SMS SPAM

Sms spam datası ile yapılan çalışmada öncelikle problemin tanımı, veri seti hikayesi ve parametrelerin neyi fade ettiği araştırıldı ve readMe file içerisine eklendi.

Yapılan incelemede veri setinde bulunan mesajların farklı bir çok dil içerdiği emojiler içerdiği, web siteleri içerdiği, telefon numarası gibi sayısal veriler içerdiği, farklı noktalama işaretleri içerdiği tespit edildi. Tespit edilen başlıklardan web siteleri, emojiler ve fazla boşluklar kaldırıldı. Bunlara ek olarak büyük küçük harf durumu da incelenerek düzenlendi. Noktalama işaretleri ve özel karakterler bazı dillerin cümle yapısı için anlam ifade ettiği gerekçesiyle bırakıldı. Sayısal veriler de anlam bütünlüğünü koruma adına bırakıldı. Hedef değişken üzerinde encoding işlemi uygulandı. Dataset multilanguage olduğu göz önünde bulundurularak mBert modelinin uygun olacağına karar verildi ve mBert modeline uygun dataset hazırlandı. Pretrained model indirilerek hazırlanan veri le Fine tune edildi. Validasyonu yapıldı ve single text’e pred atılabilecek prediction fonksiyonu hazırlandı. Elde edilen başarılar ve model çıktıları mBert.pynb dosyasında mevcuttur. İlgili dosyalar github ortamına yüklenmiştir model boyutunun bğyğk olması sebebiyle yüklenemedi.