

Tarihi el yazması belgelerin dijitalleştirilmesi sürecinde ortaya çıkan Optik Karakter Tanıma (OCR) hataları, arkaik dil yapıları ve tutarsız imla varyasyonları, geleneksel kelime tabanlı metin benzerliği yöntemlerinin performansını önemli ölçüde düşürmektedir. Bu çalışmada iki aşamalı karakter tabanlı bir metin eşleştirme mimarisi önerilmiştir. Böylece yüksek karakter hata oranlarına rağmen alıntı ve metin benzerliği tespiti yapılabilmektedir. Sistem, hesaplama maliyetini azaltmak amacıyla iki seviyeli olarak kaba aday ve hassas eşleşme olarak tasarlanmıştır. İlk aşamada metinler karakter düzeyinde bigram yapılarına ayrıştırılmıştır. k=2 shingle yöntemiyle hash temsillerine dönüştürülmüştür. Ters dizin mimarisinde saklanmıştır. Böylece kelime sınırlarına bağımlılık ortadan kaldırılmıştır. Dolayısıyla hızlı aday belge erişimi sağlanmıştır. İkinci aşamada ise aday belgeler karakter tabanlı TF-IDF yöntemi ile vektör uzayına aktarılmıştır. Kosinüs Benzerliği metriği kullanılarak hassas benzerlik analizi gerçekleştirilmiştir. Önerilen yaklaşımın performansı, ICFHR 2016 READ Project veri setine ait 9.406 gerçek satır verisi üzerinde değerlendirilmiştir. Deneysel analizlerde sorgu metinlerine %5, %15 ve %30 oranlarında yapay OCR karakter bozulmaları uygulanarak sistemin dayanıklılığı incelenmiştir. Sonuçlar, önerilen yöntemin yüksek hata oranları altında dahi güçlü performans sergilediğini göstermiştir. Sistem, %5 ve %15 CER seviyelerinde %100 Recall@10 başarısı elde ederken, %30 CER gibi ağır bozulma senaryosunda %91,87 Recall@1 ve %99,19 Recall@10 doğruluğuna ulaşmıştır. Elde edilen bulgular, önerilen karakter tabanlı mimarinin büyük ölçekli tarihi arşivler ve dijital kütüphaneler için hızlı, ölçeklenebilir ve yüksek doğruluklu bir metin benzerliği çözümü sunduğunu göstermektedir.

Anahtar Kelimeler: Dijital Arşivler, Karakter Tabanlı Eşleştirme, Metin Benzerliği Tespiti, Optik Karakter Tanıma (OCR), Tarihi El Yazmaları

[1] Eskişehir Osmangazi Üniversitesi, ahmet.demir@ogu.edu.tr, 0000-0003-2319-4592
[2] Marmara Üniversitesi, seda.kirtay@marmara.edu.tr, 0000-0003-2415-9131