****

**Air France Müşteri Yorumlarının Duygu Analizi ve Sınıflandırılması**

Merve Subaşı-132230015

Python ile Veri Bilimi

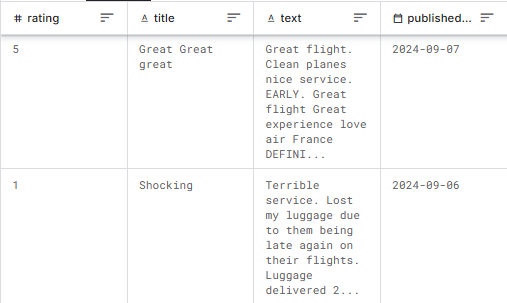
16.06.2025

**2. Giriş**

* **Proje konusu ve amacı:** Bu projede, Air France havayolu şirketine ait müşteri yorumları analiz edilerek yorumların duygu durumu (olumlu, olumsuz, nötr) sınıflandırılmıştır. Müşteri geri bildirimlerinin yapay zekâ algoritmaları ile otomatik olarak analiz edilmesi, hizmet kalitesinin objektif olarak değerlendirilmesine olanak tanımaktadır.
* **Seçilen veri madenciliği tekniği:** Projede sınıflandırma (classification) tekniği kullanılmıştır. "Sentiment" etiketi hedef değişken olarak belirlenmiştir.
* **Kısa yöntem özeti:** Veri ön işleme süreçlerinin ardından yorumlar temizlenmiş ve TF-IDF yöntemiyle sayısallaştırılmıştır. Naive Bayes, Logistic Regression ve Random Forest algoritmaları uygulanmış ve veriler %80 eğitim, %20 test olacak şekilde ayrılmıştır.

**3. Veri Seti Tanıtımı**

* **Veri setinin kaynağı:** Kaggle – [Air France Tripadvisor Reviews Dataset](https://www.kaggle.com/datasets/saharnazyaghoobpoor/air-france-reviews-dataset)
* **Veri setinin kısa açıklaması:** Bu veri seti, Air France havayolu şirketi hakkında yazılmış müşteri yorumlarını içermektedir. Yorumlar apify.com üzerinden etik web kazıma yöntemi ile toplanmış ve duygu etiketi ile etiketlenmiştir. Kişisel bilgiler ayıklanmış, veri anonimleştirilmiştir.
* **Alanlar:** Havacılık, Müşteri Hizmetleri, Sosyal Medya Analizi
* **Boyut:** 7.000+ yorum, 4 sütun (rating, title, text, publishedDate)



**4. Veri Ön İşleme**

**Eksik veriler:** Veri setindeki eksik veriler kontrol edilmiştir. rating, title, text ve publishedDate sütunlarında eksik kayıtlar bulunan satırlar analiz dışı bırakılmıştır. Eksik verilerin oranı düşük olduğu için silme yöntemi tercih edilmiştir.

**Aykırı Değerler:** Sayısal sütun olan rating değerleri 1 ile 5 arasında olduğundan, bu aralık dışındaki herhangi bir aykırı değer tespit edilmemiştir. Ancak yorum uzunlukları analiz edilmiş, aşırı kısa ya da spam benzeri içerikler (örneğin 3 kelimenin altındaki yorumlar) elenmiştir.

**Normalizasyon, kategorik verilerin dönüştürülmesi vs:** Metin verileri olan title ve text sütunları birleştirilerek zenginleştirilmiş ve küçük harfe çevrilmiştir. Noktalama işaretleri kaldırılmış, durak kelimeler (stop words) silinmiş ve lemmatizasyon uygulanmıştır. Sayısal öznitelik çıkarmak için TF-IDF yöntemi kullanılmıştır. rating kategorik olarak sınıflandırmaya dahil edilmemiş, sadece gözlemsel amaçla değerlendirilmiştir.

#### **Grafikler:**

Aşağıda açıklanan grafikler, raporun sonunda yer alan Ekler bölümündeki “Grafikler” başlığı altında detaylı şekilde sunulmuştur:

* Confusion Matrix: Logistic Regression modeli için oluşturulmuş, modelin sınıflandırma başarısını görselleştiren hata matrisi.
* Bar Grafikler: Sınıf dağılımları ve başarı oranları görsel olarak analiz edilmiştir. Ayrıca, olumlu yorum türleri ve şikayet türleri grafiklerine de yer verilmiştir.
* Word Cloud: Yorumlardaki en sık geçen kelimeler, duygu etiketlerine (pozitif/negatif) göre ayrılarak görselleştirilmiştir.
* Öne Çıkan Kelimeler (IDF): TF-IDF analizine göre en yüksek bilgi değerine sahip anahtar kelimeler sıralı şekilde sunulmuştur.

**Öncesi-Sonrası Özeti:**

| Aşama | Açıklama |
| --- | --- |
| Ham Veri | Eksik ve karışık satırlar, filtreleme yok |
| Temiz Veri | Eksikler temizlendi, metin ön işleme yapıldı |
| Vektörel Veri | TF-IDF ile dönüştürülmüş sayısal özellik matrisi |

**5. Yöntem ve Uygulama**

### **a) Sınıflandırma (Classification)**

Bu projede, Air France havayolu müşterilerine ait yorumlardan duygu (sentiment) sınıflandırması yapılmıştır. Amaç; her bir müşteri yorumunu "positive", "neutral" veya "negative" olarak otomatik etiketlemektir.

#### **Kullanılan Algoritmalar:**

* Naive Bayes
* Logistic Regression
* Random Forest Classifier

Bu algoritmalar, TF-IDF yöntemiyle sayısallaştırılmış metin verileri üzerinde eğitilmiştir.

#### **Eğitim/Test Ayrımı:**

* Veriler %80 eğitim, %20 test olacak şekilde train\_test\_split() fonksiyonu ile bölünmüştür.

#### **Başarı Ölçütleri:**

Her bir modelin başarısı aşağıdaki metriklerle değerlendirilmiştir:

* Accuracy (Doğruluk): Genel başarı oranı
* Precision: Doğru pozitif tahmin oranı
* Recall: Gerçek pozitiflerin ne kadarını yakalayabildiği
* F1-score: Precision ve Recall’un dengeli ortalaması

**Karar:**

* Logistic Regression ikili sınıflandırmada %85 başarı oranına ulaştı.
* Üçlü sınıflandırmada Random Forest dengeli performans gösterdi.

**6. Sonuç ve Yorumlar**

**Bulgular:**

* En başarılı model, üç sınıflı duygu analizi için Random Forest olmuştur.
* "Positive" ve "Negative" sınıflarında oldukça yüksek başarılar elde edilmiştir (F1-score ≈ 0.85).
* "Neutral" sınıf ise zayıf performans göstermiştir (F1-score ≈ 0.12–0.13). Bu durum, veri dengesizliği ve nötr sınıfın yorum açısından belirsizliğinden kaynaklanabilir.

**Uygulamanın Güçlü Yönleri:**

* Yüksek başarı oranı ve tutarlı sınıflandırma sonuçları
* Yorum metinlerinden anlamlı bilgi çıkarımı
* Modeller arası kıyaslama ve metriklerin detaylı kullanımı

### **Zayıf Yönleri:**

* Neutral sınıf için düşük model başarısı
* Sınıflar arası dengesizlik
* Sadece klasik ML algoritmaları kullanıldı, derin öğrenme denenmedi

### **Geliştirilebilecek Yönler:**

* Neutral sınıf için daha fazla veri toplanması
* Daha gelişmiş metin temsil yöntemleri (örneğin BERT) kullanılması
* Derin öğrenme tabanlı modellerle kıyaslama yapılması

### **Sonuçların Yorumlanması:**

* Air France müşteri yorumları, genel olarak pozitif (4–5 puan) ve negatif (1–2 puan) uçlarda yoğunlaşmaktadır. Bu durum, müşterilerin deneyimlerini net bir şekilde olumlu veya olumsuz olarak değerlendirdiğini göstermektedir.
* Nötr (3 puan) yorum sayısı görece azdır. Ayrıca bu sınıftaki yorumlar, ifade belirsizliği nedeniyle sınıflandırma modelleri açısından daha zayıf performans göstermiştir *(F1-score ≈ 0.12–0.13)*.
* TF-IDF yöntemi ile metinlerden çıkarılan öznitelikler sayesinde, yorumların içerdiği duygu tonları sayısal olarak analiz edilebilmiş, bu da model başarısını doğrudan artırmıştır.
* Özellikle yüksek IDF değerlerine sahip nadir kelimeler, detaylı analizlerde öne çıkan şikâyet veya memnuniyet unsurlarını tespit etmede etkili olmuştur.

**İşletme Açısından Stratejik Değerlendirme**

Bu analiz, yalnızca müşteri şikayetlerinin tespitiyle kalmayıp, elde edilen verilerin işletmenin operasyonel ve stratejik kararlarına nasıl yön verebileceğini göstermektedir.

Stratejik öneriler:

* Gecikme ve rötar şikayetlerinin yüksek olduğu uçuşlarda zaman yönetimi ve operasyonel iyileştirmeler yapılabilir.
* Koltuk konforu ile ilgili şikayetler dikkate alınarak kabin içi donanım geliştirmeleri planlanabilir.
* Personel ilgisi konusundaki olumsuz geri bildirimlere karşı hizmet içi eğitim programları düzenlenebilir.
* Bagaj kayıplarını azaltmaya yönelik süreçler güçlendirilerek müşteri memnuniyeti artırılabilir.
* Diğer şikayet kategorileri için hedef odaklı çözümler geliştirilerek müşteri deneyimi iyileştirilebilir.

Olumlu yorumlarda personel hizmeti ile konfor ve rahatlık konuları öne çıkmakla birlikte, olumsuz yorumlarda bu alanlara yönelik şikayetlerin daha yoğun olduğu gözlemlenmektedir. Bu durum, özellikle personel ilgisi ve kabin konforu konularında iyileştirmelerin öncelikli olarak ele alınması gerektiğini göstermektedir.

Bu sonuca ilişkin olumlu yorumların veolumsuz yorumların kategorik bar grafiklerini, *“Grafikler”* başlığı altında inceleyebilirsiniz.

Bu çıkarımlar, veri analizinden elde edilen somut sonuçlar doğrultusunda işletmenin müşteri memnuniyetini artırmaya yönelik karar alma süreçlerine önemli katkılar sağlar.

## **7. Kaynakça**

### **Veri Seti Kaynağı:**

*Air France Tripadvisor Reviews Dataset*. Kaggle.  
 Erişim: <https://www.kaggle.com/datasets/saharnazyaghoobpoor/air-france-reviews-dataset>

**Python Kütüphaneleri:**

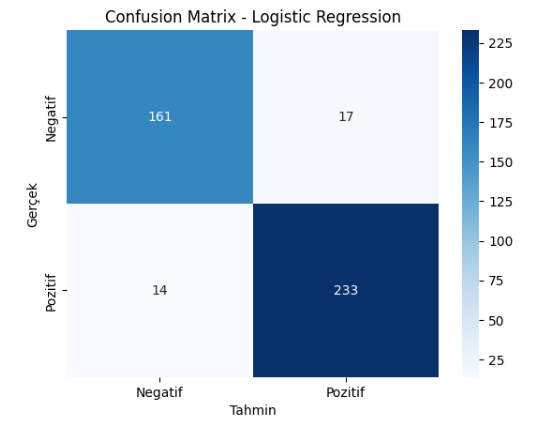
* Scikit-learn: <https://scikit-learn.org>
* Pandas: <https://pandas.pydata.org>
* Matplotlib: <https://matplotlib.org>
* Seaborn:<https://seaborn.pydata.org>
* WordCloud: <https://amueller.github.io/word_cloud/>

### **8. Ekler**

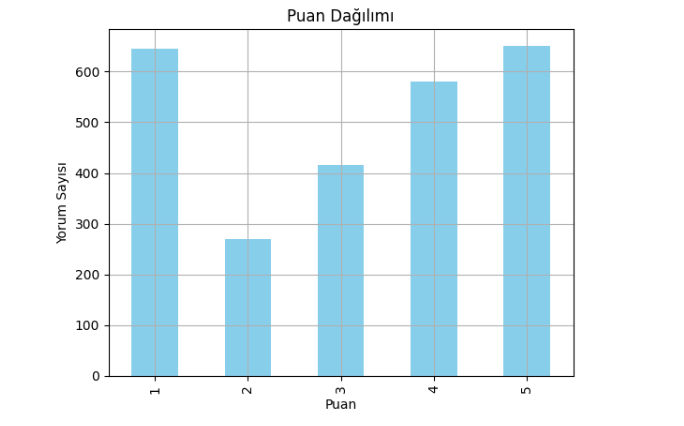
**Kodlar:** Tüm Python kodları GitHub üzerinde paylaşılmıştır:  
<https://github.com/githubmerve/NLP-Project-Sentiment-Analysis-of-Air-France-Reviews>

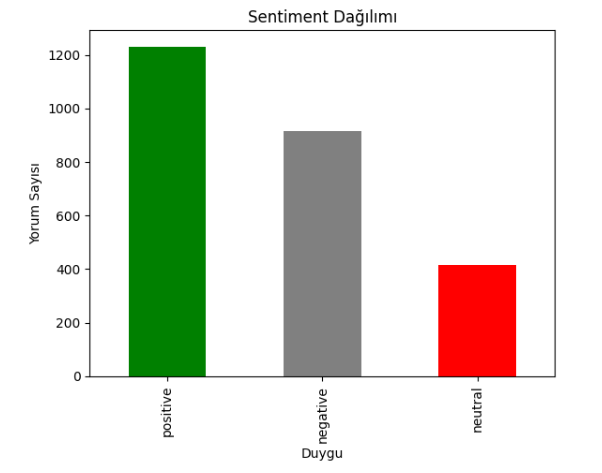
**Grafikler:**

* Confusion Matrix



* Sentiment ve rating dağılımları

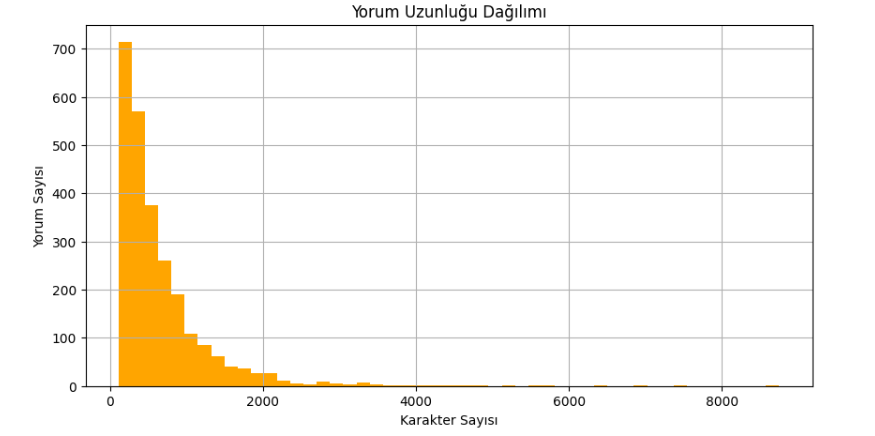




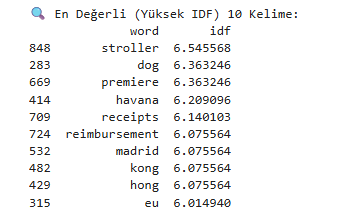
* WordCloud



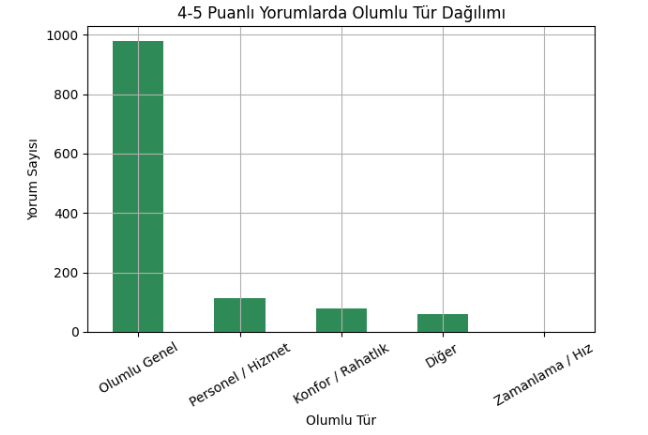
* En uzun ve kısa yorumlar listesi



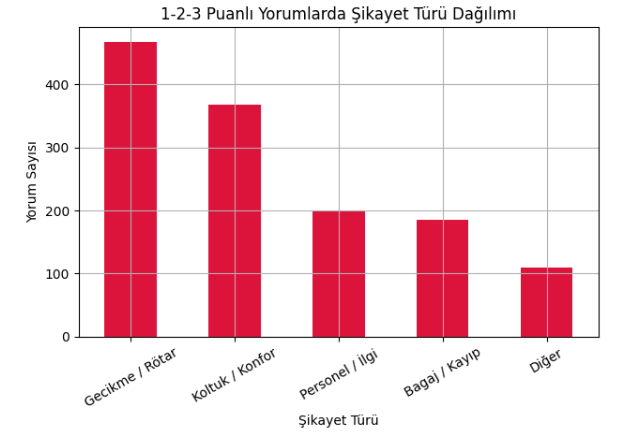
* En yüksek IDF değerine sahip kelimeler tablosu



* Olumlu Yorumların Kategorik Bar Grafiği



* Olumsuz Yorumların Kategorik Bar Grafiği



**Notlar:**

* Kodlar Jupyter Notebook (.ipynb) formatındadır.
* Çalışma Kaggle ortamında gerçekleştirilmiştir.