MHD NOUREDDIN GHABBOUR
Mail: nourghapor1@gmail.com - Tel: 53143613739 Tarih: 23.05.2025

# Makine Öğrenmesi ile Metin Sınıflandırma Projesi Raporu

## 1. Giriş

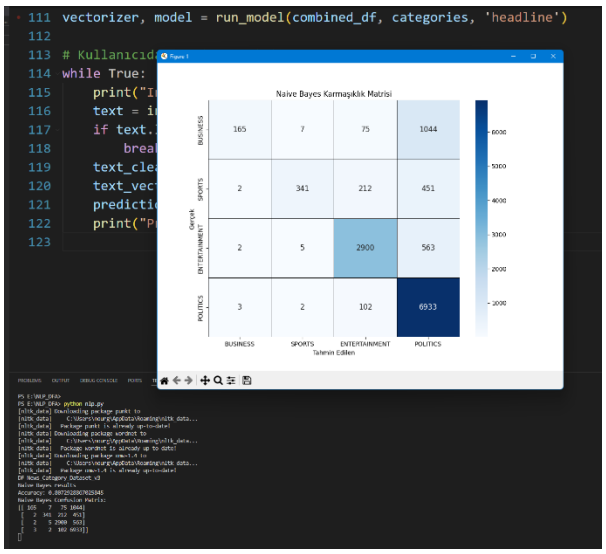
Bu proje kapsamında, doğal dil işleme (NLP) ve makine öğrenmesi teknikleri kullanılarak kısa metinlerin önceden belirlenmiş kategorilere otomatik olarak sınıflandırılması hedeflenmiştir. Metin sınıflandırma, özellikle haber metinleri gibi büyük ve hızlı üretilen veri kaynaklarında, bilgilerin düzenlenmesi ve analiz edilmesi için önemli bir yöntemdir.

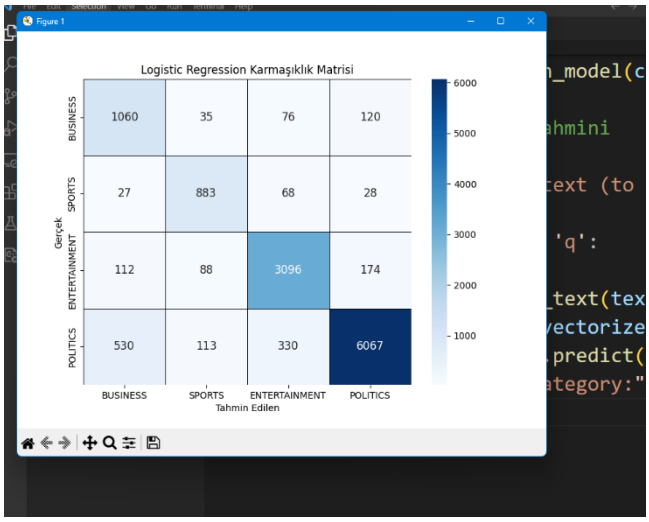
Projede, Ekonomi, Spor, Magazin ve Gündem gibi farklı kategorilere ait haber başlıkları kullanılmıştır. Metinler, temel ön işlemlerden geçirilerek sayısal özelliklere dönüştürülmüş ve Naive Bayes ile Logistic Regression gibi modellerle eğitilmiştir. Bu sayede, farklı metinler uygun kategorilere yüksek doğrulukla atanabilmiştir.

## 2. Veri Seti

1. Projede iki farklı İngilizce haber veri seti kullanılmıştır:
    1. **News Category Dataset**
        - Kaynak: Kaggle - News Category Dataset
        - Toplam örnek sayısı: yaklaşık 200.000
        - İçerik: Kullanılan Haber başlıkları (BUSINESS, SPORTS, ENTERTAINMENT, POLITICS).
    2. **AG News Classification Dataset**
        - Kaynak: Kaggle - AG News Dataset
        - Toplam örnek sayısı: yaklaşık 120.000
        - İçerik: Kullanılan Haber (POLITICS, SPORTS, BUSINESS, ENTERTAINMENT
2. Veri setlerinde kategori dağılımı dengesizdir:
    1. POLITICS ve ENTERTAINMENT kategorileri diğerlerine göre daha fazla örnek içerir.
    2. BUSINESS ve SPORTS kategorilerinde örnek sayısı daha azdır.

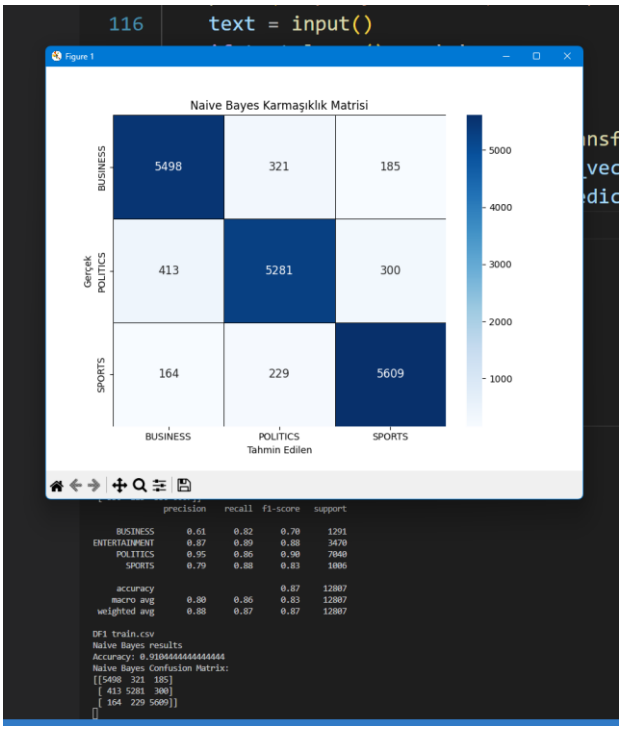
Aşağıda News Category Veri Seti sonuçları yer almaktadır:



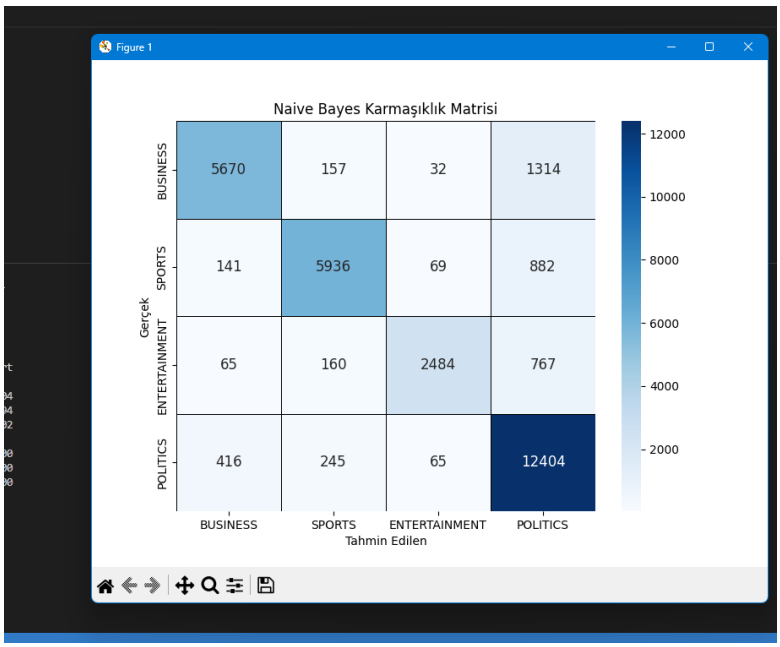



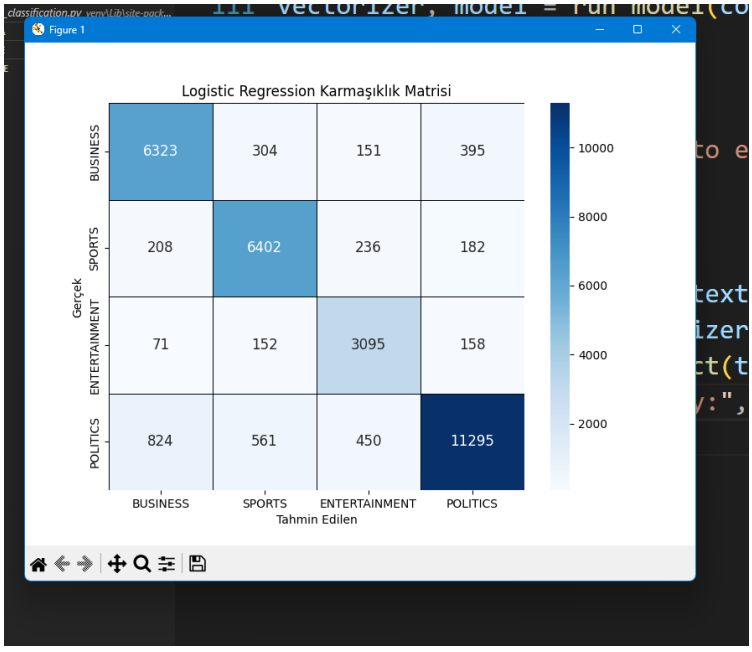
MHD NOUREDDIN GHABBOUR
Mail: nourghapor1@gmail.com - Tel: 53143613739 Tarih: 23.05.2025

Aşağıda AG News Classification Veri Seti sonuçları yer almaktadır:


Aşağıda iki farklı News Category Veri Seti'nin sonuçları yer almaktadır:

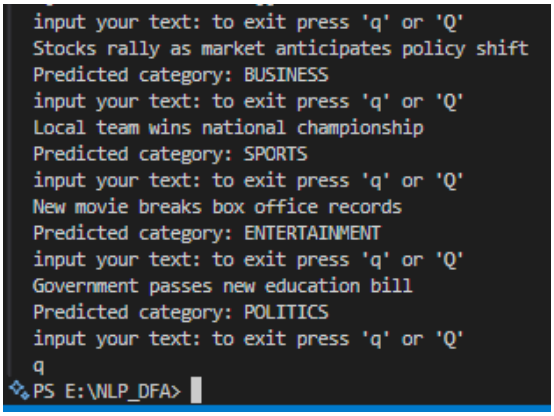


MHD NOUREDDIN GHABBOUR
Mail: nourghapor1@gmail.com - Tel: 53143613739 Tarih: 23.05.2025

Aşağıda örnek kategori dağılım grafiği ve örnek haber başlıkları gösterilmektedir:

| Örnek Haber Başlığı | Kategori |
|---|---|
| "Stocks rally as market anticipates policy shift" | BUSINESS |
| "Local team wins national championship" | SPORTS |
| "New movie breaks box office records" | ENTERTAINMENT |
| "Government passes new education bill" | POLITICS |




## 3. Proje Süreci

- Veri Toplama: Güvenilir kaynaklardan haber veri setleri indirilmiştir.
- Veriyi İnceleme ve Anlama: Veri yapısı, dağılım ve eksik değerler Pandas ile analiz edilmiştir.
- Temizleme ve Ön İşleme: Metinler küçük harfe çevrilmiş, gereksiz karakterler çıkarılmıştır.
- Tokenizasyon: Metinler kelimelere ayrılmıştır.
- Vectorizer Kullanma: TF-IDF yöntemi ile metinler sayısal verilere dönüştürülmüştür.
- Model Seçme ve Eğitme: Naive Bayes ve Logistic Regression modelleri eğitilmiştir.
- Modeli Test Etme: Doğruluk ve karışıklık matrisi ile performans ölçülmüştür.
- Kullanıcıdan Girdi Alarak Test Etme: Modelin gerçek zamanlı tahmin yeteneği test edilmiştir

| Adım No | Proje Adımı | Açıklama |
|---|---|---|
| 1 | Veri Toplama | Kaggle veya benzeri güvenilir kaynaklardan veri seti indirme. |
| 2 | Veriyi İnceleme ve Anlama | Pandas ile veri yapısı, dağılım ve eksik değerlerin analizi. |
| 3 | Temizleme ve Ön İşleme | Küçük harfe çevirme, noktalama işaretleri ve gereksiz karakterlerin temizlenmesi. |
| 4 | Tokenizasyon | Metnin kelimelere bölünerek model için uygun hale getirilmesi. |
| 5 | Vectorizer Kullanımı | TF-IDF veya CountVectorizer ile metnin sayısal verilere dönüştürülmesi. |
| 6 | Model Seçme ve Eğitme | Naive Bayes ve Logistic Regression gibi modellerin seçilmesi ve eğitilmesi. |
| 7 | Modeli Test Etme | Doğruluk, precision, recall, f1-score ve karışıklık matrisi ile model performansının değerlendirilmesi. |
| 8 | Kullanıcıdan Girdi Alarak Test Etme | Modelin gerçek zamanlı olarak kullanıcı girdileriyle test edilmesi. |

.

MHD NOUREDDIN GHABBOUR
Mail: nourghapor1@gmail.com - Tel: 53143613739 Tarih: 23.05.2025

## 4. Ön İşleme (Data Preprocessing)

Metin sınıflandırma problemlerinde ham veri genellikle doğrudan modellemeye uygun değildir. Verideki gürültü, anlamsız karakterler, büyük-küçük harf farklılıkları, noktalama işaretleri gibi öğeler, modelin performansını olumsuz etkileyebilir. Bu nedenle, model eğitimi öncesinde verinin belli standartlara göre temizlenmesi, işlenmesi gerekir. İşte bu süreç "ön işleme" olarak adlandırılır.

**Ön işlemenin başlıca amaçları şunlardır:**

- Verideki gereksiz karakterleri (noktalama işaretleri, sayılar, boşluklar vb.) temizlemek.
- Metni standartlaştırmak (örn. tümünü küçük harfe çevirmek) ve tutarlı hale getirmek.
- Kelimeleri modelin anlayacağı hale dönüştürmek için tokenizasyon yapmak.
- Benzer anlamdaki kelimeleri birleştirmek için lemmatization veya stemming uygulamak.
- Modelin karmaşıklığını ve eğitim süresini azaltmak.

Bu işlemler, modelin daha doğru ve hızlı öğrenmesini sağlar, ayrıca aşırı öğrenme (overfitting) riskini azaltır.

```python
def clean_text(text):
    # Temizleme işlemleri
    text = text.lower()
    text = text.replace('\n', ' ')
    text = re.sub(r'[^\w\s]', '', text)
    text = re.sub(r'\d+', '', text)
    tokens = text.split()  # basit tokenize
    lemmatized_tokens = [lemmatizer.lemmatize(word) for word in tokens]
    return ' '.join(lemmatized_tokens)
```

## 5. Modelleme

Bu projede iki farklı makine öğrenmesi sınıflandırma modeli kullanılmıştır:

1. **Multinomial Naive Bayes:**
   Metin sınıflandırma problemlerinde yaygın kullanılan, hızlı ve etkili bir modeldir. Kelimelerin olasılıklarına dayalı basit bir yaklaşım sunar.

2. **Logistic Regression:**
   İkili ve çok sınıflı sınıflandırmalarda kullanılan, doğrusal bir modeldir. Modelin eğitiminde max_iter=1000 parametresi iterasyon sayısını artırarak modelin daha iyi öğrenmesini sağlamıştır. Ayrıca, class_weight='balanced' parametresi kullanılarak, veri setindeki dengesiz sınıfların modelde daha eşit ağırlık alması sağlanmıştır.

Veri seti, %80 eğitim ve %20 test olacak şekilde rastgele bölünmüştür. Bu oran, modelin yeterince veri üzerinde öğrenmesini ve genel performansının test edilmesini mümkün kılar.

MHD NOUREDDIN GHABBOUR
Mail: nourghapor1@gmail.com - Tel: 53143613739 Tarih: 23.05.2025

Modeller, eğitim verisi üzerinde eğitilmiş ve test verisi üzerinde doğruluk, precision, recall gibi performans metrikleri kullanılarak değerlendirilmiştir.

```
Naive Bayes results
Accuracy: 0.8072928867025845
Naive Bayes Confusion Matrix:
[[ 165    7   75 1044]
 [   2  341  212  451]
 [   2    5 2900  563]
 [   3    2  102 6933]]
              precision    recall  f1-score   support

    BUSINESS       0.96      0.13      0.23      1291
ENTERTAINMENT      0.88      0.84      0.86      3470
    POLITICS       0.77      0.98      0.86      7040
      SPORTS       0.96      0.34      0.50      1006

    accuracy                           0.81     12807
   macro avg       0.89      0.57      0.61     12807
weighted avg       0.83      0.81      0.77     12807
```

```
Logistic Regression results
Accuracy: 0.8671820098383697
Logistic Regression Confusion Matrix:
[[1060   35   76  120]
 [  27  883   68   28]
 [ 112   88 3096  174]
 [ 530  113  330 6067]]
              precision    recall  f1-score   support

    BUSINESS       0.61      0.82      0.70      1291
ENTERTAINMENT      0.87      0.89      0.88      3470
    POLITICS       0.95      0.86      0.90      7040
      SPORTS       0.79      0.88      0.83      1006

    accuracy                           0.87     12807
   macro avg       0.80      0.86      0.83     12807
weighted avg       0.88      0.87      0.87     12807
```

## 6. Sonuçlar

- Logistic Regression modeli Naive Bayes modeline göre daha yüksek doğruluk ve dengeli performans göstermiştir.
- Logistic Regression için test doğruluğu yaklaşık %87, Naive Bayes için ise %81 olarak bulunmuştur.
- Precision, recall ve f1-score değerleri de Logistic Regression modelinde daha yüksektir.
- Model performansları confusion matrix ile görselleştirilmiştir ve sınıflar arasındaki karışıklıklar analiz edilmiştir.

## 7. Sonuçların Analizi

Projede kullanılan veri setindeki kategori dağılımının dengesiz olması, bazı sınıflarda özellikle recall değerlerinin düşük çıkmasına neden olmuştur. Örneğin BUSINESS kategorisinde örnek sayısının az olması, modelin bu sınıfı doğru tespit etmesini zorlaştırmıştır. Logistic Regression modelinde class_weight='balanced' parametresi bu durumu bir nebze iyileştirse de tam anlamıyla denge sağlanamamıştır.

Ayrıca, veri setlerinde bazen yanlış etiketlenmiş veya kategorisi net olmayan haber başlıkları bulunmuştur. Model, bu tür örnekleri genellikle mevcut kategorilerden birine atamış, dolayısıyla yanlış sınıflandırma oranı artmıştır.

Model karmaşıklığı ve basit ön işleme teknikleri, metinlerin bağlamsal ve anlamsal özelliklerini yeterince yakalayamayabilir. Bu da hatalara sebep olabilmektedir.

## 8. Kaynaklar

- News Category Dataset, Kaggle: https://www.kaggle.com/datasets/rmisra/news-category-dataset
- AG News Classification Dataset, Kaggle: https://www.kaggle.com/datasets/amananandrai/ag-news-classification-dataset
- Python, scikit-learn, nltk, imblearn kütüphaneleri

https://github.com/nourghapor/machine-learning-text-classification