T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLER ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ
TEZLİ YÜKSEK LİSANS PROGRAMI

BLM 5109: KOLLEKTİF ÖĞRENME

TEMSİLLERİN / KARARLARIN BİRLEŞTİRİLMESİ

METİN USLU
235B7014

AHMED UĞUR
23501027

Öğretim Üyesi
Prof. Dr. Mehmet Fatih AMASYALI

İSTANBUL
Aralık, 2024

Temsillerin / Kararların Birleştirilmesi

Özet

Bu çalışmada, bireysel modellerin kararlarının birleştirilerek kollektif öğrenme yöntemlerinin metin sınıflandırma problemlerindeki etkisi incelenmiştir. Çalışmanın temel amacı, farklı temsil yöntemleri ve makine öğrenmesi algoritmalarının ayrı ayrı performanslarını değerlendirmenin yanı sıra, bu bireysel modellerin sonuçlarının kolektif bir şekilde birleştirilmesinin performansa olan katkısını analiz etmektir. Türkçe metin içeren iki farklı veri kümesi kullanılarak, metinlerin vektörel temsili için beş farklı embedding yöntemi (MiniLM, Jina, E5-Large, BGE-M3, ve GTE-Large) uygulanmış ve her temsil üzerinde üç makine öğrenimi algoritması (SVM, Random Forest, ve MLP) kullanılmıştır.

Elde edilen sonuçlar, bireysel modellerin performanslarını ortaya koymanın yanı sıra, ensemble yöntemleriyle temsile göre, algoritmaya göre ve global olarak birleştirilmiş tahminlerin daha dengeli ve yüksek performanslı sonuçlar sunduğunu göstermektedir. Bu analiz, bireysel kararların kollektif öğrenme çerçevesinde birleştirilmesinin etkili bir sınıflandırma modeli oluşturmadaki potansiyelini vurgulamaktadır.

1. Giriş

Kollektif öğrenme, bireysel modellerin veya karar mekanizmalarının birleştirilerek daha güçlü ve dengeli bir öğrenme modeli oluşturma yaklaşımını ifade eder. Bu yöntem, özellikle farklı veri temsilleri veya algoritmaların belirli bir problem üzerindeki performanslarının birleşimi ile çoğunlukla doğruluk başarısını artırmaktadır. Literatürde kollektif öğrenme yaklaşımları farklı modellerle uygulanarak metin sınıflandırma problemlerinde yaygın olarak kullanılmaktadır. Bu yöntemlerin genel amacı, bireysel modellerin hatalarını dengelemek ve sınıflandırma doğruluğunu artırmaktır.

Bu çalışmada, Türkçe haber metinleri ve ürün yorumları içeren iki farklı veri kümesi üzerinde sınıflandırma problemi, hem farklı temsil yöntemleri hem de farklı makine öğrenmesi algoritmaları kullanılarak kollektif öğrenmenin etkisi açısından incelenmiştir. Makine öğrenmesi modellerinin anlaşılır bir şekilde yorumlayabilmesi için gereken sayısal veri, metinlerin vektörlerle temsil edileceği embedding yöntemleri kullanılarak dönüştürülür. Bu temsil yöntemlerinin üzerinde çalışan makine öğrenmesi algoritmaları ise sınıflandırma başarısını artırmak için kritik rol oynar. Ancak, bu bireysel temsil ve algoritmaların performansı bir arada ele alındığında, kollektif bir öğrenme yöntemiyle elde edilen sonuçlar daha dengeli ve genellenebilir hale gelmektedir.

Çalışmada beş farklı temsil yöntemi (MiniLM, Jina, E5-Large, BGE-M3, GTE-Large) ve üç makine öğrenimi algoritması (SVM, Random Forest, MLP) kullanılmıştır. Bu yöntem ve algoritmaların bireysel performansları değerlendirilmiş ve ardından ensemble yöntemleri ile birleştirilerek kollektif bir öğrenme yaklaşımı benimsenmiştir. Çalışmanın temel amacı, bireysel temsillerin ve algoritmaların güçlü ve zayıf yönlerini analiz ederek bu bireysel sonuçların kollektif olarak nasıl iyileştirilebileceğini ortaya koymaktır.

Sonuçlar, temsil yöntemlerinin ve algoritmaların bireysel olarak sınıflandırma problemlerine katkısını göstermenin yanı sıra, kollektif öğrenmenin daha yüksek bir performans ve genelleme kabiliyeti sunduğunu ortaya koymaktadır.

2. Veri Kümesi

Bu çalışmada kararları birleştirme ve kollektif öğrenmenin genelleme kabiliyetini gözlemleyebilmek için farklı sayıda sınıf içeren iki adet Türkçe veri seti kullanılmıştır. Bunlar; “Dünya”, ”Gündem”, ”Türkiye”, ”Spor”, ”Eğitim”, ”Kültür-Sanat”, ”Yaşam”, ”Savunma”, ”Ekonomi”, ”Dünya Dışı”, ”Çevre”, ”Çocuk”, ”Bilim Teknoloji” ve ”Gezi” kategorilerinden oluşan toplam 14 sınıfa ayrılmış haberler (Turkish News 2024) ile “Pozitif” ve “Negatif” olarak 2 farklı sınıfa ayrılmış ürün yorumlarından (Turkish Product Reviews) oluşmaktadır.

2.1. Türkçe Haberler Veri Kümesi

Veri kümesi, 14 farklı sınıfı içeren, toplamda 27.387 adet habere ait başlık, özet, kategori, link ve içerik bilgilerinden oluşmaktadır. Ancak, çalışmada kullanılmak üzere her bir sınıftan 1.500 haber içerecek şekilde yalnızca 5 sınıfa (Türkiye, Dünya, Spor, Gündem ve Eğitim) ait veri kullanılmıştır. Bu veri kümesi, her bir sınıftan 1.500 adet içerecek şekilde toplamda 6.500 adet eğitim ve 1.000 adet test verisi olarak dengeli bir şekilde ayrılmıştır. Şekil 1'de veri kümesinden çeşitli sınıflardaki örnekler yer almaktadır.

Şekil 1 *– Türkçe Haberler Veri Kümesi Örnekleri*

Baslik string	Ozet string	Kategori string	Link string	Icerik string
Aksaray'dan Buckingham…	Buckingham Sarayı'ndan,…	Yaşam	https://www.trthaber.com/haber/yasam/aksaraydan-buckingham-sarayina-825802.html	Dünyanın bilinen en eski halısı Türklere ait bir Hun halısı. Altay'ın Pazırık…
Malatya'da trafik kazasında 1 çocuk…	Malatya'da trambüse çarpan hafif ticari…	Türkiye	https://www.trthaber.com/haber/turkiye/malatyada-trafik-kazasinda-1-cocuk-oldu-2-kisi-yaralandi-…	B.E. idaresindeki hafif ticari araç, Malatya-Elazığ kara yolunda Battalgazi…
"Kozmik oda kumpas" davasında cezalara…	Yargıtay Ceza Genel Kurulu, Ankara…	Gündem	https://www.trthaber.com/haber/gundem/kozmik-oda-kumpas-davasinda-cezalara-onama-841481.html	Kurul, ilk derece mahkemesi sıfatıyla Yargıtay 16. Ceza Dairesince görülen ve 2…
Antalya Doğu Garajı'ndaki…	Antalya Doğu Garajı Kültür Merkezi…	Kültür-Sanat	https://www.trthaber.com/haber/kultur-sanat/antalya-dogu-garajindaki-nekropol-alani-…	Doğu Garajı Nekropol Alanı Projesi tamamlandı. Antalya Müze Müdürlüğü…
SBÜ'de Jandarma adına eğitim…	Sağlık Bilimleri Üniversitesinde…	Eğitim	https://www.trthaber.com/haber/egitim/sbude-jandarma-adina-egitim-gorecek-adaylarin-sinav-…	Jandarma ve Sahil Güvenlik Akademisinden yapılan açıklamaya göre, Sağlık Bilimleri…

2.2. Türkçe Ürün Yorumları Veri Kümesi

Veri kümesi, toplamda 235.165 adet ürün yorumuna ait cümle ve karşılığındaki pozitif veya negatif olmak üzere 2 yorum tipini içermektedir. Ancak, çalışmada kullanılmak üzere her bir sınıftan 3.750 adet yorum içerecek şekilde düzenlenmiştir. Ayrıca, veri kümesi 6.500 adet eğitim ve 1.000 adet test verisi olmak üzere üç parçaya ayrılmıştır. Her bir alt veri kümesi kendi içinde her iki sınıftan da eşit sayıda örnekler içererek dengeli bir şekilde bölünmüştür. Şekil 2'de veri kümesinden örnekler yer almaktadır.

Şekil 2 *– Türkçe Ürün Yorumları Veri Kümesi Örnekleri*

sentence string	sentiment int64
beklentimin altında bir ürün kaliteli değil	0
3. kademe hız da motor titreme yapıyor.bu sebebten 3.kademe kullanımda uzun ömürlü olacağını zannetmiyorum.elide rahatsız ediyor.	0
başlığı sabit durmuyor. arka koruma demiri üfleme hızını 2.sewiye yapınca yüksek ses cikariyor	0
ürün kaliteli ve çok kullanışlı kargo ya verilmesi ve elime çabuk ulaşması da gayet başarılı idi	1
fiyat-performans karşılaştırması yaptığınızda gayet iyi bir ürün. en önemlisi ısınma yapmıyor.	1

Bu çalışmada, MiniLM, Jina, E5-Large, BGE-M3, ve GTE-Large embedding yöntemleri kullanılarak metinler vektörek temsillere dönüştürülmüştür. Her bir yöntem, metinlerin farklı özelliklerini yakalamayı hedeflemiş ve vektör boyutları, seçilen embedding modeline bağlı olarak değişmiştir. Bu süreç, modelleme adımlarında farklı temsil yöntemlerinin performansını karşılaştırmayı mümkün kılmıştır. Tablo 1'de kullanılan temsil yöntemlerinin vektör boyutları verilmiştir.

Tablo 1 *– Temsil Yöntemlerinin Vektör Boyutları*

Model	Vektör Boyutu
Model 1 - MiniLM	384
Model 2 - Jina	1024
Model 3 - E5-Large	1024
Model 4 - BGE-M3	1024
Model 5 - GTE-Large	1024

3. Deneysel Analiz

Bu çalışmada kullanılan beş farklı embedding yöntemi (MiniLM, Jina, E5-Large, BGE-M3, ve GTE-Large), metinlerin vektörel temsiline yönelik literatürde yaygın olarak kullanılan, etkili ve başarılı modellerdir. Bu yöntemler, metinlerin anlamını daha iyi yakalamak ve makine öğrenimi algoritmaları için metinleri sayısal veri olarak temsil etmek amacıyla kullanılmıştır. MiniLM, yüksek performans ve düşük hesaplama maliyeti ile öne çıkarken, Jina ve E5-Large gibi modeller, çoklu dil desteği ve geniş kapsamlı metin temsili sağlama yetenekleriyle tercih edilmiştir. BGE-M3 ve GTE-Large ise metinlerin bağlamsal ilişkilerini yakalamadaki başarıları nedeniyle çalışmaya dahil edilmiştir.

Metinler, her bir embedding yöntemiyle Tablo 1'de belirtilen sabit boyutlu vektörlere dönüştürülmüştür. Embedding modelleri, önceden eğitilmiş dil modellerine dayanmaktadır ve kelime, cümle veya paragraf seviyesindeki bağlamsal bilgileri yakalamak için tasarlanmıştır. Bu süreç, metinlerin model tarafından anlamlı bir şekilde işlenmesini sağlayarak sınıflandırma algoritmalarının doğruluğunu artırmayı hedeflemiştir.

Çalışmada; Support Vector Machines (SVM), Random Forest (RF) ve Multi-Layer Perceptron (MLP) makine öğrenmesi algoritmaları kullanılmıştır. Bu algoritmalar için hiper parametre optimizasyonu yapılmış ve doğrulama verisi üzerinde en iyi sonuçların elde edildiği parametreler ile test verisi üzerindeki sınıflandırma performansı karşılaştırılmıştır.

3.1. Hiper Parametre Optimizasyonu

Makine öğrenimi modellerinin performansında hiper parametrelerin doğru şekilde ayarlanmasının etkisi yüksektir. Bu çalışmada, her algoritma ve veri kümesi için GridSearch ve RandomSearch yöntemleri kullanılarak hiper parametre optimizasyonu gerçekleştirilmiştir. Hiper parametrelerin Türkçe ürün yorumları eğitim verisi üzerindeki başarıları Tablo 2'de gösterilmiştir.

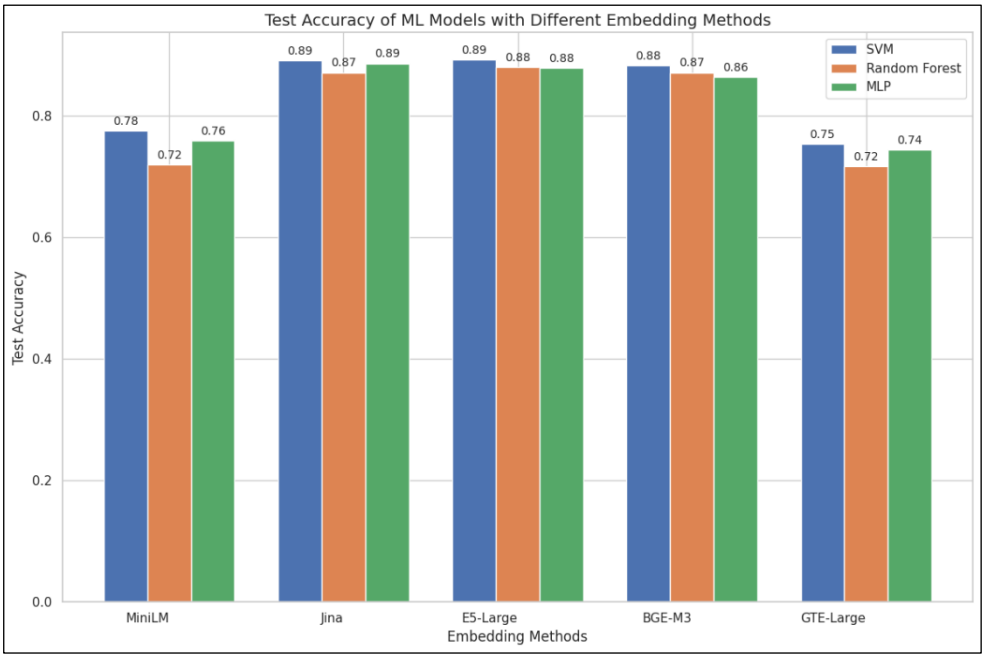
Tablo 2 *– Türkçe Ürün Yorumları Eğitim Verisi Üzerinde Hiper Parametre Optimizasyonu Sonuçları*

Model	Parametre	Denenen Değerler	Embedding	En Başarılı Değerler	Accuracy
SVM	C kernel gamma	0.1, 1, 10 linear, rbf, poly scale,auto	MiniLM	1, poly, scale	0.78
			Jina	1, linear, scale	0.89
			E5-Large	1, poly, scale	0.89
			BGE-M3		0.88
			GTE-Large	10, rbf, scale	0.75
RF	n_estimators max_depth min_samples_split	200, 500 10, None 2,3	MiniLM	500, None, 3	0.72
			Jina	500, None, 2	0.87
			E5-Large	500, 10, 2	0.88
			BGE-M3	500, None, 3	0.87
			GTE-Large	500, None, 2	0.72
MLP	hidden_layer_sizes activation learning_rate alpha	(100,), (100, 50) tanh, relu constant,adaptive 0.0001, 0.001	MiniLM	(100,), relu, constant, 0.001	0.76
			Jina	(100,), tanh, constant, 0.001	0.89
			E5-Large	(100,), relu, constant, 0.001	0.88
			BGE-M3		0.86
			GTE-Large		0.74

3.2. Modelin Eğitim ve Tahmini

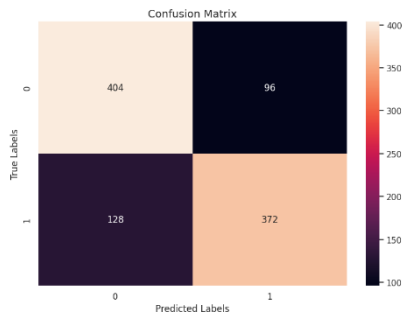
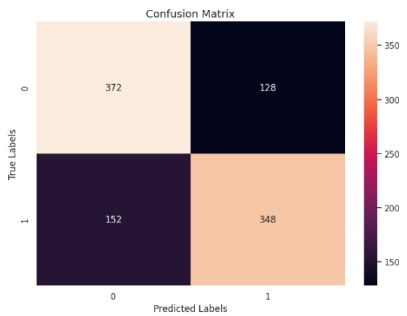
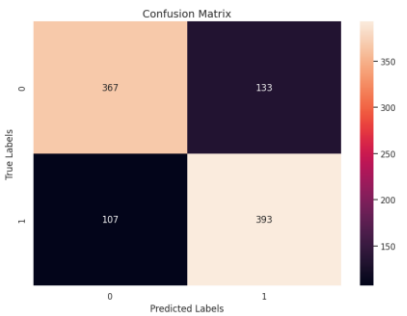
İki farklı veri kümesinde GridSearch ve RandomSearch yöntemleri ile elde edilen hiper parametreler kullanılarak model oluşturulmuştur. Her bir temsil yöntemi kullanılarak üç farklı makine öğrenmesi algoritması ile modelin test verisi üzerindeki değerlendirmesi accuracy ve F1 skoru kullanılarak yapılmıştır. Farklı temsil yöntemleri ve makine öğrenmesi algoritmalarının Türkçe ürün yorumları test verisi üzerindeki başarı Şekil 3'te gösterilmiştir.

Şekil 3 *– Türkçe Ürün Yorumları Test Verisi Başarısı*

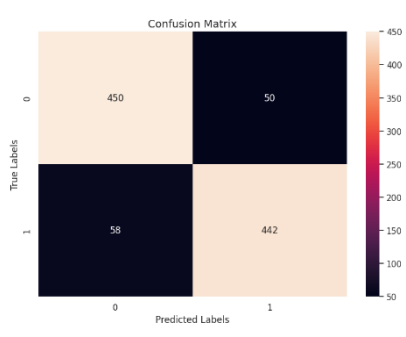
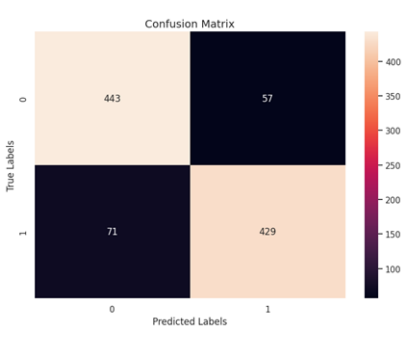
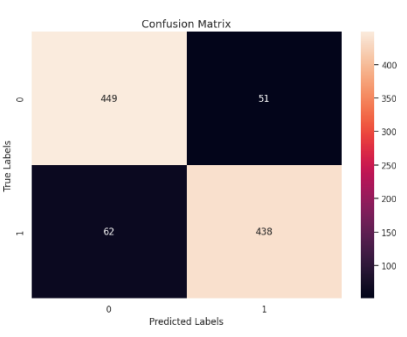


Modelin farklı temsil yöntemleri ve makine öğrenmesi algoritmaları kullandığında Türkçe ürün yorumları test verisi üzerindeki doğru ve hatalı tahminlerini gösteren karışıklık matrisi, Tablo [3-7]'de gösterilmiştir.

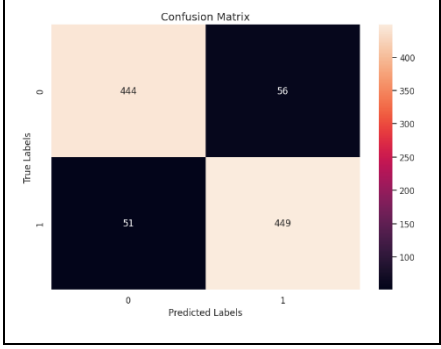
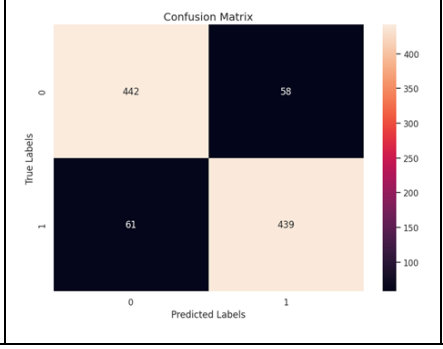
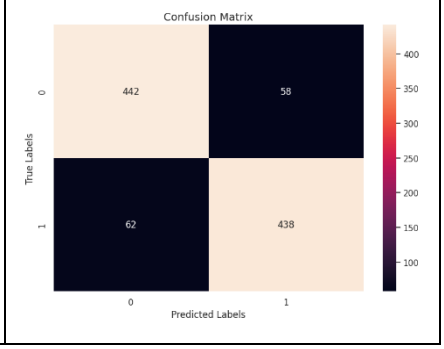
Tablo 3 *– MiniLM Temsil Yöntemi ile Farklı Makine Öğrenmesi Algoritmalarının Türkçe Ürün Yorumları Test Verisi Üzerindeki Tahminleri*

SVM	RF	MLP
		

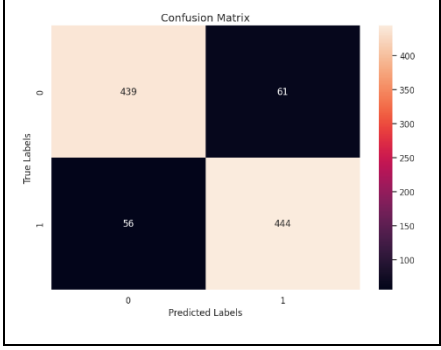
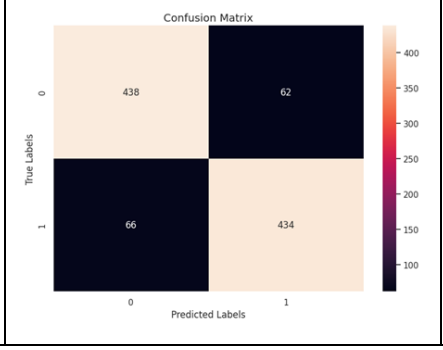
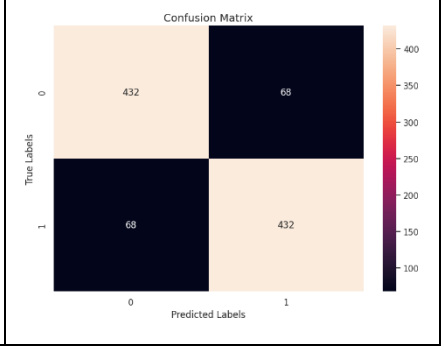
Tablo 4 *– Jina Temsil Yöntemi ile Farklı Makine Öğrenmesi Algoritmalarının Türkçe Ürün Yorumları Test Verisi Üzerindeki Tahminleri*

SVM	RF	MLP
		

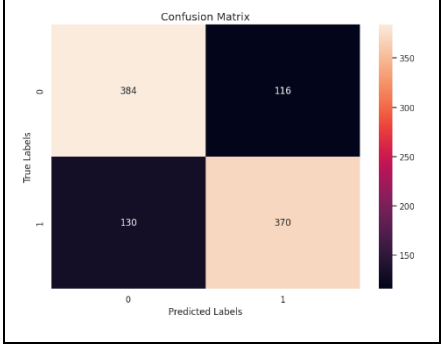
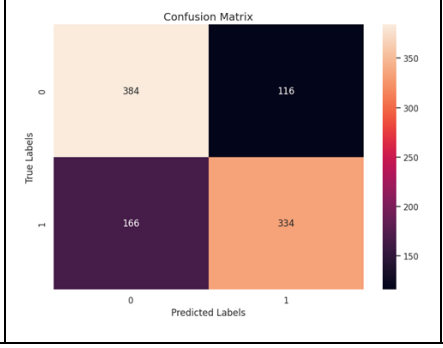
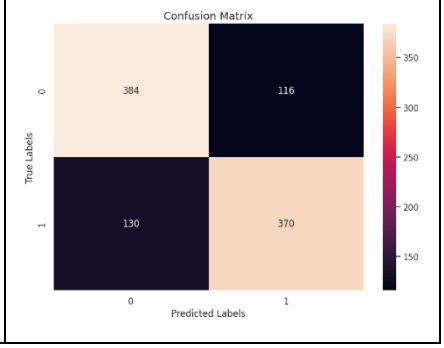
Tablo 5 *– E5-Large Temsil Yöntemi ile Farklı Makine Öğrenmesi Algoritmalarının Türkçe Ürün Yorumları Test Verisi Üzerindeki Tahminleri*

SVM	RF	MLP
		

Tablo 6 *– BGE-M3 Temsil Yöntemi ile Farklı Makine Öğrenmesi Algoritmalarının Türkçe Ürün Yorumları Test Verisi Üzerindeki Tahminleri*

SVM	RF	MLP
		

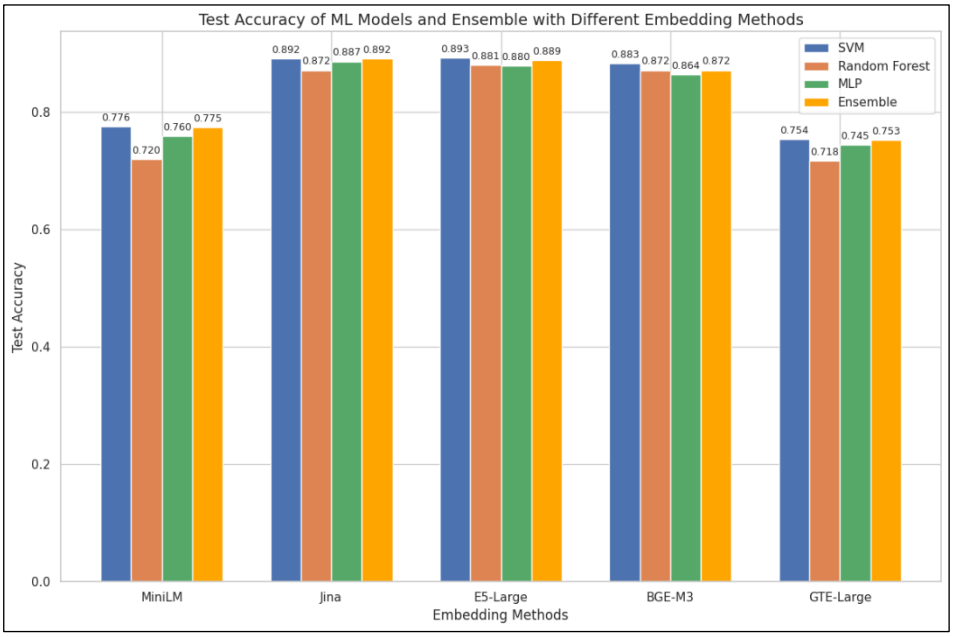
Tablo 7 *– GTE-Large Temsil Yöntemi ile Farklı Makine Öğrenmesi Algoritmalarının Türkçe Ürün Yorumları Test Verisi Üzerindeki Tahminleri*

SVM	RF	MLP
		

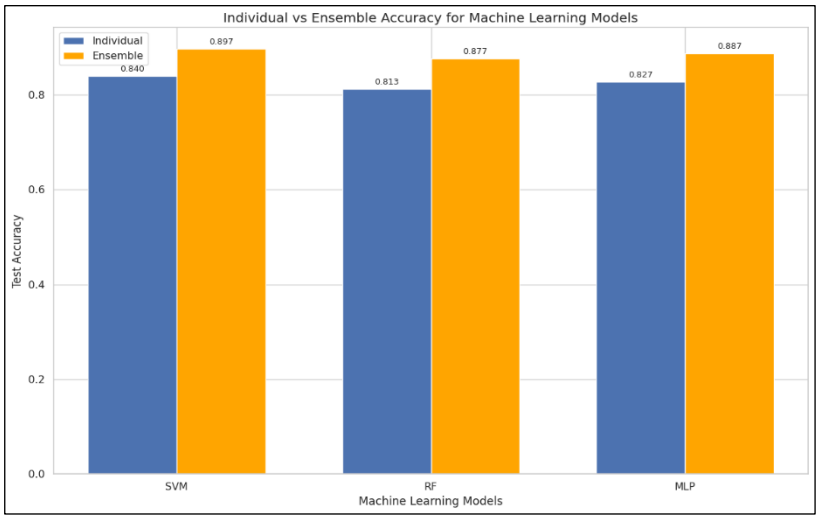
3.2. Ensemble Yöntemleri

Ensemble yöntemleri, farklı modellerin çıktılarının birleştirilmesiyle daha dengeli ve genellenebilir tahminler sağlamayı amaçlamaktadır. Bu çalışmada, temsile göre, makine öğrenmesi algoritmasına göre ve global olmak üzere üç farklı ensemble stratejisi oluşturulmuştur. Bireysel modellerin kararlarını birleştirmesinde demokrasi yöntemi (majority voting) kullanılmıştır. Türkçe ürün yorumları test verisi üzerindeki aynı temsil yöntemlerine ait kararlar birleştirildiğinde ensemble sonuçları Şekil 4'te, aynı makine öğrenmesi algoritmalarına ait kararlar birleştirildiğinde ensemble sonuçları Şekil 5'te, tüm kararlar birleştirildiğindeki sonuçlar ise Şekil 6'da gösterilmiştir. Sonuçlara göre bireysel modellerin sonuçlarının birleştirilmesiyle elde edilen ensemble sonuçları performansı artırmıştır.

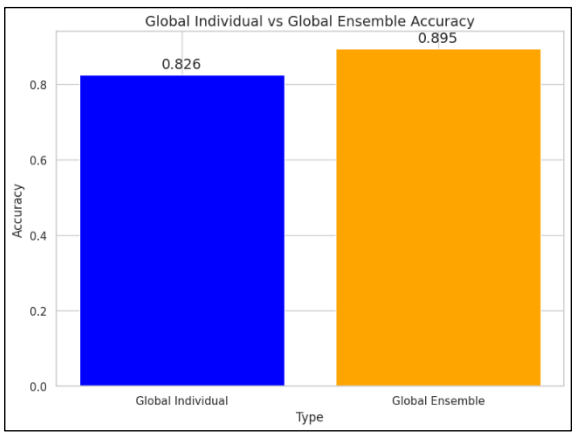
Şekil 4 *– Türkçe Ürün Yorumları Test Verisi Üzerinde Aynı Temsil Yöntemlerine Ait Kararların Birleştirilmesi*



Şekil 5 *– Türkçe Ürün Yorumları Test Verisi Üzerinde Aynı Makine Öğrenmesi Algoritmalarına Ait Kararların Birleştirilmesi*



Şekil 6 *– Türkçe Ürün Yorumları Test Verisi Üzerinde Tüm Kararların Birleştirilmesi*



4. Sonuç

Bu çalışmada, bireysel temsiller ve makine öğrenimi algoritmalarının performanslarının yanı sıra bu bireysel modellerin birleştirilmesinin etkisi kapsamlı bir şekilde incelenmiştir. Çalışmada kullanılan iki farklı Türkçe veri kümesi, temsillerin ve ensemble yöntemlerinin genelleme kabiliyetini değerlendirmek için uygun bir zemin sağlamıştır.

Bireysel modeller üzerinde yapılan deneysel analizler, her bir temsil yönteminin ve algoritmanın farklı güçlü yönlere sahip olduğunu göstermiştir. Özellikle E5-Large, Jina ve BGE M3 gibi embedding yöntemleri, geniş bağlamsal temsilleri yakalayabilme kapasiteleri sayesinde yüksek sınıflandırma başarısı elde etmiştir. Bununla birlikte, bireysel modellerin sonuçlarının birleştirilmesiyle elde edilen ensemble modelleri, performansta belirgin bir iyileşme sağlamış ve daha dengeli sonuçlar ortaya koymuştur. Kararların birleştirilmesinin bireysel başarı ortalamasının daha da performanslı hale geldiği ve bu sayede sonuçların genellenebilirliği arttığı gözlemlenmiştir. Bu sonuçlar, bireysel modellerin hatalarını dengelemek ve doğruluğu artırmak için kollektif öğrenme yöntemlerinin etkili bir yaklaşım olduğunu göstermektedir.

5. Kaynakça

Wang, S., Dong, L., & Hua, H. (2020, Aralık). Parameter optimization of support vector machine based on improved grid algorithm. In Journal of Physics: Conference Series (Vol. 1693, No. 1, p. 012108). IOP Publishing.

GridSearchCV kullanarak SVM Hiperparametre Ayarlama (2023, Ocak 11). ”GeeksforGeeks”. Erişim 29 Kasım 2024

Duran, A., & Bakır, H. (2023). Hiperparametreleri Ayarlanmış Makine Öğrenimi Algoritmalarını Kullanarak Android Sistemlerde Kötü Amaçlı Yazılım Tespiti. Uluslararası Sivas Bilim ve Teknoloji Üniversitesi Dergisi, 2(1), 1-19.

Kokoç, M., & Kokoç, F. (2021). Tahmin performansını arttırmak için kolektif öğrenme yaklaşımı: Engelli bireylere yönelik deneysel çalışma. Niğde Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi, 10(1), 100-114.