YILDIZ TEKNİK ÜNİVERSİTESİ
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ
KOLEKTİF ÖĞRENME DERSİ
ÖDEVİ
2024-2025 GÜZ

Temsillerin / Kararların Birleştirilmesi

DERSİ VEREN Öğretim Üyesi : **Prof. Dr. Mehmet Fatih AMASYALI**
ÖDEV VERİLİŞ TARİHİ : **19.11.2024**
ÖDEV TESLİM TARİHİ : **03.12.2024**

Öğrencinin Adı - Soyadı	Öğrencinin Numarası
Zeynep ASLAN	23501135

İçindekiler

1. Amaç

Bu çalışmada, farklı metin temsil yöntemleriyle (embeddings) farklı makine öğrenimi algoritmalarını kullanarak performans karşılaştırması yapmak ve bu yöntemlerin çıktılarının birleştirilerek (ensemble) sonuçlara etkisini incelemek amaçlanmıştır. Temsil yöntemleri, özellikle Türkçe metin veri kümelerinde metin sınıflandırma performansını değerlendirmek için seçilmiştir.

2. Veri Setleri

2.1. Türkçe Sosyal Medya Paylaşımı Veri Seti

Türkçe tweetlerden oluşan bir sosyal medya paylaşımı veri setidir. 11.000'den fazla tweet barındırmaktadır. %45 i negatif, %55 i pozitiftir.

2.2. Duygu Analizi Veri Seti

Çeşitli elektronik mağazalardan toplanan Türkçe duygu analizi veri setidir. 11.000'den fazla veri barındırmaktadır. 4253 olumlu, 4238 olumsuz ve 2938 tarafsız sınıfından veri içermektedir.

3. Veri Ön İşleme

Veri setlerine yapılan ön işlemeler aşağıdaki gibidir.

1. **Veri Filtreleme:**
   - Anlam ifade etmeyen veya çok kısa metinler çıkarıldı.
   - Belirli bir sınıf dağılımı eşitlenerek dengesiz veri seti normalize edildi.
2. **Sınıf Dağılımının İncelenmesi:**
   - Her bir sınıfın veri sayısı analiz edildi.
   - Veri sayısı az olan sınıflar sınıf dağılımının istatistiğine bakılarak veri setinden çıkarıldı.
3. **Satır Başına Kelime Sayısı Analizi:**
   - Metinlerin, kelime sayıları çıkarıldı.
   - Maksimum metin uzunluğu belirlendi ve daha uzun metinler kırpıldı.
4. **Metin Temizliği:**
   - Gereksiz boşluklar, özel karakterler, sayılar ve noktalama işaretleri temizlendi, tüm metinler küçük harfe dönüştürüldü.
   - Stopwords ler çıkarıldı.
5. **Metin Temizliği:**
   - Veri kümeleri %80 eğitim ve %20 test olarak ayrıldı.
   - Sınıf bazında stratify parametresi kullanıldı.

4. Embedding Yöntemleri

Çalışmada kullanılan embedding modelleri, metin verilerini anlam ve bağlama göre sayısal vektörlere dönüştüren güçlü temsil yöntemleridir. Aşağıda her bir embedding modeli ve özellikleri açıklanmıştır:

1. dbmdz/bert-base-turkish-uncased

- **Model Türü:** BERT tabanlı Türkçe dil modeli.
- **Özellikler:**
  - Küçük harf duyarlılığına sahip olmayan (uncased) bir modeldir, yani metindeki büyük-küçük harf farklarını göz ardı eder.
  - Özellikle Türkçe metinler için optimize edilmiştir ve Türkçe dil bilgisi kurallarına göre çalışır.
- **Kullanım Alanı:** Türkçe dil işleme görevlerinde, özellikle metin sınıflandırma ve duygu analizi gibi problemlerde başarılıdır.

2. sentence-transformers/all-MiniLM-L12-v2

- **Model Türü:** Küçük ve hızlı bir Transformer modeli.
- **Özellikler:**
  - Düşük kaynak tüketimiyle yüksek doğruluk sağlar.
  - Çok dilli bir modeldir ve Türkçe dahil birçok dilde iyi performans gösterir.
  - Özellikle cümle ve metinlerin anlamlarını yakalamak için tasarlanmıştır.
- **Kullanım Alanı:** Metin benzerliği, arama, sıralama gibi görevlerde yaygın olarak kullanılır.

3. intfloat/multilingual-e5-large-instruct

- **Model Türü:** Multilingual (çok dilli) büyük dil modeli.
- **Özellikler:**
  - Çoklu dil desteği ile geniş çapta kullanım sağlar.
  - Büyük boyutlu bir model olduğu için daha derin anlam ve bağlam yakalamada başarılıdır.
  - Çeşitli dil işleme görevleri için eğitim verilmiştir ve birçok farklı görevde uygulanabilir.
- **Kullanım Alanı:** Çeviri, duygu analizi, sınıflandırma gibi dil işleme görevlerinde güçlü bir performans sunar.

4. **dbmdz/electra-base-turkish-cased-discriminator**
   - **Model Türü:** ELECTRA tabanlı Türkçe dil modeli.
   - **Özellikler:**
     - Büyük-küçük harf duyarlılığına sahip bir modeldir (cased).
     - Dönüştürücülere dayalı "replaced token detection" özelliğiyle verimli bir şekilde çalışır.
     - Türkçe dil yapısını derinlemesine öğrenerek daha doğru tahminler yapar.
   - **Kullanım Alanı:** Türkçe metinlerde detaylı analiz ve sınıflandırma görevlerinde kullanılır.

5. **cardiffnlp/twitter-xlm-roberta-base**
   - **Model Türü:** XLM-RoBERTa tabanlı çok dilli model.
   - **Özellikler:**
     - Twitter gibi kısa metinlerde güçlü bir performans sergiler.
     - Türkçe dahil birçok dili destekler ve sosyal medya metinlerinde özelleşmiştir.
     - Duygu analizi, kullanıcı niyeti tahmini gibi sosyal medya metin işleme görevleri için optimize edilmiştir.
   - **Kullanım Alanı:** Kısa metin analizi, sosyal medya yorumları sınıflandırma gibi görevlerde idealdir.

5. Makine Öğrenmesi Algoritmaları

Verilen dokümanlarda kullanılan makine öğrenmesi algoritmaları aşağıda açıklanmıştır:

1. Destek Vektör Makineleri (SVM)

- **Temel Prensip:**
  - SVM, sınıflandırma problemlerinde veriyi ayırmak için en iyi sınır çizgisini (hiper düzlem) belirler.
  - Doğrusal olarak ayrılabilen ve ayrılmayan veri kümeleri için çekirdek (kernel) fonksiyonları kullanarak esnek bir şekilde çalışır.
- **Özellikler:**
  - Özellikle yüksek boyutlu verilerde etkili sonuç verir.
  - Gürültülü verilerle çalışırken kararlı performans gösterir.
- **Avantajlar:**
  - Karmaşık olmayan bir yapıya sahiptir ve aşırı öğrenmeye (overfitting) karşı dirençlidir.
- **Kullanım Alanı:**
  - Metin sınıflandırma, duygu analizi, sahtekarlık tespiti gibi görevlerde yaygın olarak kullanılır.

2. Rastgele Ormanlar (RF)

- **Temel Prensip:**
  - RF, birden fazla karar ağacından oluşan bir ensemble yöntemidir.
  - Her ağacın sonucu oylanır ve çoğunluğun kararı son tahmin olarak alınır.
- **Özellikler:**
  - Aşırı öğrenmeye karşı dayanıklıdır.
  - Karmaşık veri yapıları ve çok sayıda değişkenle çalışabilir.
- **Avantajlar:**
  - Hızlı hesaplama süresi ve kolay uygulanabilirliği ile öne çıkar.
  - Dengesiz veri kümelerinde iyi performans gösterir.
- **Kullanım Alanı:**
  - Metin sınıflandırma, özellik seçimi ve tahmin gibi görevlerde tercih edilir.

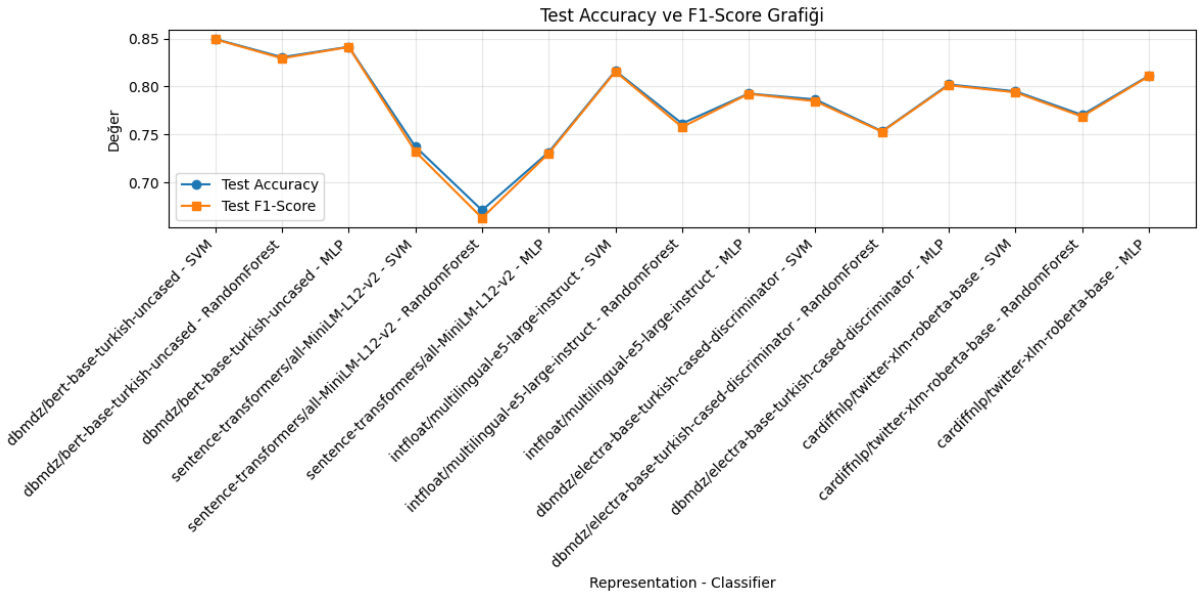
3. Çok Katmanlı Algılayıcılar (MLP)

- **Temel Prensip:**
  - MLP, yapay sinir ağları ailesinden bir modeldir.
  - Girdi ve çıktı arasında bir veya daha fazla gizli katmana sahiptir ve bu katmanlar aracılığıyla doğrusal olmayan ilişkileri öğrenir.
- **Özellikler:**
  - Geri yayılım algoritmasıyla ağırlıklarını optimize eder.
  - Daha karmaşık ve esnek modeller oluşturabilir.
- **Avantajlar:**
  - Çeşitli veri türleriyle çalışabilir ve doğrusal olmayan sınıflandırma problemlerinde etkili çözüm sunar.
- **Kullanım Alanı:**
  - Görüntü ve metin sınıflandırma, tahmin ve regresyon gibi görevlerde kullanılır.

6. Sonuçlar

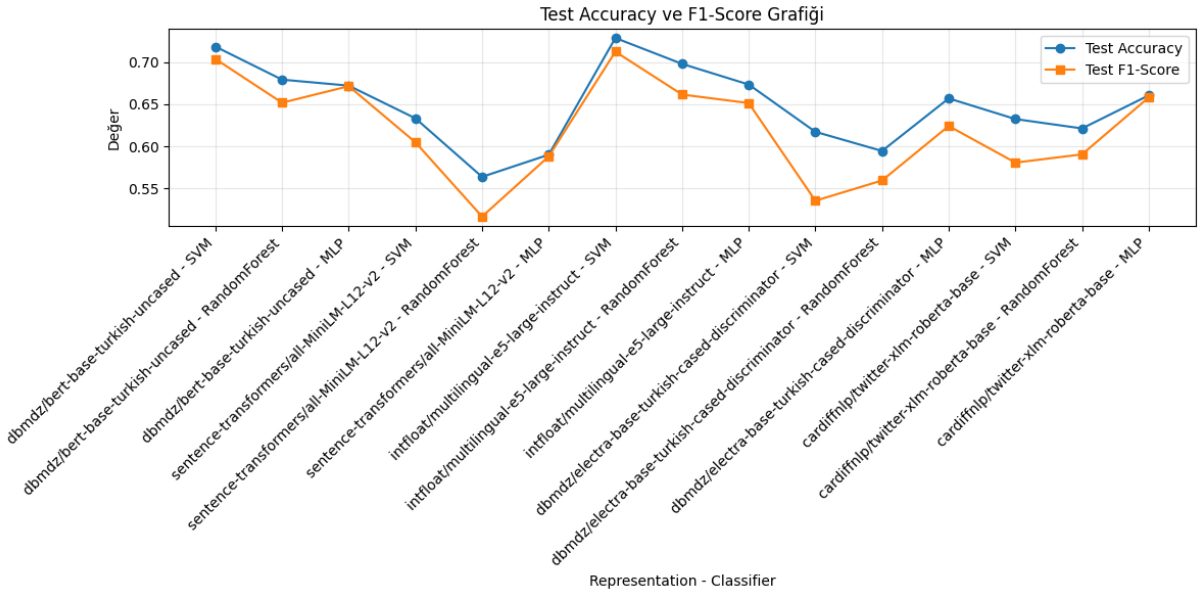
- **5 Temsil ve 3 Algoritmanın Beraber Karşılaştırılması**

Sosyal Medya Veri Seti:



1. **`dbmdz/bert-base-turkish-uncased`:**
   - Bu temsil yöntemi, tüm algoritmalar için en yüksek **accuracy** ve **F1-Score** değerlerini sağlamış.
   - En iyi performans **SVM** algoritmasıyla (Accuracy: 0.849, F1: 0.849) elde edilmiş.
2. **`sentence-transformers/all-MiniLM-L12-v2`:**
   - Performansı diğer temsil yöntemlerine göre daha düşük.
   - En iyi sonuç yine **SVM** ile elde edilmiş (Accuracy: 0.737, F1: 0.732).
   - **RandomForest** bu temsil yöntemiyle düşük sonuçlar vermiş.
3. **`intfloat/multilingual-e5-large-instruct`:**
   - Performansı orta seviyede, özellikle **SVM** ile iyi sonuçlar elde edilmiş (Accuracy: 0.816, F1: 0.815).
   - **MLP** ile de makul bir performans sergilemiş (Accuracy: 0.793, F1: 0.792).
4. **`dbmdz/electra-base-turkish-cased-discriminator`:**
   - Performans olarak dengeli, ancak diğer yöntemlere göre daha geride.
   - En iyi sonuç **MLP** ile alınmış (Accuracy: 0.802, F1: 0.801).
5. **`cardiffnlp/twitter-xlm-roberta-base`:**
   - Sosyal medya verilerine uygun bir model olmasına rağmen, genel sınıflandırma problemlerinde iyi sonuç vermiş.
   - En iyi sonuç yine **MLP** ile alınmış (Accuracy: 0.811, F1: 0.811).

Duygu Analizi Veri Seti:



Sonuçların Kısa Yorumu:

1. **`dbmdz/bert-base-turkish-uncased`**:
   - Bu temsil yöntemi, tüm algoritmalar arasında **SVM** ile en iyi sonuçları sağlamış (**Accuracy: 0.718, F1: 0.703**).
   - Ancak, önceki sonuçlara göre daha düşük bir performans sergilemiş. Özellikle **RandomForest** ve **MLP** algoritmalarında başarı sınırlı kalmış.
2. **`sentence-transformers/all-MiniLM-L12-v2`**:
   - Performansı diğer temsil yöntemlerine göre oldukça düşük.
   - En iyi sonuç **SVM** ile alınmış (**Accuracy: 0.633, F1: 0.604**).
   - **RandomForest** bu temsil yönteminde en düşük sonuçları vermiş (**Accuracy: 0.563, F1: 0.516**).
3. **`intfloat/multilingual-e5-large-instruct`**:
   - Performansı daha dengeli ve yüksek.
   - **SVM** ile en iyi sonuçları sağlamış (**Accuracy: 0.728, F1: 0.712**).
   - **MLP** ve **RandomForest** ile de makul sonuçlar alınmış ancak **SVM** kadar iyi değil.
4. **`dbmdz/electra-base-turkish-cased-discriminator`**:
   - Genel performans, diğer temsil yöntemlerine göre daha düşük.
   - **MLP** ile en iyi sonucu sağlamış (**Accuracy: 0.657, F1: 0.624**).
   - **SVM** ve **RandomForest** ile performans görece daha zayıf.
5. **`cardiffnlp/twitter-xlm-roberta-base`**:
   - Performansı sınırlı, ancak dengeli sonuçlar göstermiş.
   - **MLP**, en iyi performansı sağlamış (**Accuracy: 0.660, F1: 0.658**).

- **Ensemble Modellerin Karşılaştırılması**

Farklı makine öğrenimi modelleri ve temsil yöntemleriyle elde edilen tahminleri birleştirerek ensemble (birlikte öğrenme) yöntemi kullanılarak sonuçlar karşılaştırılmıştır. Aşağıda tahminlerin nasıl birleştirildiği açıklanmıştır.

1. Aynı Temsil Yöntemi için Ensemble

- **Amaç:** Aynı temsil yöntemi (embedding modeli) kullanılarak elde edilen farklı sınıflandırıcıların (SVM, RandomForest, MLP) tahminlerini birleştirmek.
- **Nasıl Çalışır?**
  - Her temsil yöntemi için tüm sınıflandırıcıların (SVM, RF, MLP) tahminleri toplanır.
  - Tahminler arasında çoğunluk oylaması (**majority voting**) uygulanır.
  - Çoğunlukla tahmin edilen sınıf, temsil yöntemi için ensemble sonucu olarak kaydedilir.
- **Sonuç:** Her bir temsil yöntemi için birleştirilmiş (ensemble) tahminler oluşturulur.

2. Aynı Algoritma için Ensemble

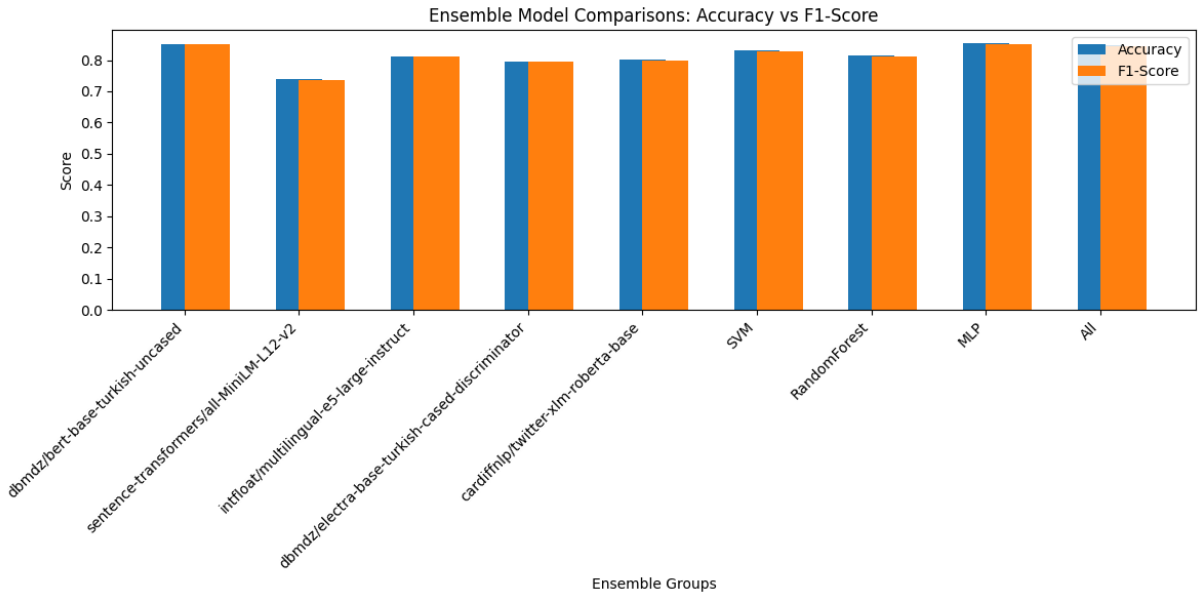
- **Amaç:** Aynı sınıflandırıcıyı (SVM, RandomForest, MLP) kullanarak farklı temsil yöntemlerinden elde edilen tahminleri birleştirmek.
- **Nasıl Çalışır?**
  - Her sınıflandırıcı için farklı temsil yöntemlerinin tahminleri toplanır.
  - Çoğunluk oylaması uygulanır.
  - Çoğunlukla tahmin edilen sınıf, ilgili sınıflandırıcı için ensemble sonucu olarak kaydedilir.
- **Sonuç:** Her bir sınıflandırıcı için temsil yöntemlerinden bağımsız birleştirilmiş (ensemble) tahminler oluşturulur.

3. Tüm Sonuçları Birleştirerek Ensemble

- **Amaç:** Tüm temsil yöntemleri ve sınıflandırıcıları birleştirerek genel bir ensemble tahmin oluşturmak.
- **Nasıl Çalışır?**
  - Tüm temsil yöntemlerinden ve sınıflandırıcılardan elde edilen tahminler tek bir listede toplanır.
  - Çoğunluk oylaması uygulanır.
  - Çoğunlukla tahmin edilen sınıf, nihai ensemble tahmini olarak belirlenir.
- **Sonuç:** Tüm yöntem ve algoritmaların birleştirilmesiyle oluşturulmuş nihai bir tahmin seti elde edilir.

Aşağıda veri setlerinde ensemble ların karşılaştırılabileceği grafikler verilmiştir.

Sosyal Medya Veri Seti:



Sonuçların Kısa Yorumu:

En Yüksek Performans:

- **dbmdz/bert-base-turkish-uncased** yöntemi, hem Accuracy hem de F1-Score açısından en yüksek performansı göstermiş. Bu temsil yöntemi, SVM sınıflandırıcısıyla (Accuracy: 0.849, F1: 0.849) diğer kombinasyonlardan açık ara önde.

Sentence-Transformers (all-MiniLM-L12-v2):

- Bu temsil yöntemi, diğerlerine kıyasla daha düşük performans göstermiş. Özellikle RandomForest sınıflandırıcısıyla elde edilen sonuçlar (Accuracy: 0.671, F1: 0.663) oldukça düşüktür.

intfloat/multilingual-e5-large-instruct:

- Bu temsil yöntemi, **SVM** ile iyi bir performans sergilemiş (Accuracy: 0.816, F1: 0.815). Ancak RandomForest ve MLP sonuçları daha düşük düzeyde kalmıştır.

ELECTRA (dbmdz/electra-base-turkish-cased-discriminator):

- Görece dengeli bir performans sergilemiştir. En iyi sonuç, MLP sınıflandırıcısıyla elde edilmiştir (Accuracy: 0.802, F1: 0.801).

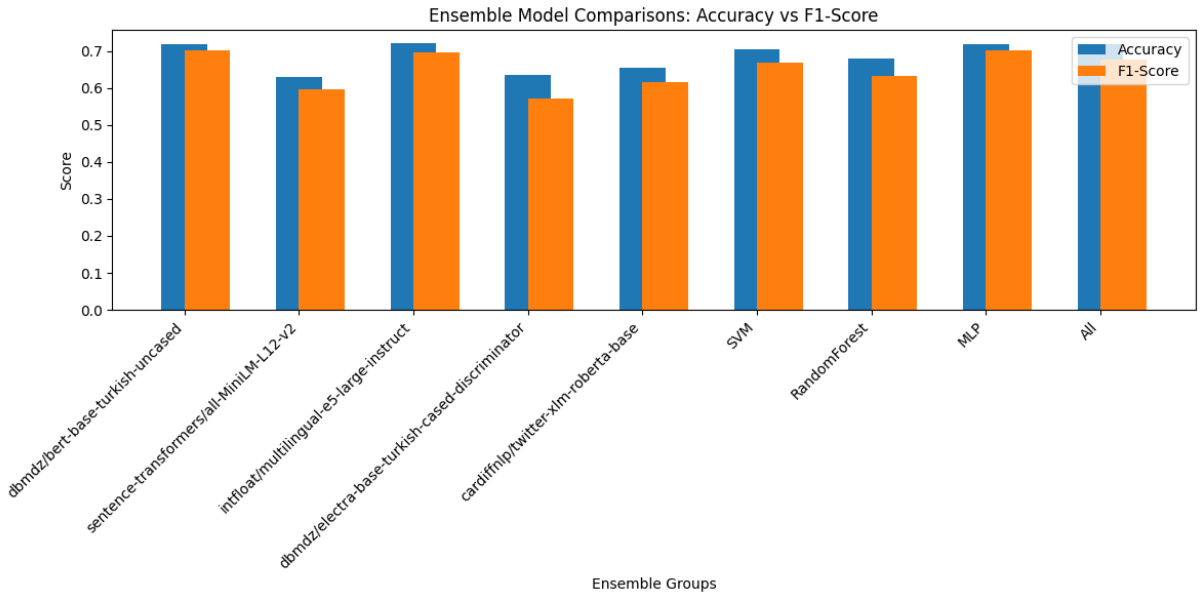
Cardiffnlp (twitter-xlm-roberta-base):

- Sosyal medya verileri için optimize edilmiş bu temsil yöntemi, özellikle MLP sınıflandırıcısıyla iyi sonuçlar vermiş (Accuracy: 0.811, F1: 0.811).

Genel Trendler:

- **SVM**: Çoğu temsil yöntemi için en yüksek performansı sağlıyor.
- **RandomForest**: Temsil yöntemlerinden bağımsız olarak daha düşük performans sergilemiş.
- **MLP**: Bazı temsil yöntemlerinde iyi sonuçlar vermiş, ancak genelde **SVM** kadar güçlü değil.

Duygu Analizi Veri Seti:



Sonuçların Kısa Yorumu:

En Yüksek Performans:

- **dbmdz/bert-base-turkish-uncased** yöntemi, hem Accuracy hem de F1-Score açısından en yüksek performansı göstermiş. Bu temsil yöntemi, SVM sınıflandırıcısıyla (Accuracy: 0.849, F1: 0.849) diğer kombinasyonlardan açık ara önde.

Sentence-Transformers (all-MiniLM-L12-v2):

- Bu temsil yöntemi, diğerlerine kıyasla daha düşük performans göstermiş. Özellikle RandomForest sınıflandırıcısıyla elde edilen sonuçlar (Accuracy: 0.671, F1: 0.663) oldukça düşüktür.

intfloat/multilingual-e5-large-instruct:

- Bu temsil yöntemi, **SVM** ile iyi bir performans sergilemiş (Accuracy: 0.816, F1: 0.815). Ancak RandomForest ve MLP sonuçları daha düşük düzeyde kalmıştır.

ELECTRA (dbmdz/electra-base-turkish-cased-discriminator):

- Görece dengeli bir performans sergilemiştir. En iyi sonuç, MLP sınıflandırıcısıyla elde edilmiştir (Accuracy: 0.802, F1: 0.801).

Cardiffnlp (twitter-xlm-roberta-base):

- Sosyal medya verileri için optimize edilmiş bu temsil yöntemi, özellikle MLP sınıflandırıcısıyla iyi sonuçlar vermiş (Accuracy: 0.811, F1: 0.811).

Genel Trendler:

- **SVM**: Çoğu temsil yöntemi için en yüksek performansı sağlıyor.
- **RandomForest**: Temsil yöntemlerinden bağımsız olarak daha düşük performans sergilemiş.
- **MLP**: Bazı temsil yöntemlerinde iyi sonuçlar vermiş, ancak genelde **SVM** kadar güçlü değil.

7. Genel Karşılaştırma

Bu bölümde, farklı temsil yöntemleri (embeddings) ve sınıflandırıcıların performansları Accuracy ve F1-Score metrikleri üzerinden karşılaştırılmıştır. Ayrıca, ensemble yöntemlerinin etkisi değerlendirilmiştir.

Temsil Yöntemlerinin Performans Karşılaştırması:

1. **dbmdz/bert-base-turkish-uncased:**
    - En yüksek performansı sağlayan temsil yöntemidir.
    - SVM ile hem Accuracy hem de F1-Score açısından en iyi sonuçları vermiştir (Accuracy: ~0.85, F1-Score: ~0.85).
    - Türkçe diline özel optimize edilmiş olması bu başarının temel nedenidir.
2. **intfloat/multilingual-e5-large-instruct:**
    - Performans açısından ikinci sırada yer almıştır.
    - Çok dilli çalışmalara uygun bir temsil yöntemi olup SVM ile dikkat çekici sonuçlar sağlamıştır (Accuracy: ~0.81, F1-Score: ~0.81).
3. **sentence-transformers/all-MiniLM-L12-v2:**
    - Daha düşük performans göstermiştir.
    - Accuracy ve F1-Score değerleri diğer temsil yöntemlerine göre geride kalmıştır (Accuracy: ~0.73, F1-Score: ~0.73).

4. **dbmdz/electra-base-turkish-cased-discriminator:**
    - Dengeli bir performans sergilemiştir, ancak BERT modeline kıyasla daha düşük sonuçlar elde etmiştir.
5. **cardiffnlp/twitter-xlm-roberta-base:**
    - Sosyal medya metinleri için optimize edilmiş olmasına rağmen, genel sınıflandırma görevlerinde daha düşük bir performans sergilemiştir.

Algoritmaların Performans Karşılaştırması:

1. **SVM (Support Vector Machines):**
    - Çoğu temsil yöntemi ile en iyi sonuçları sağlamıştır.
    - Accuracy ve F1-Score açısından dengeli ve tutarlı bir performans göstermiştir.
2. **RandomForest:**
    - Genel olarak diğer algoritmalara göre daha düşük performans sergilemiştir.
    - Özellikle temsil yöntemlerinin performansını tam olarak yansıtamamıştır.
3. **MLP (Multi-Layer Perceptron):**
    - BERT ve diğer güçlü temsil yöntemlerinde başarılı sonuçlar sağlamış, ancak genelde SVM kadar yüksek performans göstermemiştir.

Ensemble Yöntemlerinin Performansı:

1. **Aynı Temsil Yöntemi için Ensemble:**
    - SVM, RF ve MLP gibi farklı sınıflandırıcıların tahminleri birleştirildiğinde, bireysel sınıflandırıcılardan daha dengeli ve kararlı sonuçlar elde edilmiştir.
2. **Aynı Algoritma için Ensemble:**
    - Farklı temsil yöntemlerinden elde edilen tahminlerin birleşimi, bireysel temsil yöntemlerine kıyasla daha yüksek genel performans sağlamıştır.
3. **Tüm Sonuçların Ensemble'ı:**
    - Temsil yöntemleri ve algoritmaların tamamını birleştiren bu yaklaşım, Accuracy ve F1-Score açısından en dengeli sonuçları sağlamıştır.
    - Model çeşitliliği sayesinde bireysel modellerin eksiklerini tamamlamış ve daha kararlı sonuçlar elde edilmiştir.

Genel Bulgular ve Öneriler:

- **En Başarılı Kombinasyon: dbmdz/bert-base-turkish-uncased + SVM**, Accuracy ve F1-Score açısından en iyi sonuçları sağlamıştır.
- **Ensemble Yöntemleri:** Tahmin sonuçlarını birleştiren ensemble yöntemleri, bireysel modellerin performansını artırarak daha tutarlı sonuçlar üretmiştir.