BBC Haber Veri Seti Sınıflandırma Analizi Raporu

Mehmet Ali YILDIZ

Mayıs 2025

Abstract

Bu raporda, BBC haber veri setindeki metinlerin kategorilere (business, entertainment, politics, sport, tech) ayrılması amacıyla çeşitli makine öğrenimi modellerinin performansı analiz edilmiştir. Sekiz farklı model (Linear SVM, Logistic Regression, MLP, Naive Bayes, Random Forest, XGBoost, KNN, Decision Tree) kullanılmış ve performansları karmaşıklık matrisleri, sınıf bazlı metrikler (precision, recall, F1-score) ve genel metrikler (accuracy, specificity, MCC) üzerinden değerlendirilmiştir. Linear SVM ve Logistic Regression en iyi sonuçları verirken, Decision Tree en düşük performansı göstermiştir. Sport kategorisi tüm modellerde en yüksek başarıyı elde etmiştir.

1 Giriş

Bu çalışma, Kaggle üzerinden elde edilen BBC haber veri setindeki metinlerin kategorilere ayrılması amacıyla gerçekleştirilmiştir. Veri seti, beş farklı kategoriye (business, entertainment, politics, sport, tech) ait haber metinlerini içermektedir. Amaç, farklı makine öğrenimi modellerinin bu sınıflandırma görevindeki performanslarını karşılaştırmaktır. Analiz, Python ve scikit-learn kütüphaneleri kullanılarak yapılmış, modeller 5 katlı çapraz doğrulama (Stratified K-Fold) ile değerlendirilmiştir.

2 Yöntem

2.1 Veri ve Ön İşleme

Veri seti, `bbc-text.csv` dosyasından yüklenmiş ve eksik değerler temizlenmiştir. Metinler, TF-IDF vektörleştirme yöntemiyle sayısal özelliklere dönüştürülmüştür (`max_features=5000`, İngilizce stop kelimeler hariç tutulmuştur). Kategoriler, LabelEncoder ile sayısal değerlere çevrilmiştir.

2.2 Modeller

Aşağıdaki sekiz model kullanılmıştır:

- Linear SVM
- Logistic Regression
- MLP (Multi-Layer Perceptron)

- Naive Bayes
- Random Forest
- XGBoost
- KNN (K-Nearest Neighbors)
- Decision Tree

2.3 Değerlendirme Metrikleri

Modellerin performansı şu metrikler üzerinden değerlendirilmiştir:

- Accuracy (Doğruluk)
- Precision (Kesinlik)
- Recall (Duyarlılık)
- Specificity (Özgüllük)
- F1 Score
- MCC (Matthews Correlation Coefficient)

Ayrıca, her model için karmaşıklık matrisleri ve sınıf bazlı metrikler (precision, recall, F1-score) hesaplanmıştır.

3 Sonuçlar

3.1 Karmaşıklık Matrisleri

Aşağıda, her modelin karmaşıklık matrisleri ayrı ayrı tablo formatında sunulmuştur. Satırlar gerçek sınıfları, sütunlar tahmin edilen sınıfları temsil eder. Diyagonaldeki değerler doğru tahminleri gösterir.

Linear SVM	**Tahmin**				
Gerçek	business	entertainment	politics	sport	tech
business	492	3	8	1	6
entertainment	2	381	2	0	1
politics	9	2	404	0	2
sport	1	0	0	510	0
tech	5	4	1	1	390

Table 1: Karmaşıklık Matrisi - Linear SVM

Logistic Regression	**Tahmin**				
Gerçek	business	entertainment	politics	sport	tech
business	495	0	6	1	8
entertainment	2	380	3	0	1
politics	11	2	402	0	2
sport	1	0	0	510	0
tech	7	3	0	2	389

Table 2: Karmaşıklık Matrisi - Logistic Regression

MLP	**Tahmin**				
Gerçek	business	entertainment	politics	sport	tech
business	492	3	10	1	4
entertainment	3	378	4	0	1
politics	10	4	401	0	2
sport	1	0	0	510	0
tech	1	5	1	1	393

Table 3: Karmaşıklık Matrisi - MLP

Naive Bayes	**Tahmin**				
Gerçek	business	entertainment	politics	sport	tech
business	492	1	9	0	8
entertainment	3	375	4	0	4
politics	11	0	404	0	2
sport	1	0	0	510	0
tech	4	2	2	4	389

Table 4: Karmaşıklık Matrisi - Naive Bayes

Random Forest	**Tahmin**				
Gerçek	business	entertainment	politics	sport	tech
business	488	2	12	1	7
entertainment	8	365	5	4	4
politics	13	3	393	5	3
sport	3	0	0	508	0
tech	9	5	2	8	377

Table 5: Karmaşıklık Matrisi - Random Forest

XGBoost	**Tahmin**				
Gerçek	business	entertainment	politics	sport	tech
business	476	4	15	3	12
entertainment	7	366	6	3	4
politics	12	4	390	5	6
sport	2	0	1	506	2
tech	8	8	1	3	381

Table 6: Karmaşıklık Matrisi - XGBoost

KNN	**Tahmin**				
Gerçek	business	entertainment	politics	sport	tech
business	452	7	28	5	18
entertainment	14	351	8	3	10
politics	17	4	390	4	2
sport	7	3	6	495	0
tech	10	9	7	0	375

Table 7: Karmaşıklık Matrisi - KNN

Decision Tree	**Tahmin**				
Gerçek	business	entertainment	politics	sport	tech
business	389	21	45	19	36
entertainment	20	310	13	21	22
politics	32	22	343	10	10
sport	14	14	5	475	3
tech	36	25	15	9	316

Table 8: Karmaşıklık Matrisi - Decision Tree

3.2 Sınıf Bazlı Metrikler

Sınıf bazlı metrikler, Logistic Regression modeli için şu şekildedir:

Kategori	**Precision**	**Recall**	**F1-Score**
Business	0.96	0.97	0.96
Entertainment	0.99	0.98	0.99
Politics	0.98	0.96	0.97
Sport	0.99	1.00	1.00
Tech	0.97	0.97	0.97

Table 9: Sınıf Bazlı Precision, Recall ve F1 Skorları (Logistic Regression)

Sport kategorisi mükemmel bir performans (F1=1.00) gösterirken, politics kategorisinde recall (0.96) biraz daha düşüktür.

3.3 Genel Performans Matrisi

Modellerin genel performans metrikleri aşağıdaki tabloda özetlenmiştir:

Model	Accuracy	Precision	Recall	Specificity	F1 Score	MCC
Linear SVM	0.978	0.978	0.978	0.995	0.978	0.973
Logistic Regression	0.978	0.978	0.977	0.994	0.978	0.972
MLP	0.977	0.976	0.977	0.994	0.977	0.971
Naive Bayes	0.975	0.975	0.975	0.994	0.975	0.969
Random Forest	0.958	0.956	0.956	0.989	0.957	0.947
XGBoost	0.952	0.952	0.951	0.988	0.952	0.940
KNN	0.927	0.927	0.927	0.982	0.927	0.909
Decision Tree	0.824	0.821	0.821	0.956	0.821	0.779

Table 10: Sınıflandırma Modelleri Performans Matrisi

Linear SVM ve Logistic Regression, tüm metriklerde en yüksek performansı göstermiştir.

4 Tartışma ve Sonuç

Yapılan analizler, Linear SVM ve Logistic Regression modellerinin BBC haber veri setinde en iyi performansı sergilediğini göstermektedir. Sport kategorisi tüm modellerde en yüksek başarıyı elde ederken, entertainment ve tech kategorilerinde bazı modeller (örneğin, Decision Tree) daha düşük performans göstermiştir. Decision Tree, tüm metriklerde en düşük sonuçları verdiği için bu veri seti için uygun bir model değildir.

4.1 Öneriler

- Entertainment ve tech kategorilerindeki recall değerlerini artırmak için veri setinde bu kategorilere ait daha fazla örnek eklenebilir.
- Decision Tree yerine Random Forest veya XGBoost gibi daha karmaşık modeller tercih edilmelidir.
- Linear SVM ve Logistic Regression modelleri, bu tür metin sınıflandırma görevleri için önerilir.

Bu çalışma, metin sınıflandırma projelerinde model seçiminin ve veri ön işleme adımlarının önemini ortaya koymuştur. Gelecek çalışmalar, hiperparametre optimizasyonu ve daha büyük veri setleri ile genişletilebilir.

5 Kaynakça

https://www.kaggle.com/datasets/moazeldsokyx/bbc-news