EMOTION DETECTION in ENGLİSH TEXT

Bu dosya, konusu Emotion detection in English text (İngilizce metinlerde duygu tespiti) olan 5 farklı makine öğrenmesi algoritmalarının raporlarını içermektedir.

Projede Görevlisi

İsmail Bayhan Yaltırık

KONYA TEKNİK ÜNİVERSİTESİ YAPAY ZEKA VE MAKİNE ÖĞRENMESİ MÜHENDİSLİĞİ

Kullanılan Veri Setinin Linki:
https://www.kaggle.com/datasets/ishantjuyal/emotions-in-text

Konu: Karar Ağacı ile Metin Sınıflandırması

Amaç: Metinlerde geçen ifadelerin hangi duyguya ait olduğunu sınıflandırmak ve bu sınıflandırmanın doğruluk oranı ile performans metriklerini değerlendirmektir.

Veri Kümesi ve Detayları: 21405 veri ve 2 kolonu (text ve emotion) bulunan emotion_final.csv veri dosyası kullanılmıştır. Maksimum 5000 kelimeye TF-IDF ile ağırlık atanmıştır.

Kullanılan Modeller: Veriler test ve eğitim setlerine ayrılır. TfidfVectorizer yöntemi ile metinler ağırlıklandırılarak sayısal verilere dönüştürülür ve veriler gereksiz kelimelerden arındırılır. Karar ağacı metinleri belirli sınıfa atayan ayrım noktaları üzerinden sınıflandırma yapar. Karışıklık matrisi, accuracy, f-measure, recall, precision, sensivity, specifity değerleri bulunmaya çalışılır.

Elde Edilen Sonuçlar:

1- Confusion Matrix(Karışıklık Matrisi): Karışıklık matrisi modelin tahmin ettiği sınıfların doğru ve yanlış dağılımını göstermektedir.

```
[[ 527  23   29    1   27     3]
 [  28  418   18    1   19    16]
 [  30   15 1223   62   59   14]
 [   9    1   55  256    2     0 ]
 [  71   29   71    9  1086   11]
 [   7   18   10    0    4   140 ]]
```

2- Doğruluk(Accuary): Modelin doğruluk oranı yaklaşık %0.85 çıkmıştır.

3-Sınıflandırma Raporu Sonucu:

	precision	recall	f1-score	support
anger	0.78	0.86	0.82	610
fear	0.83	0.84	0.83	500
happy	0.87	0.87	0.87	1403
love	0.78	0.79	0.79	323
sadness	0.91	0.85	0.88	1277
surprise	0.76	0.78	0.77	179

4- Sınıf Bazlı Performans: Her sınıf için hesaplanan metrikler:

Class 0 - Sensitivity (Recall): 0.8639
Class 0 - Specificity: 0.9606
Class 0 - F1-Score: 0.8222

Class 1 - Sensitivity (Recall): 0.8360
Class 1 - Specificity: 0.9773
Class 1 - F1-Score: 0.8327

Class 2 - Sensitivity (Recall): 0.8717
Class 2 - Specificity: 0.9367
Class 2 - F1-Score: 0.8708

Class 3 - Sensitivity (Recall): 0.7926
Class 3 - Specificity: 0.9816
Class 3 - F1-Score: 0.7853

Class 4 - Sensitivity (Recall): 0.8504
Class 4 - Specificity: 0.9632
Class 4 - F1-Score: 0.8779

Class 5 - Sensitivity (Recall): 0.7821
Class 5 - Specificity: 0.9893
Class 5 - F1-Score: 0.7713

Konu: Metin Sınıflandırmasında Lojistik Regresyon Uygulaması

Amaç: Lojistik regresyon ile metin verilerinin işlenmesi ve duygu sınıflarının tahmin edilmesidir. Performans metriklerin değerleri bulunmaya çalışılır.

Veri Kümesi ve Detayları: 21405 veri ve 2 kolonu (text ve emotion) bulunan emotion_final.csv veri dosyası kullanılmıştır. Veri setleri %80 eğitim ve %20 test oranıyla ayrılmıştır. Maksimum 5000 kelimeye TF-IDF ile ağırlık atanmıştır.

Kullanılan Modeller: TF-IDF ile metinler vektörleştirilir ve anlamlı özellikler çıkarılır. Ayrıca bununla veri seti gereksiz kelimelerden temizlenir. Model eğitim veri setiyle eğitilir ve test veri setiyle değerlendirilir. Karışıklık matrisi, accuracy, f-measure, recall, precision, sensivity, specifity değerleri bulunmaya çalışılır.

Elde edilen Sonuçlar:
1- Doğruluk (Accuary): Modelin doğruluk oranı yaklaşık % 0.84 çıkmıştır.

2-Karışıklık Matrisi: Karışıklık matrisi modelin tahmin ettiği sınıfların doğru ve yanlış dağılımını göstermektedir.

```
[[ 439   13   40    3   56    0]
 [  28  395   52    0   60    8]
 [   8    2 1308   20   49    1]
 [   4    0  130  194   30    0]
 [  27    7   74    1 1185    1]
 [   5   18   29    0   26   79]]
```

3- Recall (Duyarlılık) ve Specificity (Özgüllük) Parametre Değerleri:

Sınıf: anger
Duyarlılık (Recall): 0.7967
Özgüllük (Specificity): 0.9808

Sınıf: sadnees
Duyarlılık (Recall): 0.9151
Özgüllük (Specificity): 0.9151

Sınıf: fear
Duyarlılık (Recall): 0.7274
Özgüllük (Specificity): 0.9893

Sınıf: surprise
Duyarlılık(Recall): 0.5032
Özgüllük (Specificity): 0.9976

Sınıf: happy
Duyarlılık (Recall): 0.9424
Özgüllük (Specificity): 0.8881

Sınıf: love
Duyarlılık (Recall): 0.5419
Özgüllük (Specificity): 0.9939

Konu: KNN Algoritması ile Metin Duygu Analizi

Amaç: KNN algoritmasını kullanarak metinlerin hangi duygusal kategoriye (örneğin mutlu, üzgün, öfkeli gibi) ait olduğunu belirlemek.

Veri Kümesi ve Detayları: 21405 veri ve 2 kolonu (text ve emotion) bulunan emotion_final.csv veri dosyası kullanılmıştır.

Kullanılan Modeller: Veriler işlenir. Veriler eğitim ve test kümelerine bölünür. TfidfVectorizer ile kullanılarak sayısal verilere dönüştürülür. K-Nearest Neighbors (KNN) sınıflandırıcısı oluşturulur. Eğitim verileriyle model eğitilir ve test verileriyle tahminler yapılır. Karışıklık matrisi, accuracy, f-measure, recall, precision, sensivity, specifity değerleri bulunmaya çalışılır.

Elde Edilen Sonuçlar:

1- Karışıklık Matrisi Sonuçları:

```
[[22  0  5  0 10  0]
 [ 4 16  3  0  4  0 ]
 [ 7  3 48  1  5  0 ]
 [ 4  0  6  9  1  0 ]
 [ 6  2  6  1 45  0 ]
 [ 1  2  1  0  1  2 ]]
```

2- Doğruluk(Accuarcy): Doğruluk değeri yaklaşık olarak %0.66 çıkmıştır.

3- Sınıflandırma Raporu Sonuçları:

	precision	recall	f1-score	support
anger	0.50	0.59	0.54	37
fear	0.70	0.59	0.64	27
happy	0.70	0.75	0.72	64
love	0.82	0.45	0.58	20
sadness	0.68	0.75	0.71	60
surprise	1.00	0.29	0.44	7

4- Recall (Duyarlılık) ve Specificity (Özgüllük): Duyarlılık değeri yaklaşık %0.80 ve özgüllük değeri 1.0 çıkmıştır.

*Doğruluk değerini kod üzerinde her ne kadar oynama yapsakta yine de istediğimiz yüksek değerlerde göremedik.

Konu: Metin Duygu Analizi için K-means ile Kümeleme

Amaç: Projenin amacı metin verilerini sayısallaştırmak ve K-means algoritması ile metinlerin duygusal anlamda hangi kümelerde olabileceğini tahmin etmektir.

Veri Kümesi ve Detayları: 21405 veri ve 2 kolonu (text ve emotion) bulunan emotion_final.csv veri dosyası kullanılmıştır.

Kullanılan Modeller: Etiketsiz verilerde benzerliklere dayanarak kümeler oluşturmak için K-means algoritması kullanılmıştır. Metin verilerini sayısallaştırmak ve gereksiz sözcükleri çıkartmak için TfidfVectorizer yöntemi kullanılmıştır. Verileri 2 boyuta indirgemek ve bu verinin 2D düzlemde görselleştirmek için PCA (Principal Component Analysis) kullanılmıştır. Karışıklık matrisi, accuracy, f-measure, recall, precision, sensivity, specifity değerleri bulunmaya çalışılır.

Elde Edilen Sonuçlar:
1-) K-means algoritması veriyi toplamda 5 kümeye ayırmıştır.

2- PCA ile indirgenen veriler görselleştirilmiştir.

3- Karışıklık Matrisi Sonuçları:

```
[[2657    0    0    0    0  3608]
 [1357    0    0    0    0  1636]
 [ 548    0    0    0    0  1093]
 [ 454    0    0    0    0   425]
 [1297    0    0    0    0  1355]
 [2439    0    0    0    0  4590]]
```

4- Sınıflandırma Sonucu Raporu:

	precision	recall	f1-score	support
sadness	0.30	0.42	0.35	6265
anger	0.00	0.00	0.00	2993
love	0.00	0.00	0.00	1641
surprise	0.00	0.00	0.00	879
fear	0.00	0.00	0.00	2652
happy	0.36	0.65	0.47	7029

5- Doğruluk(Accuarcy): Doğruluk değeri yaklaşık %0.34 çıkmıştır.

*Performans ve doğruluk değerleri istediğimiz gibi çıkmamış, düşük kalmıştır. Sebebini araştırdığımızda K-Means algoritmasının veri kümesi ile uyumsuzluğu veya veri kümesi kaynaklı olduğu görülmüştür.

Konu: Random Forest ile Metinlerde Duygu Sınıflandırması

Amaç: Metinlerdeki ifadeleri analiz ederek bu ifadelerin hangi duygu kategorisinde (örneğin mutluluk, üzüntü gibi) ait olduğunu bulmak.

Veri Kümesinin Detayları: 21405 veri ve 2 kolonu (text ve emotion) bulunan emotion_final.csv veri dosyası kullanılmıştır. Bu veriler daha sonralarda işleme tabii tutulmuştur.

Kullanılan Modeller: TfidfVectorizer yöntemi ile metinlerdeki kelimelerin önem düzeyine göre ağırlıklar atayarak veri vektörleştirilmiş ve gereksiz kelimelerden temizlenmiştir. Veriler eğitim ve teste tabii tutulur. Random Forest Algoritması ile birden fazla karar ağacı birleştirilmeye çalışılır bununla sınıflandırma doğruluğunu arttırmayı hedefleriz. Daha sonra model test edilir, sonuçlar değerlendirilir ve özelliklerin önemi görselleştirilir.

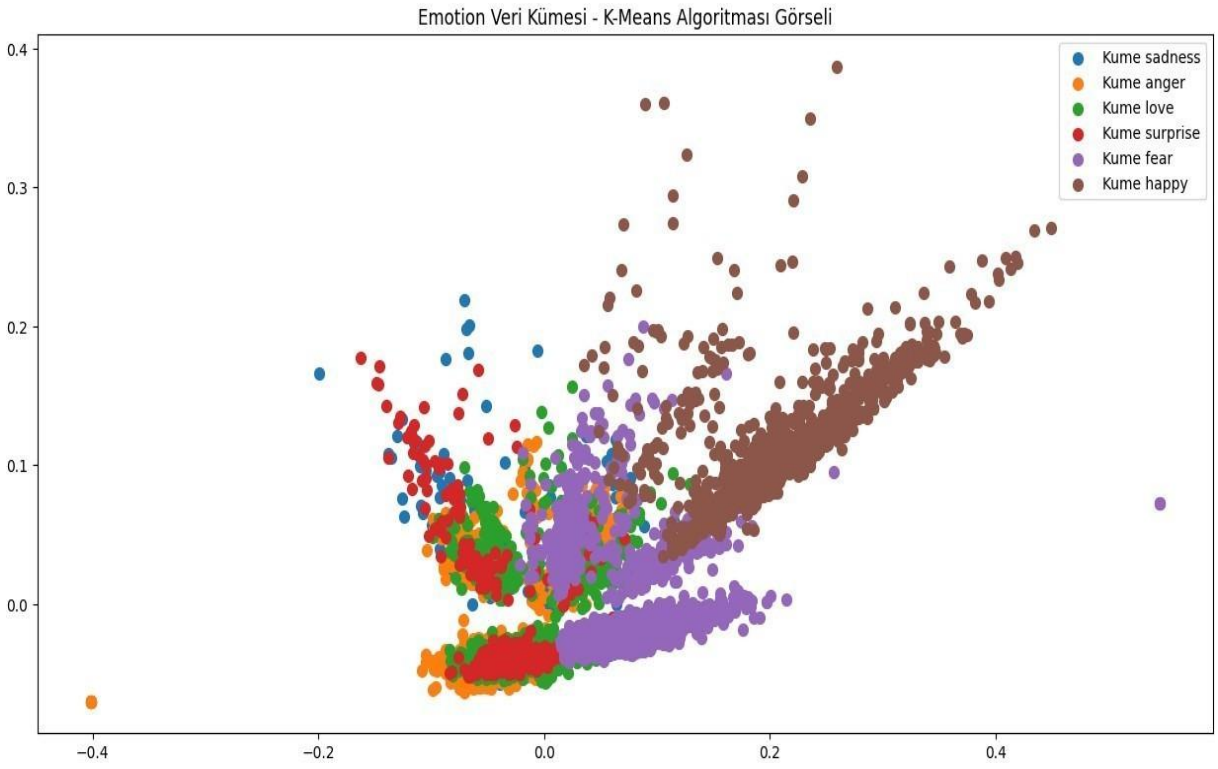
Elde Edilen Sonuçlar:
1- Doğruluk (Accuracy):Modelin doğruluk oranı yaklaşık olarak %0.88 çıkmıştır.

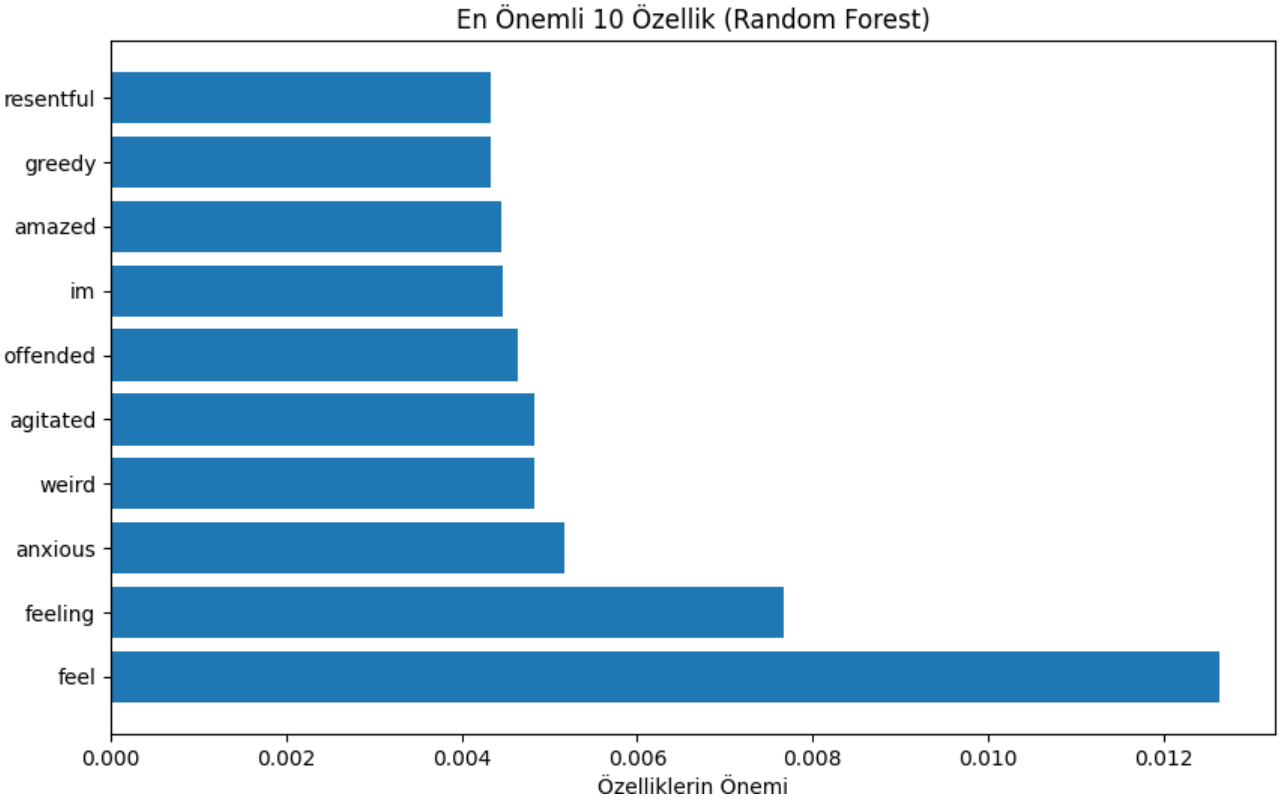
2- Sınıflandırma Raporu:

	precision	recall	f1-score	support
anger	0.91	0.85	0.88	909
fear	0.85	0.87	0.86	796
happy	0.86	0.94	0.90	2059
love	0.86	0.72	0.78	492
sadness	0.92	0.90	0.91	1929
surprise	0.83	0.70	0.76	253

3- Recall (Duyarlılık) ve Specificity (Özgüllük): Duyarlılık değeri yaklaşık olarak %0.97 özgüllük değeri yaklaşık olarak %0.96 çıkmıştır

4- Görselleştirme: En önemli 10 kelimeye ait değerler çubuk grafiği ile görselleştirilmiş.


Emotion Veri Kümesi - K-Means Algoritması Görseli
Kume sadness
Kume anger
Kume love
Kume surprise
Kume fear
Kume happy
0.4
0.3
0.2
0.1
0.0
−0.4
−0.2
0.0
0.2
0.4


En Önemli 10 Özellik (Random Forest)
resentful
greedy
amazed
im
offended
agitated
weird
anxious
feeling
feel
0.000
0.002
0.004
0.006
0.008
0.010
0.012
Özelliklerin Önemi