Temsillerin/Kararların Birleştirilmesi

Müşerref Özkan - 235B7053

1.Spam Mesaj Tahmini

1.1.Giriş

Bu çalışmada, Türkçe mesajların spam veya normal olarak sınıflandırılması hedeflenmiştir. Proje kapsamında, metin verileri farklı temsil yöntemleriyle vektörleştirilmiş ve Destek Vektör Makineleri, Rastgele Ormanlar, Çok Katmanlı Algılayıcı gibi algoritmalarla modellenmiştir. Bunun yanında, temsil yöntemleri ve sınıflandırıcılar üzerinden ensemble teknikleri uygulanarak sonuçlar kıyaslanmıştır. Amaç, sınıflandırma performansını artırarak dengeli bir sonuç elde etmektir.

1.2. Veri Seti ve Özellikler

Veri seti, 5768 satır ve 3 sütundan oluşmaktadır. “Unnamed: 0” isimli sütun index bilgilerini, “text” sütunu mesaj metnini ve hedef değişken “sonuç” sütunu ise mesajın spam ya da normal olması bilgisini bulundurmaktadır. Sınıflar 3051 spam, 2717 normal olmak üzere dengeli sayılabilecek bir dağılıma sahiptir. Veri setinde eksik değer bulunmamaktadır.

1.3. Temsil Yöntemleri

Projede kullanılan temsil yöntemleri aşağıda sıralanmıştır.

- **all-MiniLM-L12-v2:** Düşük hesaplama gereksinimlerine rağmen yüksek performans sağlar. Özellikle semantik anlam benzerliklerini bulmak için optimize edilmiştir.
- **multilingual-e5-large-instruct:** Çok dilli metin verilerini işlemek üzere tasarlanmış bir metin gömme modelidir.
- **gte-large:** Özellikle büyük ölçekli metin verileri üzerinde etkili bir şekilde çalışabilmek için tasarlanmış güçlü bir gömme modelidir. Yüksek boyutlu vektör temsilleri sağlar ve geniş veri kümelerinde anlamlı sonuçlar üretir.
- **bert-base-turkish-uncased:** Türkçe diline özgü ve BERT tabanlıdır. Türkçe metin analizi, duygu analizi ve metin sınıflandırma gibi görevlerde yüksek performans sağlar.
- **jina-embeddings-v3:** Gniş bir kullanım yelpazesi için optimize edilmiş bir gömme modelidir.

1.4. Model

Destek vektör makineleri, Rastgele Ormanlar, Çok Katmanlı Algılayıcı modelleri kullanılmıştır. Optimizasyon işleminde kullanılacak olan model parametreleri aşağıda paylaşılmıştır.

Tablo1. Optimizasyon İçin Model Parametre Değerleri

Destek Vektör Makineleri		Rastgele Ormanlar		Çok Katmnalı Algılayıcı	
C:	0.1, 1	n_estimators:	500,100	hidden_layer_size	(50,), (100,)
Kernel:	lienar, rbf	max_depth:	10, None	max_iter	300,500

1.5. Uygulama

Veri seti %80 eğitim - %20 test seti olarak ayrılmıştır. Ayrım yapıldıktan sonra her iki veri seti içinde sınıf dağılımları kontrol edilmiştir. Sınıf değerleri için label encoder işlemi uygulanmıştır.

Her bir temsil yöntemi için embedding işlemi uygulaması için fonksiyon oluşturulup sonuçlar bir klasöre kaydedilmiştir.

Bireysel modeller üzerinden eğitimler gerçekleştirilmiştir. GridSearchCV yöntemi ile optimizasyon işlemi uygulanmıştır. Optimizasyon işleminde katman sayısı üç olarak belirlenmiştir.

Aynı temsil yöntemine ait sonuçları, aynı modele ait tüm sonuçları ve tüm sonuçları birleştirerek ensemble yöntemi uygulanmıştır. Modellerinin doğruluk hesabı için accuracy, recall, precision, f1 score ve confusion matrix hesaplaması yapılmıştır.

1.6. Sonuçlar

Sonuçlar bölümünde, kullanılan temsil yöntemleri ve modeller üzerinden alınan sonuçlar detaylı bir şekilde incelenmiş ve ensemble yöntemleriyle elde edilen başarı oranları analiz edilmiştir.

Tablo 2. Aynı Temsil Yöntemine Ait Sonuçların Birleştirilmesi

	all-MiniLM-L12-v2	**multilingual-e5-large-instruct**	**gte-large**	**bert-base-turkish-uncased**	**jina-embeddings-v3**
SVM	0.96	0.98	0.95	0.98	0.98
RF	0.91	0.97	0.92	0.96	0.97
MLP	0.95	0.98	0.94	0.98	0.97
Ensemble	0.95	0.98	0.95	0.98	0.98

Farklı temsil yöntemlerine göre bireysel modellerin ve ensemble sonuçlarının arasında belirgin faklılık yoktur. Bireysel modellerin tahmin başarısı da yüksektir.

Tablo 3. Aynı Modele Ait Sonuçların Birleştirilmesi

Model	Accuracy
Ensemble SVM	0.97
Ensemble RF	0.97
Ensemble MLP	0.97

Aynı model için birleştirilme sonuçlarına göre modeller arasında bir fark olmadığı görüşmüştür.

Tüm sonuçlar birleştirildiğinde test veri seti için accuracy değeri 0.97 olarak hesaplanmıştır.

Tablolardan elde edilen sonuçlara göre, temsil yöntemleri ve modeller arasında belirgin performans farklılıkları görülmemektedir. Örneğin, 'multilingual-e5-large-instruct'' 'bert-base-turkish-uncased' ve 'jina-embeddings-v3' temsil yöntemleri özellikle yüksek başarı oranları sergilemiştir. Ensemble yöntemleri ise genel olarak bireysel modellerden daha dengeli ve başarılı sonuçlar üretmiştir.

2. Duygu Analizi

2.1.Giriş

Bu çalışmada, yapılan alışverişler sonucunda müşterinin ürün hakkında yorumu ve duygu durumu bilgileri üzerine çalışılmıştır. Proje kapsamında, metin verileri farklı temsil yöntemleriyle vektörleştirilmiş ve Destek Vektör Makineleri, Rastgele Ormanlar, Çok Katmanlı Algılayıcı gibi algoritmalarla modellenmiştir. Bunun yanında, temsil yöntemleri ve sınıflandırıcılar üzerinden ensemble teknikleri uygulanarak sonuçlar kıyaslanmıştır. Amaç, sınıflandırma performansını artırarak dengeli bir sonuç elde etmektir.

2.2. Veri Seti ve Özellikler

Veri seti, 8491 satır ve 2 sütundan oluşmaktadır. "Görüş" sütunu müşterinin ürün hakkında yorumlarını, "Durum" sütunu müşterinin yorumu için duygu bilgisini bulundurmaktadır. "Görüş" sütununda bulunan iki eksik değer silinmiştir. Sınıflar 4252 olumlu, 4237 olumsuz olmak üzere dengeli sayılabilecek bir dağılıma sahiptir.

2.3. Temsil Yöntemleri

Projede kullanılan temsil yöntemleri aşağıda sıralanmıştır.

- **all-MiniLM-L12-v2:** Düşük hesaplama gereksinimlerine rağmen yüksek performans sağlar. Özellikle semantik anlam benzerliklerini bulmak için optimize edilmiştir.
- **multilingual-e5-large-instruct:** Çok dilli metin verilerini işlemek üzere tasarlanmış bir metin gömme modelidir.
- **gte-large:** Özellikle büyük ölçekli metin verileri üzerinde etkili bir şekilde çalışabilmek için tasarlanmış güçlü bir gömme modelidir. Yüksek boyutlu vektör temsilleri sağlar ve geniş veri kümelerinde anlamlı sonuçlar üretir.
- **bert-base-turkish-uncased:** Türkçe diline özgü ve BERT tabanlıdır. Türkçe metin analizi, duygu analizi ve metin sınıflandırma gibi görevlerde yüksek performans sağlar.
- **jina-embeddings-v3:** Gniş bir kullanım yelpazesi için optimize edilmiş bir gömme modelidir.

2.4. Model

Destek vektör makineleri, Rastgele Ormanlar, Çok Katmanlı Algılayıcı modelleri kullanılmıştır. Optimizasyon işleminde kullanılacak olan model parametreleri aşağıda paylaşılmıştır.

Tablo 4. Optimizasyon İçin Model Parametre Değerleri

Destek Vektör Makineleri		**Rastgele Ormanlar**		**Çok Katmnalı Algılayıcı**	
C:	0.1, 1	n_estimators:	500,100	hidden_layer_size	(50,), (100,)
Kernel:	lienar, rbf	max_depth:	10, None	max_iter	300,500

2.5. Uygulama

Veri seti %80 eğitim - %20 test seti olarak ayrılmıştır. Ayrım yapıldıktan sonra her iki veri seti içinde sınıf dağılımları kontrol edilmiştir. Sınıf değerleri için label encoder işlemi uygulanmıştır.

Her bir temsil yöntemi için embedding işlemi uygulaması için fonksiyon oluşturulup sonuçlar bir klasöre kaydedilmiştir.

Bireysel modeller üzerinden eğitimler gerçekleştirilmiştir. GridSearchCV yöntemi ile optimizasyon işlemi uygulanmıştır. Optimizasyon işleminde katman sayısı üç olarak belirlenmiştir.

Aynı temsil yöntemine ait sonuçları, aynı modele ait tüm sonuçları ve tüm sonuçları birleştirerek ensemble yöntemi uygulanmıştır. Modellerinin doğruluk hesabı için accuracy, recall, precision, f1 score ve confusion matrix hesaplaması yapılmıştır.

2.6. Sonuçlar

Sonuçlar bölümünde, kullanılan temsil yöntemleri ve modeller üzerinden alınan sonuçlar detaylı bir şekilde incelenmiş ve ensemble yöntemleriyle elde edilen başarı oranları analiz edilmiştir.

Tablo 5. Aynı Temsil Yöntemine Ait Sonuçların Birleştirilmesi

	all-MiniLM-L12-v2	**multilingual-e5-large-instruct**	**gte-large**	**bert-base-turkish-uncased**	**jina-embeddings-v3**
SVM	0.84	0.94	0.83	0.92	0.93
RF	0.76	0.93	0.79	0.89	0.93
MLP	0.82	0.94	0.83	0.91	0.93
Ensemble	0.84	0.94	0.83	0.92	0.94

“multilingual-e5-large-instruct” ve “jina-embeddings-v3” temsil yöntemlerinin birleşimi en iyi sonucu vermiştir. Bu yöntemler için modellerin bireysel sonuçları da diğer modellerle kıyaslandığında iyi olduğu görülmüştür.

Tablo6. Aynı Modele Ait Sonuçların Birleştirilmesi

Model	Accuracy
Ensemble SVM	0.93
Ensemble RF	0.93
Ensemble MLP	0.93

Aynı model için birleştirilme sonuçlarına göre modeller arasında bir fark olmadığı görüşmüştür.

Tüm sonuçlar birleştirildiğinde test veri seti için accuracy değeri 0.93 olarak hesaplanmıştır.

Tablolardan elde edilen sonuçlara göre, temsil yöntemleri ve modeller arasında belirgin performans farklılıkları görülmektedir. Örneğin, "multilingual-e5-large-instruct" ve "jina-embeddings-v3" temsil yöntemleri özellikle yüksek başarı oranları sergilemiştir. Ensemble yöntemleri ise genel olarak bireysel modellerden daha dengeli ve başarılı sonuçlar üretmiştir.