KAHRAMANMARAŞ SÜTÇÜ İMAM ÜNİVERSİTESİ
MÜHENDİSLİK VE MİMARLIK FAKÜLTESİ
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ
DOĞAL DİL İŞLEME DERSİ

ÜRÜN YORUMLARI DUYGU ANALİZİ

ALPEREN USLU 20110131048

ÖMER PARLAKYİĞİT 22110131307

Doç.Dr. ZEYNEP BANU ÖZGER

Mayıs 2024

1. İş Problemi

Amazon üzerinden satışlarını gerçekleştiren kulaklık ürünlerine gelen yorumları analiz ederek ve aldığı şikayetlere göre özelliklerini geliştirerek satışlarını artırmayı hedeflemektedir. Bu hedef doğrultusunda yorumlara duygu analizi ile etiketleme yapılacak ve etiketlenen veri ile sınıflandırma modeli oluşturulacaktır.

2. Veri Seti Hikayesi

Veri seti belirli bir ürün grubuna ait yapılan yorumları, yorum başlığını, yıldız sayısını, yorum tarihini ve yapılan yorumu kaç kişinin faydalı bulduğunu belirten değişkenlerden oluşmaktadır.

Bu veri seti, Amazon'daki kulaklık incelemeleri veri kümesinin küçük bir alt kümesidir. 1500 tane veri içeren İngilizce bir veri setidir.
Bu veri kümesinde 6 sütun vardır:

- Müşteri Adı : ürünü satın alan müşterinin adı
- Yorum başlığı : kısaca inceleme
- Renk : ürünün rengi
- Yorum_tarihi : müşterinin derecelendirme verdiği tarih, örneğin: 05-Eylül-21
- Yorumlar : müşteriler ürün hakkında müşterinin ne hissettiğini yorumluyor
- Değerlendirmeler : müşteri 5 yıldız üzerinden nasıl derecelendirilir, örneğin: 4/5

3. Metin Ön İşleme

Adım 1: headphone_datn.csv verisini okutma.

Adım 2: COMMENTS değişkeni üzerinde ;

a. Tüm harfleri küçük harfe çevirdik.

b. Noktalama İşaretlerini çıkardık.

c. Yorumlarda bulunan sayısal ifadeleri çıkardık.

d. Bilgi içermeyen kelimeleri (stopwords) veriden çıkardık.

e. 2'den az geçen kelimeleri veriden çıkardık.

f. Lemmatization işlemini uyguladık.

4.Metin Görselleştirme

Adım 1: Barplot görselleştirme işlemi için;

a. "COMMENTS" değişkeninin içerdiği kelimelerin frekanslarını hesaplayıp, tf olarak kaydettik.

b. Tf dataframe'inin sütunlarını yeniden adlandırdık: "words", "tf" şeklinde

c. "tf" değişkeninin değeri 100'den çok olanlara göre filtreleme işlemi yaparak barplot ile görselleştirme işlemini yaptık.

Adım 2: WordCIoud görselleştirme işlemi için;

a. "COMMENTS" değişkeninin içerdiği tüm kelimeleri "text" isminde string olarak kaydettik.

b. WordCloud kullanarak şablon şeklini belirleyip kaydettik.

c. Kaydettiğiniz wordcloud'u ilk adımda oluşturduğunuz string İle generate ettik.

d. Görselleştirme adımlarını tamamladık ve wordcloud görseli oluştu.

5. Duygu Analizi

Adım 1:
Python içerisindeki NLTK paketinde tanımlanmış olan SentimentIntensityAnalyzer nesnesini oluşturduk.

Adım 2:
SentimentIntensityAnalyzer nesnesi ile polarite puanlarının incelenmesi;

a. "COMMENTS" değişkeni için polarity_scores() hesapladık.
b. İncelenen verileri compound skoruna göre filtreleyerek tekrar gözlemledik.
c. Gözlemler için compound skorları 0 dan büyükse "pos" değilse "neg" şeklinde güncelledik.
d. Bu gözlemler sonucu pos-neg ataması yaparak yeni bir değişken olarak dataframe'e ekledik.
e. Yeni gözlemleri LabelEncoder() ile 0 ve 1 çevirdik.

SentimentIntensityAnalyzer ile yorumları etiketleyerek, yorum sınıflandırma makine öğrenmesi modeli için bağımlı değişken oluşturulmuş oldu.

6. Makine Öğrenmesine Hazırlık

Adım 1: Bağımlı ve bağımsız değişkenlerimizi belirleyerek datayı train test olarak ayırdık.

Adım 2: Makine öğrenmesi modeline verileri verebilmemiz için temsil şekillerini sayısala çevirmemiz gerekmekte;

a. CountVectorizer ve TfidfVectorizer kullanarak bir nesne oluşturduk.
b. CountVectorizer ve TfidfVectorizer için words lere göre ve ngramlara göre ayrı ayrı oluşturduk. 4 tane ayrı nesnemiz oldu.
c. Daha önce ayırmış olduğumuz train datamızı kullanarak oluşturduğumuz nesnelerle fit ettik.
d. Oluşturmuş olduğumuz vektörleri train ve test datalarına transform işlemini uygulayıp kaydettik.

7. Modelleme(Lojistik Regresyon, Random Forest)

Adım 1: Lojistik regresyon modelini kurarak train dataları ile fit ettik.
Adım 2: Kurmuş olduğunuz model ile tahmin işlemleri gerçekleştirdik;

a. Predict fonksiyonu ile test datasını tahmin ederek kaydettik.
b. classification_report ile tahmin sonuçlarınızı raporlayıp gözlemledik.
c. cross validation fonksiyonunu kullanarak ortalama accuracy değerini hesapladık.

Adım 3: Random Forest modeli ile tahmin sonuçlarının gözlenmesi;

a. RandomForestClassifier modelini kurup fit ettik.
b. Cross validation fonksiyonunu kullanarak ortalama accuracy değerini hesapladık.

c. Lojistik regresyon modeli ile sonuçları karşılaştırdık.
d. Karşılaştırma sonucu CountVectorizer words ile oluşturulan model en iyi sonucu accuracy 0.867 ile vermiştir .

8. Modele Yorum Sorulması

Adım 1: Veride bulunan yorumlardan rastgele seçerek modele sorulması;

a. Yeni yorumlar yazarak yeni bir değere atadık.
b. sample fonksiyonu ile "COMMENTS" değişkeni içerisinden örneklem seçilerek yeni bir değere atadık.
c. Kurmuş olduğunuz modele örneklemi vererek tahmin sonucunu kaydettik.
d. Örneklemi ve tahmin sonucunu ekrana yazdırdık.

KAYNAKLAR

1. https://www.kaggle.com/datasets/mdwaquarazam/headphone-dataset-review-analysis


T.C.
KAHRAMANMARAŞ
SÜTÇÜ İMAM ÜNİVERSİTESİ
1992