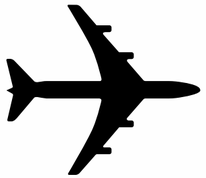
LET'S PLAY

TRIPADVISOR YORUM ANALIZI

NATUREL LANGUAGE PROCESSING İLE

MURAT DEMİRBAŞ

KONULAR

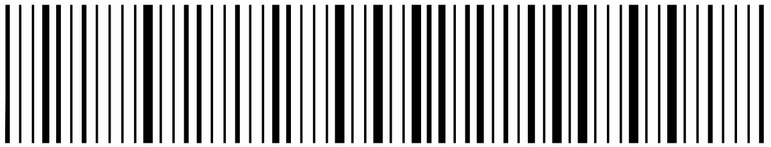


01

Problem & Genel Bakış

02

Veri Analizi



03

Önişleme & Vektörleme

04

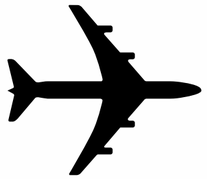
Modelleme

01

PROBLEM

"DUSUK", "ORTALAMA" VE "HARIKA" OLARAK KATEGORIZE ETMEK. BU, TÜKETICILERIN VE IŞLETME SAHIPLERININ HER BIR INCELEME HISSIYATINI DAHA IYI ANLAMALARINA YARDIMCI OLUR. HÂLIHAZIRDA ÖNCEDEN AYARLANMIŞ INCELEME VEYA DERECELENDIRME SISTEMINE SAHIP OLMAYAN SISTEMLER IÇIN FAYDALIDIR.

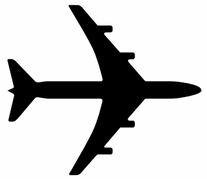
TRIPADVISOR



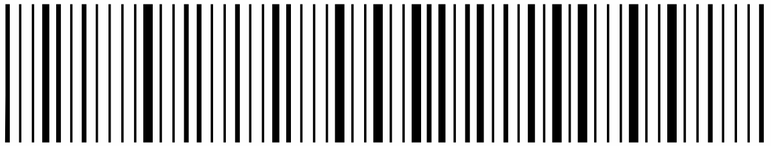
Tripadvisor Dünyanın en büyük seyahat rehberlik platformu Gezginlerin plan yapmalarına, rezervasyon yapmalarına ve seyahat etmelerine yardımcı olur Gezginlerin nerede kalacaklarını, yemek yiyeceklerini ve uyuyacaklarını keşfetmelerine yardımcı olur Dünya genelinde 8 milyon işletme hakkında 884 milyondan fazla yorum

---

VERİ SETİ



"Tripadvisor'dan 3 farklı şehirden 3 farklı restorana ait yorumlar çekilerek veri seti oluşturuldu."



02

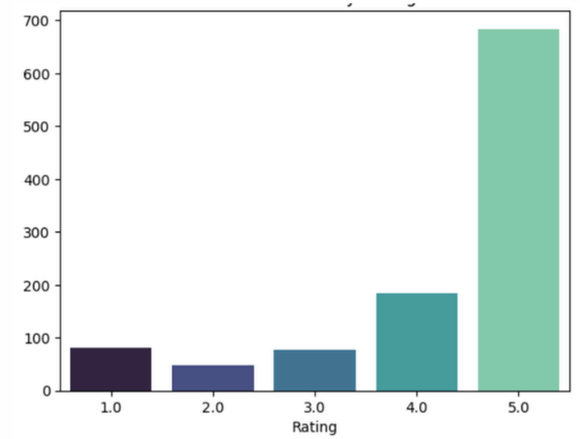
VERI ANALIZI

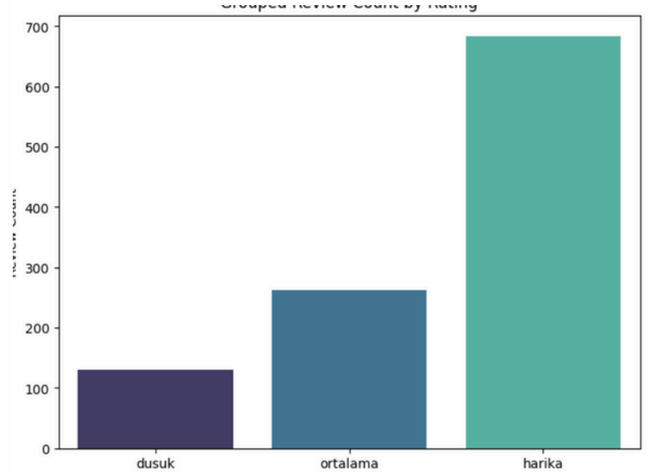
ORIJINAL VERI SETINDEKI SINIF DENGESIZLIĞI NEDENIYLE, DERECELENDIRME PUANLARINI BIR ARAYA GETIRILDI:

1 VE 2 = 'DUSUK'

3 VE 4 = 'ORTALAMA'

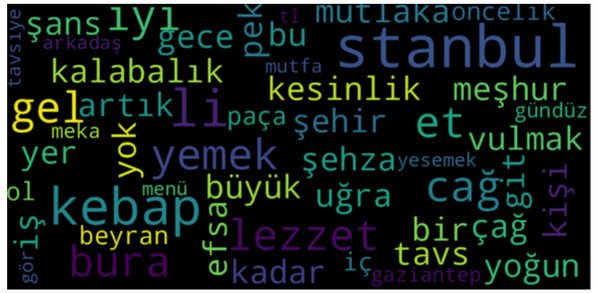
5 = 'HARIKA’





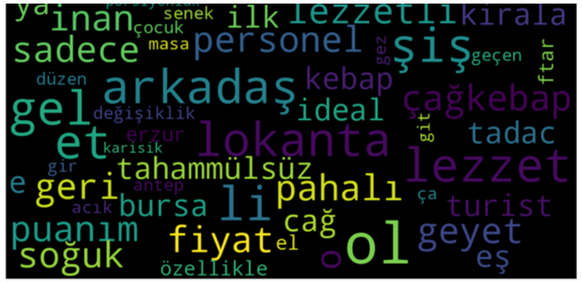
KELIME BULUTU

Her kategori için belirli kelimelerin ne kadar yaygın olduğunu gösteren kelime bulutları oluşturuldu



harika

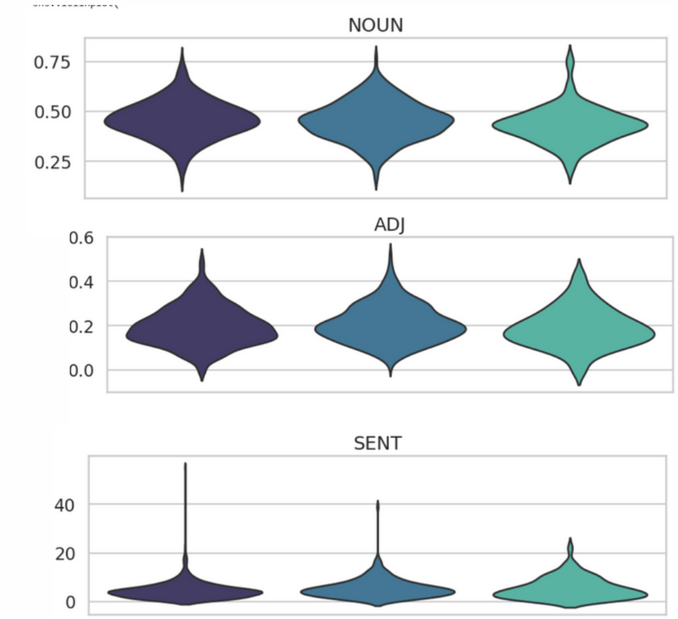
ortalama



dusuk

VİOLİN PLOT

İsimlerin, sıfatların ve fiillerin frekansı, üç inceleme türü arasında dağılımına bakıldı. Kelime sayısı, karakter sayısı ve cümlelerin ortalama uzunluğu 'dusuk' incelemelerde daha yüksek olma eğilimindedir.



03

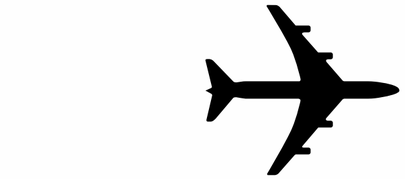
ÖNISLEME VE VEKTORLEME

Önişleme

- Noktalama işaretlerini kaldırma
- Kelimeleri küçük harfe dönüştürme
- Stop kelimelerini kaldırma
- Sözdizim etiketleri atama
- Kelimeleri köklerine ayırma (lemmatizasyon)
- Kalan kelimeleri tokenize etme

Vektörleştirici

- TF-IDF Vektörleştirici



04

MODELLEME

name	accuracy score	f1 score
Naive Bayes	0.712963	0.645777
Logistic Regression	0.597222	0.621896
Logistic Regression (PCA)	0.597222	0.623525
Decision Tree	0.662037	0.593373
Decision Tree (PCA)	0.680556	0.568102
Random Forest	0.675926	0.545222
Light GBM	0.726852	0.703356
KNN	0.694444	0.586114

Light GBM modeli, en yüksek doğruluk ve F1 skoruna sahiptir.

- Tüm modeller, optimal hiperparametreleri belirlemek için grid-search ile tarandı.
- Odak noktası accuracy score du.
- En iyi model: Light GBM
  - accuracy score: 0.726852
  - F1 score: 0.703356

En İyi 3 Model belirlendi:
Navia Bayes, Light GBM ve KNN