MAKİNE ÖĞRENMESİ İLE METİN SINIFLANDIRMA

Proje Adı: DFA Teknoloji Stajyer Teknik Seçim Görevleri - Görev 2: Makine Öğrenmesi ile Metin Sınıflandırma

Aday: Barış Aslan

Tarih: 23 Mayıs 2025

1. Giriş

Projenin temel amacı, kısa metinlerden oluşan bir veri seti kullanarak, her bir metni "Ekonomi", "Spor", "Magazin" ve "Gündem" gibi önceden belirlenmiş kategorilere otomatik olarak sınıflandıran bir Python tabanlı makine öğrenmesi modeli geliştirmektir. Bu çalışma, temel makine öğrenmesi modelleri ile metin analizi yapabilme becerisini değerlendirmeyi hedeflemektedir. Çalışma kapsamında, metinler üzerinde uygun ön işlemler uygulanmış ve en az iki farklı model kullanılarak performans karşılaştırması sunulmuştur.

2. Veri Seti

2.1. Kaynak Bu projede kullanılan veri seti, Kaggle platformunda yer alan "Turkish Headlines Dataset" adlı kamuya açık ve referans verilebilir bir veri setidir.

- **Veri Seti Adresi:** https://www.kaggle.com/datasets/anil1055/turkish-headlines-dataset

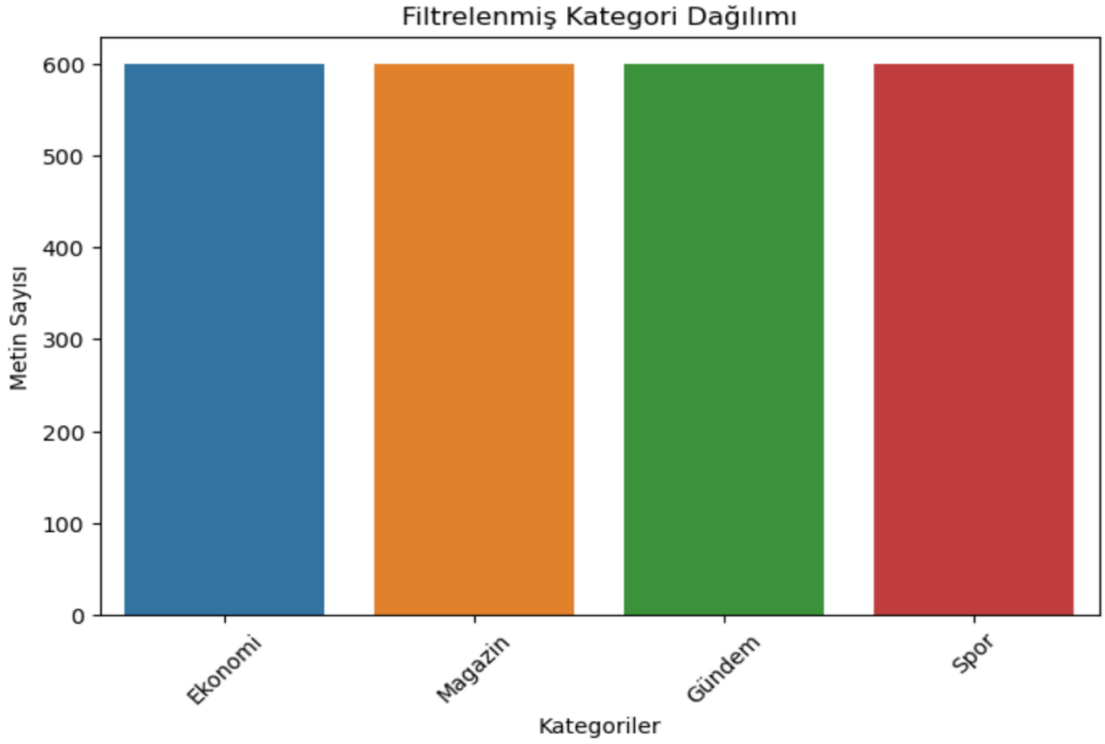
2.2. Açıklama "Turkish Headlines Dataset", çeşitli Türk haber sitelerinden toplanmış haber başlıklarını içermektedir. Orijinal veri seti 7 farklı kategori (ekonomi, siyaset, yaşam, teknoloji, magazin, sağlık, spor) ve her bir kategori için 600 başlık olmak üzere toplam 4200 haber başlığı içermektedir.

Bu proje kapsamında, görev tanımında belirtilen "Ekonomi", "Spor", "Magazin" ve "Gündem" kategorileriyle eşleşen veriler kullanılmıştır. Orijinal veri setindeki 'economy' kategorisi "Ekonomi", 'sport' kategorisi "Spor", 'magazine' kategorisi "Magazin" ve 'politics' kategorisi "Gündem" olarak eşleştirilmiştir. Bu dört kategoriye ait toplam 2400 haber başlığı (her kategoriden 600 adet) model geliştirme ve değerlendirme süreçlerinde kullanılmıştır.

```
Kullanılacak Kategoriler: ['Ekonomi', 'Spor', 'Magazin', 'Gündem']
Filtrelenmiş Veri Seti Boyutu: 2400

Eksik Veri Kontrolü (Filtrelenmiş):
HABERLER    0
ETIKET      0
dtype: int64

Filtrelenmiş ve Temizlenmiş 'ETIKET' Sütunundaki Kategori Dağılımı:
ETIKET
Ekonomi    600
Magazin    600
Gündem     600
Spor       600
Name: count, dtype: int64
```



3. Ön İşlemler

Metin verilerinin makine öğrenmesi modelleri tarafından işlenebilir hale getirilmesi için aşağıdaki temel ön işlem adımları uygulanmıştır:

- **Veri Temizleme:**
  - Tüm metinler küçük harfe dönüştürülmüştür.
  - Noktalama işaretleri, sayılar ve özel karakterler metinlerden çıkarılmıştır.
  - Fazla boşluklar kaldırılarak metinler standart hale getirilmiştir.
- **Tokenizasyon (Tokenization):**
  - Temizlenmiş metinler kelimelerine (token) ayrılmıştır.
- **Vektörleştirme (Vectorization):**
  - Tokenize edilmiş metinler, makine öğrenmesi algoritmalarının anlayabileceği sayısal formata dönüştürülmüştür. Bu aşamada, metinlerdeki kelimelerin önemini temsil eden TF-IDF (Term Frequency-Inverse Document Frequency) yöntemi kullanılmıştır. TF-IDF, bir kelimenin bir dokümandaki frekansını, o kelimenin tüm koleksiyondaki (corpus) genel frekansına göre ağırlıklandırır.

4. Modeller ve Eğitim

Proje kapsamında, metin sınıflandırma problemine uygun ve farklı çalışma prensiplerine sahip en az iki makine öğrenmesi modeli seçilerek eğitilmiş ve performansları karşılaştırılmıştır.

4.1. Veri Bölme Ön işlemlerden geçirilmiş ve vektörleştirilmiş veri seti, modelin eğitimi ve performansının objektif bir şekilde değerlendirilmesi için eğitim (%80) ve test (%20) olmak üzere iki alt kümeye ayrılmıştır.

4.2. Kullanılan Modeller

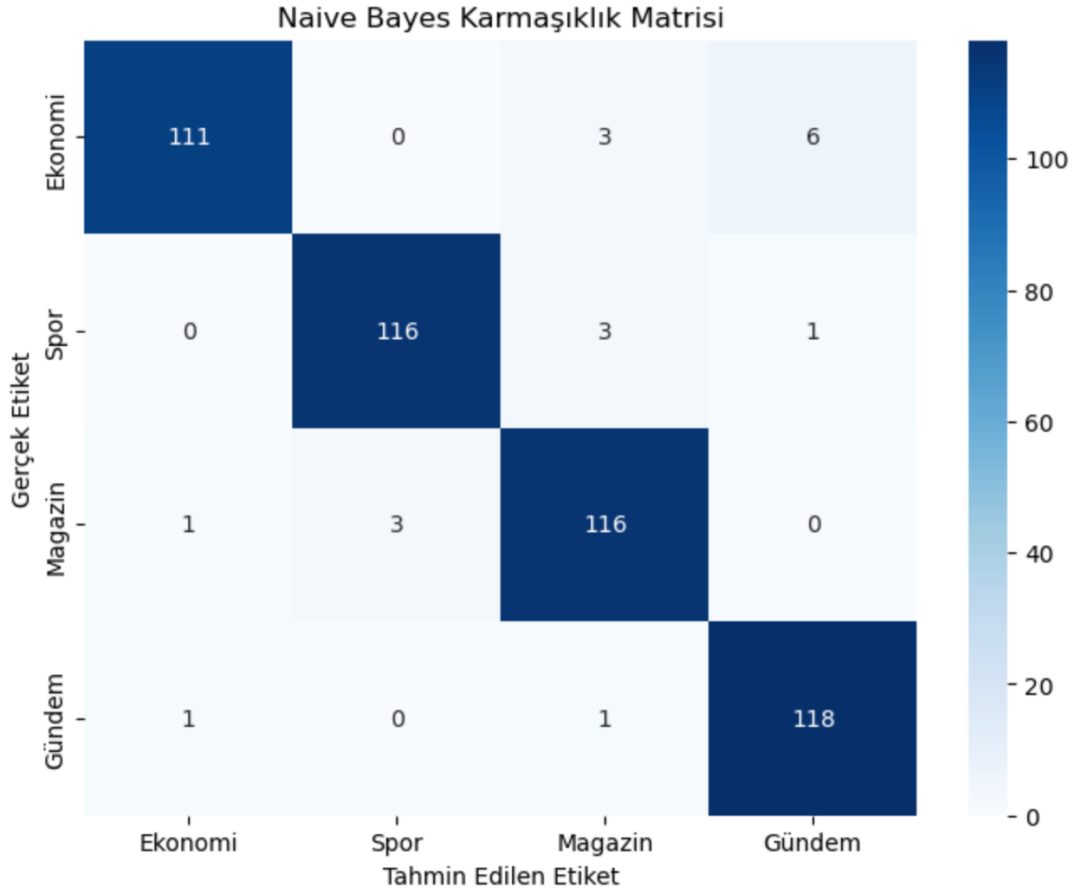
- **Model 1: Multinomial Naive Bayes (MNB)**
  - **Açıklama:** Naive Bayes sınıflandırıcıları, Bayes teoremine dayanan olasılıksal modellerdir. Özellikle metin sınıflandırma gibi görevlerde basitliği ve etkinliği ile bilinir. Multinomial Naive Bayes, kelime sayıları gibi ayrık özelliklerle iyi çalışır.
  - **Eğitim:** TF-IDF ile vektörleştirilmiş eğitim verileri kullanılarak MNB modeli eğitilmiştir.

```
Naive Bayes Sınıflandırma Raporu:
              precision    recall  f1-score   support

     Ekonomi       0.98      0.93      0.95       120
        Spor       0.94      0.98      0.96       120
     Magazin       0.94      0.97      0.95       120
      Gündem       0.97      0.97      0.97       120

    accuracy                           0.96       480
   macro avg       0.96      0.96      0.96       480
weighted avg       0.96      0.96      0.96       480
```

Naive Bayes Karmaşıklık Matrisi:



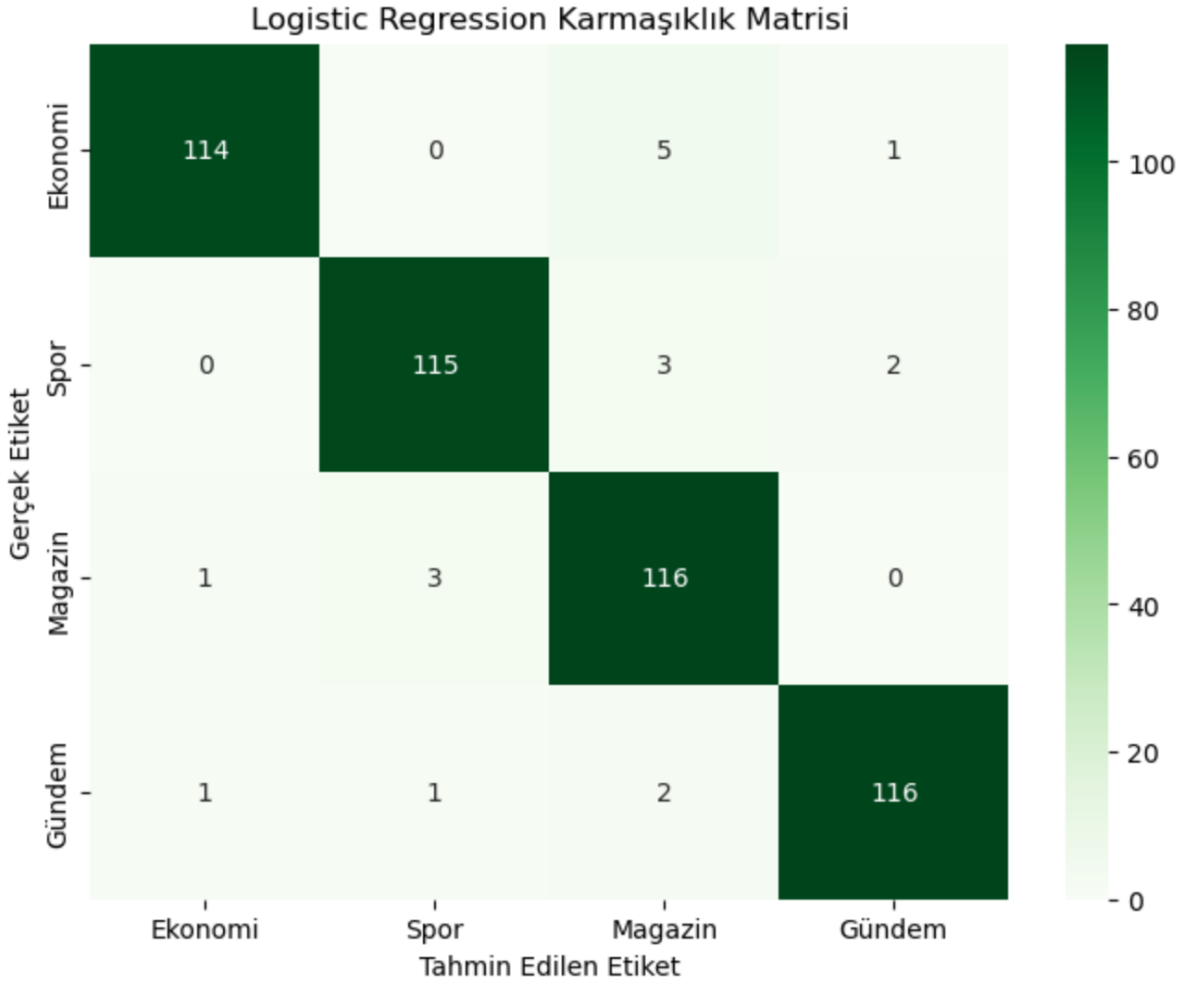
- **Model 2: Destek Vektör Makineleri - LinearSVC (Support Vector Machines - SVM)**
    - **Açıklama:** SVM, verileri en iyi ayıran hiperdüzlemi bulmayı amaçlayan güçlü bir sınıflandırma algoritmasıdır. Özellikle yüksek boyutlu uzaylarda (metin verilerinde olduğu gibi) etkili sonuçlar verebilir. LinearSVC, doğrusal bir çekirdek (kernel) kullanan bir SVM varyasyonudur.
    - **Eğitim:** TF-IDF ile vektörleştirilmiş eğitim verileri kullanılarak LinearSVC modeli eğitilmiştir.

```
Logistic Regression Sınıflandırma Raporu:
              precision    recall  f1-score   support

     Ekonomi       0.98      0.95      0.97       120
        Spor       0.97      0.97      0.97       120
     Magazin       0.92      0.97      0.94       120
      Gündem       0.97      0.96      0.96       120

    accuracy                           0.96       480
   macro avg       0.96      0.96      0.96       480
weighted avg       0.96      0.96      0.96       480
```

Logistic Regression Karmaşıklık Matrisi:



5. Değerlendirme ve Sonuçlar

Modellerin performansı, test seti üzerinde "Doğruluk (Accuracy)" metriği ve "Karmaşıklık Matrisi (Confusion Matrix)" kullanılarak değerlendirilmiştir.

5.1. Performans Metrikleri

- **Doğruluk (Accuracy):** Modelin doğru sınıflandırdığı örnek sayısının toplam örnek sayısına oranıdır.
- **Precision (Kesinlik):** Modelin pozitif olarak tahmin ettiği örneklerden kaçının gerçekten pozitif olduğudur. Her sınıf için ayrı ayrı hesaplanır.
- **Recall (Duyarlılık):** Gerçekte pozitif olan örneklerden kaçının model tarafından doğru bir şekilde pozitif olarak tahmin edildiğidir. Her sınıf için ayrı ayrı hesaplanır.
- **F1-Skoru:** Precision ve recall metriklerinin harmonik ortalamasıdır. Modelin dengeli bir performans gösterip göstermediğini anlamak için kullanılır.

5.2. Model Performansları

Aşağıdaki tabloda, eğitilen modellerin test seti üzerindeki doğruluk skorları ve karmaşıklık matrisleri sunulmuştur.

Tablo 1: Model Doğruluk Skorları

Model Adı	Doğruluk (Accuracy)
Multinomial Naive Bayes	%96.04
LinearSVC (SVM)	%96.04

Karmaşıklık Matrisi Raporu:

Her bir model için karmaşıklık matrisleri aşağıda sunulmuştur. Matrisler, satırlarda gerçek sınıfları, sütunlarda ise model tarafından tahmin edilen sınıfları göstermektedir.

Multinomial Naive Bayes - Karmaşıklık Matrisi :

Kategori	Precision	Recall	F1-Score	Support
Ekonomi	0.98	0.93	0.95	120
Spor	0.94	0.98	0.96	120
Magazin	0.94	0.97	0.95	120
Gündem	0.97	0.97	0.97	120
Genel	0.96	0.96	0.96	480

LinearSVC (SVM) - Karmaşıklık Matrisi :

Kategori	Precision	Recall	F1-Score	Support
Ekonomi	0.98	0.95	0.97	120
Spor	0.97	0.97	0.97	120
Magazin	0.92	0.97	0.94	120
Gündem	0.97	0.96	0.96	120
Genel	0.96	0.96	0.96	480

5.3. Sonuçların Yorumlanması

Her iki model de (Multinomial Naive Bayes ve Logistic Regression) test seti üzerinde %96.04 gibi yüksek ve aynı genel doğruluk oranını elde etmiştir. Bu, seçilen problem ve veri seti için her iki modelin de oldukça etkili olduğunu göstermektedir.

Sınıf bazlı performanslara bakıldığında küçük farklılıklar gözlemlenmektedir:

- **Multinomial Naive Bayes:** "Ekonomi" sınıfında çok yüksek precision, "Spor" sınıfında ise çok yüksek recall göstermiştir. Diğer sınıflarda da dengeli bir performans sergilemiştir.
- **Logistic Regression:** "Ekonomi" ve "Spor" sınıflarında çok dengeli ve yüksek metrikler sunarken, "Magazin" sınıfında precision değeri Naive Bayes'e göre biraz daha düşüktür.

Her iki modelin de makro ortalama (macro avg) ve ağırlıklı ortalama (weighted avg) F1 skorları 0.96'dır, bu da genel sınıflandırma performanslarının çok benzer ve başarılı olduğunu teyit eder. Verilen veri seti için, her iki model de tercih edilebilir. Model seçiminde, belirli bir sınıf için precision veya recall'un önceliğine ya da modelin eğitim hızı/kaynak kullanımı gibi faktörlere göre karar verilebilir. Naive Bayes genellikle daha hızlı eğitilirken, Logistic Regression da yorumlanabilirliği yüksek bir modeldir.

6. Sonuç

Bu projede, "Turkish Headlines Dataset" kullanılarak Türkçe haber başlıklarını "Ekonomi", "Spor", "Magazin" ve "Gündem" kategorilerine ayıran metin sınıflandırma modelleri geliştirilmiş ve karşılaştırılmıştır. Veri ön işleme adımları uygulanmış, ardından Multinomial Naive Bayes ve Logistic Regression modelleri eğitilerek performansları doğruluk ve detaylı sınıflandırma raporları ile değerlendirilmiştir.

Her iki model de %96.04'lük yüksek bir genel doğruluk oranı elde ederek, bu görev için uygun olduklarını kanıtlamışlardır. Sınıf bazında küçük farklılıklar olsa da, genel performansları birbirine çok yakındır. Bu sonuçlar, temel makine öğrenmesi modelleriyle dahi etkili metin sınıflandırma sistemleri geliştirilebileceğini göstermektedir.