Metin Sınıflandırma Raporu

Hazırlayan: İzzet Efe Demirci

Giriş

Bu projede, farklı haber kategorilerine ait kısa metinlerin sınıflandırılması problemi ele alınmıştır. Amacımız, doğal dil işleme (NLP) ve makine öğrenmesi tekniklerini kulalnarak bir haberin içeriğine göre hangi kategoriye ait olduğunu yüksek doğrulukla tahmin edebilen bir model geliştirmektir. Bu doğrultuda açık kaynaklı AG News veri seti kullanılarak çeşitli modellerle metin sınıflandırma yapılmış ve sonuçları karşılaştırılmıştır.

Veri Seti

Proje kapsamında kullanılan veri seti, Kaggle – AG News Classification Dataset kaynağından temin edilmiştir. Veri seti;

- 4 farklı haber kategorisini (World, Sports, Business, Sci/Tech) içermektedir.
- Her veri örneği bir haber başlığı (Title) ve açıklamasından (Description) oluşmaktadır.
- Eğitim ve test veri seti ayrı dosyalar (train.csv ve test.csv) olarak sunulmuştur.

Çözüm Yaklaşımı

Proje çözümü aşağıdaki temel adımları içermektedir:

1) **Veri Ön İşleme**
   Metinler aşağıdaki işlemlerden geçirilmiştir:
   - Tüm metinler küçük harflere dönüştürülmüştür.
   - Noktalama işaretleri ve özel karakterler temizlenmiştir.
   - Gereksiz boşluklar, URL'ler ve rakamlar kaldırılmıştır.
   - Metinler, TfidfVectorizer kullanılarak sayısal vektörlere dönüştürülmüştür.

2) **Modelleme**
   Aşağıdaki gözetimli öğrenme algoritmaları kullanılarak modeller eğitilmiştir:
   - **Naive Bayes (MultinomialNB)**
   - **Logistic Regression**
   - **Linear Support Vector Classifier (LinearSVC)**

   Her bir model aynı eğitim ve test veri seti üzerinde değerlendirilmiş, karşılaştırmalı sonuçlar elde edilmiştir.

3) **Değerlendirme**
   - **Doğruluk (Accuracy):** Modelin test verisindeki genel başarı durumu.
   - **Confusion Matrix:** Her sınıf için doğru ve yanlış tahminlerin analizi.

Değerlendirme

Yapılan deneyler sonucunda en iyi performansı LinearSVC modeli göstermiştir. Aşağıda her modelin yaklaşık doğruluk oranları verilmiştir:

Model	Doğruluk
Logistic Regression	~90.44%
Linear SVC	~90.28%
LNaive Bayes	~89.03%

Elde edilen sonuçlar, temel ön işleme ve TF-IDF yaklaşımıyla bile yüksek doğruluk elde edilebileceğini göstermektedir. Daha ileri seviyedeki modeller (örneğin BERT gibi transformer tabanlı modeller) ile bu doğruluk oranları daha da artırılabilir.

Sonuç

Bu projede, kısa metin sınıflandırma problemini çözmek için uygulanan yaklaşımlar detaylı bir şekilde analiz edilmiş ve farklı modellerin başarı oranları karşılaştırılmıştır. Elde edilen sonuçlar, temel makine öğrenmesi tekniklerinin doğru ön işleme adımlarıyla birlikte uygulandığında etkili sonuçlar verebildiğini göstermektedir.