Samir Ganbarli Kolektif Öğrenme - Ödev 2 Rapor

Temsillerin / Kararların Birleştirilmesi

Ödev kapsamında iki metin veri kümesi 5 farklı temsil yöntemiyle sayısal olarak ifade edilip 3 farklı makine öğrenmesi algoritması (SVM, RF, MLP) ile eğitilmiştir. Birinci veri seti toplamda 10588, ikinci ise 10000 satırdan oluşmaktadır. Her iki veriseti eğitim ve test için 0.2 oranında bölünmüştür (20% test, 80% eğitim).

1) İlk veriseti iki sınıftan (pozitif, negatif) oluşan duygu analizi modeli için kullanılan cümlelerden (ilk sütun cümle, ikinci sütun ise duygu sınıfı) oluşmaktadır. İkinci veriseti ise metin sınıflandırma için kullanılan ve 32 farklı sınıftan oluşan haber metinlerinden (ilk sütun sınıf ismi ikinci sütun haber metni) oluşmaktadır.

Ödevde kullanılan temsil yöntemleri : all-MiniLM-L12-v2, jina-embeddings-v3, multilingual-e5-large-instruct, nomic-embed-text-v1, gte-large. Kullanılan algoritmaların 3-ü de çoğu zaman sınıflandırma görevleri için tercih edilen algoritmalardır. Raporda temsil yöntemleri aynı zamanda model olarak da isimlendirilmiştir.

Model1 -> all-MiniLM-L12-v2: dimension = 384

Model2 -> jina-embeddings-v3: dimension 1024

Model3 -> multilingual-e5-large-instruct: dimension 1024

Model4 -> nomic-embed-text-v1: dimension 1024

Model5 -> gte-large: dimension 1024

İlk veri seti için yapılan çalışmada öncelikle her temsil yöntemi her algoritma için accuracy ve F1 puanları alınmıştır. Bundan sonra birleştirme (ensemble) tekniği kullanarak temsil modelleri ve algoritmalar için iyileştirmeler tespit edilmiştir. (Bütün sonuçların (15) birleşdiği total ensemble puanı da tespit edilmiştir). Her sonuç için classification report kod dosyasındadır.

SVM	Model1	Acc-0.89, F1-0.89
RF	Model1	Acc-0.84, F1-0.84
MLP	Model1	Acc-0.88, F1-0.88
SVM	Model2	Acc-0.96, F1-0.96
RF	Model2	Acc-0.95, F1-0.95
MLP	Model2	Acc-0.96, F1-0.96
SVM	Model3	Acc-0.96, F1-0.96
RF	Model3	Acc-0.95, F1-0.95
MLP	Model3	Acc-0.96, F1-0.96
SVM	Model4	Acc-0.90, F1-0.90
RF	Model4	Acc-0.86, F1-0.86
MLP	Model4	Acc-0.89, F1-0.89
SVM	Model5	Acc-0.88, F1-0.88
RF	Model5	Acc-0.84, F1-0.84

MLP	Model5	Acc-0.87, F1-0.87

Ensemble Accuracy SVM: 0.9367 (5 temsil yöntemi ile)

Ensemble Accuracy RF: 0.9202 (5 temsil yöntemi ile)

Ensemble Accuracy MLP: 0.9433 (5 temsil yöntemi ile)

Ensemble Accuracy M1: 0.8881 (3 algoritma ile)

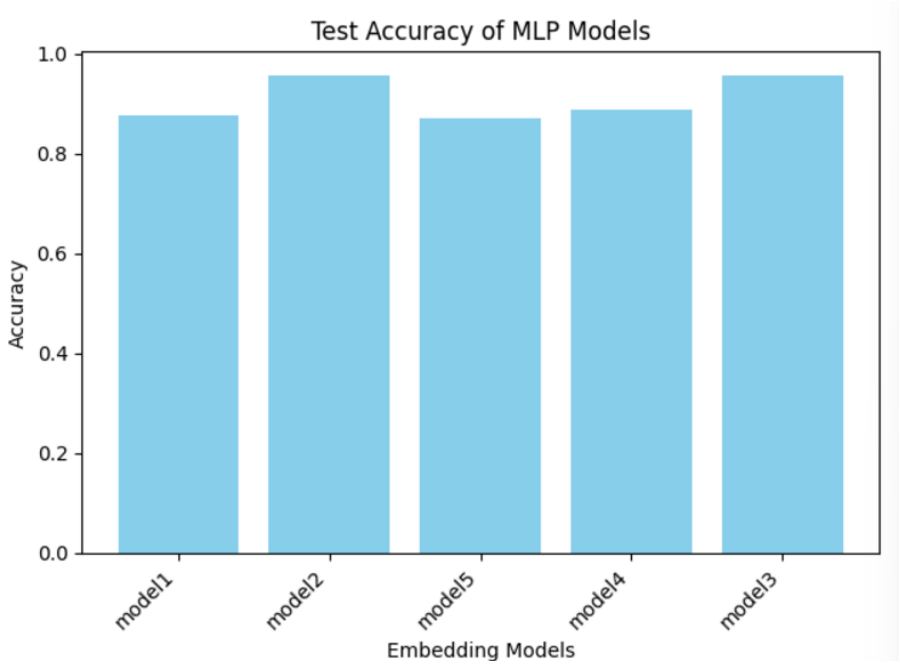
Ensemble Accuracy M2: 0.9622 (3 algoritma ile)

Ensemble Accuracy M3: 0.9618 (3 algoritma ile)

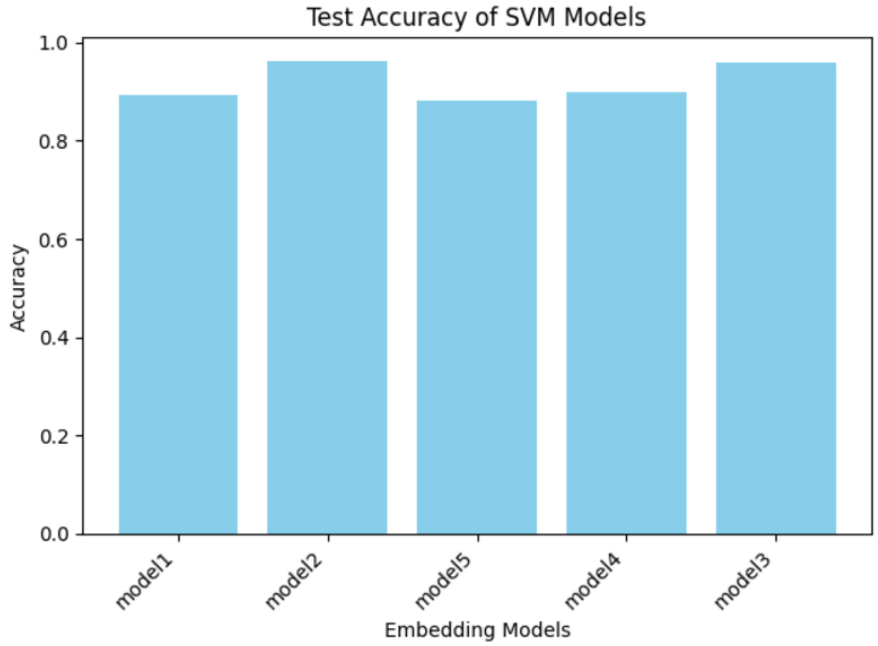
Ensemble Accuracy M4: 0.8966 (3 algoritma ile)

Ensemble Accuracy M5: 0.8829 (3 algoritma ile)

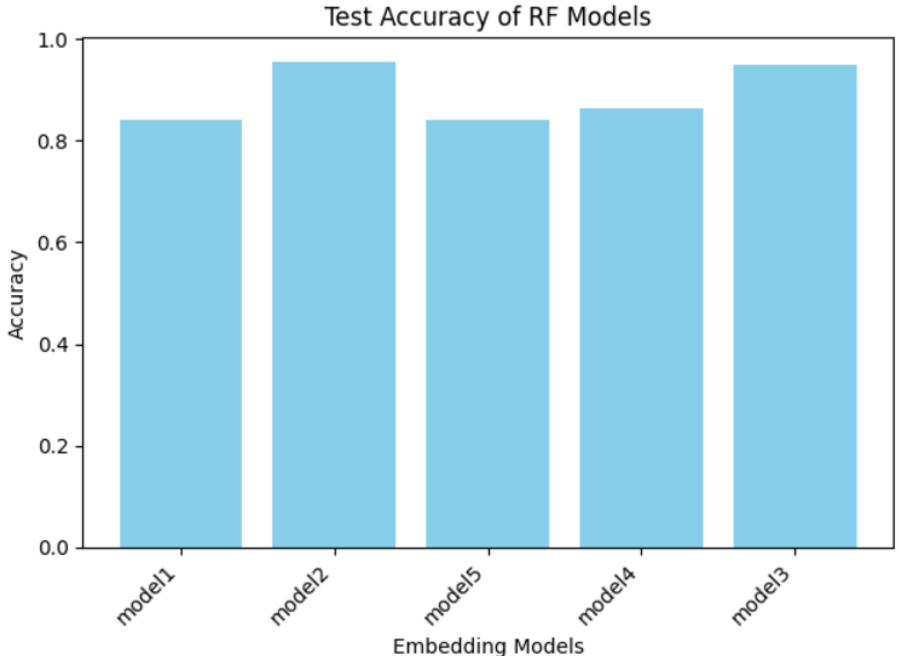
Total Ensemle Accuracy: 0.94



Şekil 1. Temsil yöntemlerinin MLP modeli ile ensemble grafiği



Şekil 2. SVM modelinin temsil yöntemleri ile acc grafiği



Şekil 3. RF modelinin temsil yöntemleri ile acc grafiği

Bireysel olarak yüksek olan F1 ve accuracy değerleri ikinci ve 3 cü temsil yöntemleri (jina-embeddings, multilingual-e5-large-instruct) ile elde edilmiştir.

Ensemble yönteminde ise temsil yöntemlerinin birleşmesinde en yüksek başarı MLP için alınmıştır. Algoritmaların birleşmesinde ise ikinci ve üçüncü model için yüksek başarı alınmıştır.

Tüm sonuçların birleşmesinden alınan accuracy 0.94 olmuştur.

2) İkinci veri seti için de aynı temsil yöntemleri ile metin verileri vektörlere çevrilmiştir.

Her sonuç için detaylı classification report kod dosyasındadır.

SVM	Model1	Acc- 0.68 , F1- 0.68
RF	Model1	Acc- 0.52, F1- 0.52
MLP	Model1	Acc- 0.64, F1- 0.64
SVM	Model2	Acc- 0.88, F1- 0.88
RF	Model2	Acc- 0.82, F1- 0.82
MLP	Model2	Acc- 0.85, F1- 0.85
SVM	Model3	Acc- 0.90, F1- 0.90
RF	Model3	Acc- 0.80, F1- 0.80
MLP	Model3	Acc- 0.86, F1- 0.86
SVM	Model4	Acc- 0.76, F1- 0.76
RF	Model4	Acc- 0.54, F1- 0.54
MLP	Model4	Acc- 0.72, F1- 0.72
SVM	Model5	Acc- 0.74, F1- 0.74
RF	Model5	Acc- 0.59, F1- 0.59
MLP	Model5	Acc- 0.72, F1- 0.72

Bireysel sonuçlarda en iyi puanlar yine aynı şekilde ikinci ve beşinci temsil yöntemi modellerinde elde edilmiştir (jina-embeddings, multilingual-e5-large-instruct).

Ensemble Accuracy SVM: 0.8663 (5 temsil yöntemi modeli ile)

Ensemble Accuracy RF: 0.7997 (5 temsil yöntemi modeli ile)

Ensemble Accuracy MLP: 0.8858 (5 temsil yöntemi modeli ile)

Ensemble Accuracy M1: 0.6640 (3 algoritma ile)

Ensemble Accuracy M2: 0.8728 (3 algoritma ile)

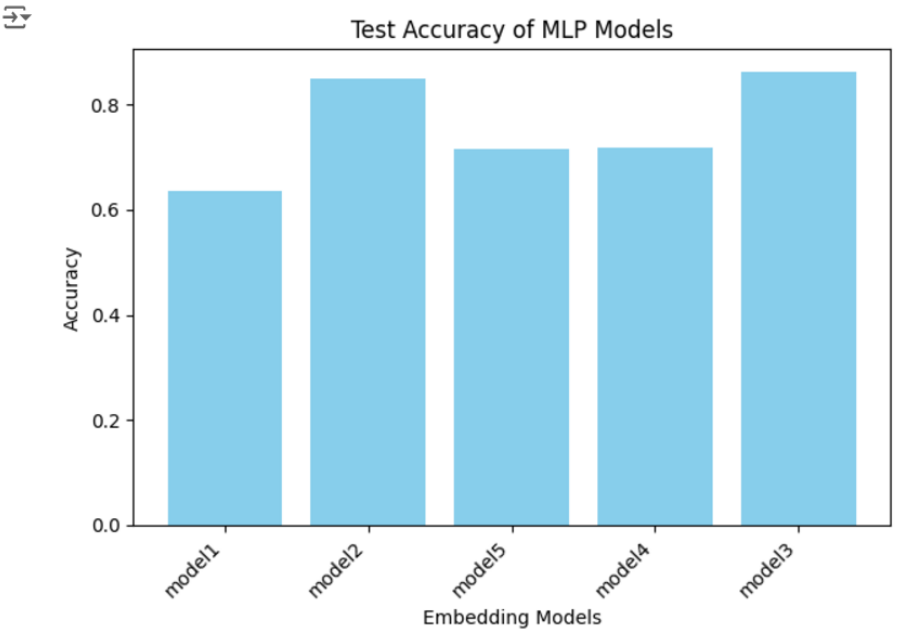
Ensemble Accuracy M3: 0.8908 (3 algoritma ile)

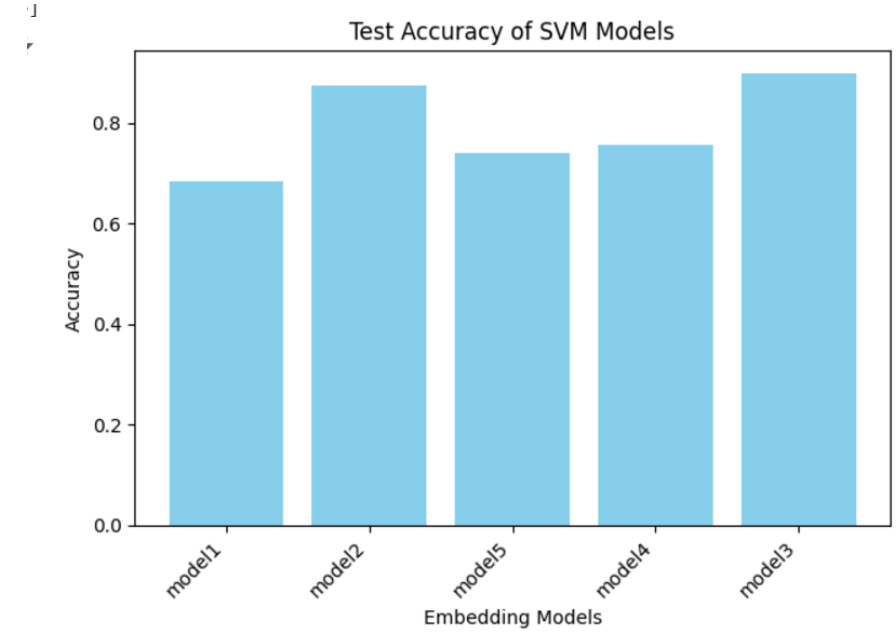
Ensemble Accuracy M4: 0.7341 (3 algoritma ile)

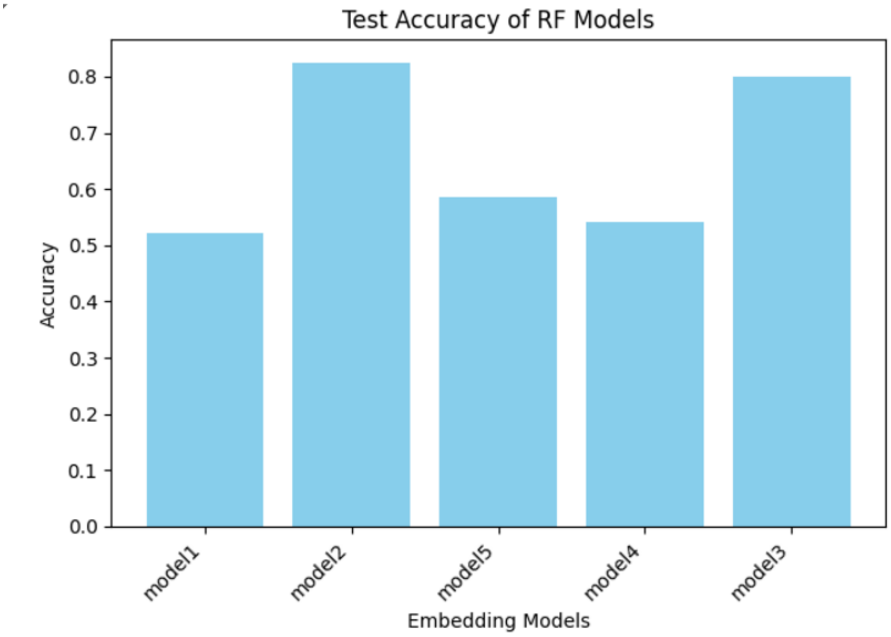
Ensemble Accuracy M5: 0.7241 (3 algoritma ile)

Total Ensemble Accuracy: 0.89

Ensemble yönteminde ise 5 temsil yönteminin ensemblinde en iyi sonuç yine MLP-den alınmıştır. 3 algoritmanın ensemblinde ise en iyi puanlar ikinci ve üçüncü temsil yöntemi modellerinde alınmıştır.


Test Accuracy of MLP Models
Accuracy
0.0
0.2
0.4
0.6
0.8
model1
model2
model5
model4
model3
Embedding Models


Test Accuracy of SVM Models
Accuracy
0.0
0.2
0.4
0.6
0.8
model1
model2
model5
model4
model3
Embedding Models


Test Accuracy of RF Models
Accuracy
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
model1
model2
model5
model4
model3
Embedding Models