BLM5109 2. ÖDEV

TEMSİLLERİN VE KARARLARIN BİRLEŞTİRİLMESİ

Büşra Medine GÜRAL
20011038
medine.gural@std.yildiz.edu.tr

1. GİRİŞ

Bu çalışmada, farklı metin temsil yöntemleri ve makine öğrenmesi algoritmalarını kullanarak sınıflandırma problemleri çözülmüş ve bu modellerin performansları analiz edilmiştir. Beş farklı dil modeli (BERT [1], MiniLM [2], GTE [3], BGE [4], Jina [5]) ve üç farklı algoritma (SVM, RF, MLP) bir araya getirilerek toplamda 15 modelin sonuçları üretilmiştir. Daha sonra, aynı temsil yöntemine dayalı sonuçlar (5 ensemble), aynı algoritmaya dayalı sonuçlar (3 ensemble) ve tüm sonuçlar (1 ensemble) birleştirilerek karşılaştırmalar yapılmıştır. Performans ölçütü olarak F1 skoru ve accuracy kullanılmış, sonuçlar grafik ve tablolarla desteklenmiştir. Çalışma, 5000 ve 10000 örnek içeren iki farklı Türkçe metin veri kümesi üzerinde gerçekleştirilmiş ve temsil/karar birleştirme yöntemlerinin sınıflandırma başarısına etkisi analiz edilmiştir.

2. VERİ SETLERİ

Bu çalışmada, iki farklı Türkçe metin veri seti kullanılmış ve her biri üzerinde sınıflandırma modelleri eğitilerek temsil/karar birleştirme yöntemlerinin performansı değerlendirilmiştir. Her iki veri seti için, test kümesi tüm işlemlerden önce ayrılmış ve tüm yöntemlerde aynı test kümesi kullanılmıştır. Bu sayede, elde edilen sonuçların karşılaştırılabilirliği sağlanmıştır. Veri setleri hakkında detaylı bilgiler aşağıda sunulmuştur.

2.1 Turkish Product Reviews

Bu veri seti [6], kullanıcıların farklı ürünler hakkında bıraktıkları yorumlardan oluşmaktadır. Toplamda 235165 gözlemden oluşan veri seti, *sentence* (yorum metni) ile *sentiment* (duygu) olmak üzere iki temel özniteliğe sahiptir. *sentiment* sütunu, yorumların olumlu (1) veya olumsuz (0) olduğunu ifade etmektedir. 220284 olumlu yorum, 14881 olumsuz yorum bulunmaktadır.

Çalışmada, veri setinin daha hızlı işlenebilmesi ve analiz edilmesi için toplamda 5000 örnek seçilmiştir. Bu seçimin dengeli bir yaklaşımla yapılması sağlanmış ve veri setinden 2500 olumlu ve 2500 olumsuz yorum rastgele seçilmiştir. Böylelikle eğitim ve test kümelerinde sınıf dağılımının eşitliği korunarak modellerin her iki sınıfta eşit performans göstermesi hedeflenmiştir.

2.2 Interpress News Category TR

Bu veri seti [7], Türkçe haber metinlerinden oluşmaktadır ve *Content* (haber metni), *CategoryCode* (kategorinin kodu) dahil olmak üzere toplamda 6 sütun içermektedir. Veri seti; spor, ekonomi, teknoloji, kültür-sanat gibi 17 farklı kategoriye ait haberlerden oluşmaktadır ve bu bağlamda *CategoryCode* sütunu 0 ile 16 arasında değer almaktadır. Toplamda 218880 eğitim ve 54721 test verisi içermektedir.

Çalışmada, bu veri setinden 10000 örnek rastgele seçilmiştir. Ancak, veri setinin sınıf dağılımı dengesiz olduğundan (bazı kategoriler diğerlerine göre daha fazla sayıda örneğe sahiptir), eğitim ve test kümeleri oluşturulurken veri setindeki dağılım korunmuş ve *stratify* yöntemi kullanılarak ayrım yapılmıştır.

3. YÖNTEM

3.1 Veri

Bu çalışmada, farklı metin temsil yöntemleri ve makine öğrenimi algoritmalarının performansını karşılaştırmak için kullanılan iki farklı veri setinde izlenen yöntemler birbirine benzerdir. Bu nedenle, bu başlık altında temel adımlar detaylandırılmıştır.

Her iki veri setinde de analizlerin hızlı ve verimli bir şekilde gerçekleştirilebilmesi için tüm örnekler içinden 5000 ve 10000 örnek rastgele seçilmiştir. İlk veri setinde (Turkish Product Reviews), dengeli bir yaklaşım benimsenmiş ve her iki sınıftan (olumlu ve olumsuz) eşit sayıda örnek alınmıştır. İkinci veri setinde (Interpress News Category TR) ise veri setinin dengesiz yapısı korunmuş ve sınıflar arasındaki doğal dağılımda 10000 örnek seçilmiştir. Seçilen bu örnekler, %80 eğitim ve %20 test olacak şekilde stratify yöntemiyle ayrılmıştır. Bu, test kümesindeki sınıf dağılımının eğitim kümesindekiyle aynı kalmasını sağlamıştır.

3.2 Dil Modelleri ve Temsil Oluşturulması

Metinleri sayısal temsillere dönüştürmek için her iki veri setinde de aynı beş dil modeli kullanılmıştır:

- **Bert**: dbmdz/bert-base-turkish-uncased
- **Minilm**: sentence-transformers/all-MiniLM-L12-v2
- **Gte**: thenlper/gte-large
- **Jina**: jinaai/jina-embeddings-v3
- **Bge**: BAAI/bge-m3

bert-base-turkish-uncased, Türkçe diline özel olarak eğitilmiş bir BERT modelidir. Bavarian State Library tarafından geliştirilen bu model, 35 GB'lık bir veri kümesi üzerinde eğitilmiştir. Veri kümesi; Türkçe OSCAR corpus, Wikipedia, OPUS corpus ve Kemal Oflazer tarafından sağlanan özel bir veri setini içermektedir. Model, 12 katmanlı bir yapı, 768 gizli birim ve 12 dikkat başlığına sahiptir ve toplamda yaklaşık 110 milyon parametre içerir. Türkçe metin anlamlandırma görevlerinde bağlama duyarlı performansı ile öne çıkar.

all-MiniLM-L12-v2, hızlı ve etkili cümle temsilleri üretmek için optimize edilmiş, kompakt bir transformer modelidir. Model, 12 katmana ve 384 boyutlu embeddinglere sahiptir. Küçük boyutuyla, özellikle semantik benzerlik ve kümelenme gibi görevler için uygundur.

gte-large, Alibaba DAMO Akademisi tarafından geliştirilen ve çok aşamalı kontrastif öğrenme yöntemiyle eğitilmiş 330 milyon parametreli bir metin gömme modelidir [8]. BERT altyapısına dayanan GTE modelleri, üç farklı boyutta sunulmaktadır: GTE-large, GTE-base ve GTE-small. Bu modeller, geniş kapsamlı ve farklı alanları kapsayan büyük ölçekli bir metin çifti corpus üzerinde eğitilmiştir. Bu sayede, bilgi getirme, anlamsal metin benzerliği ve metin sıralama gibi metin gömme görevlerinde yüksek performans sergilemektedir.

jina-embeddings-v3, Jina AI tarafından geliştirilen ve 570 milyon parametreye sahip çok dilli bir metin gömme modelidir. Farklı diller ve alanlarda yüksek kaliteli metin temsilleri oluşturmak için geliştirilmiş bir modeldir. Bilgi erişimi ve semantik benzerlik gibi görevlerde üstün performans sergilemektedir. Model, metinlerin anlamlarını etkili bir şekilde yakalayan embeddingler üretir ve bilgi tabanlı sistemlerde güçlü bir performans sağlar.

bge-m3, Pekin Yapay Zeka Akademisi (BAAI) tarafından geliştirilmiş, çok işlevli ve çok dilli 560 milyon parametreli bir modeldir. Model, yoğun bilgi erişimi, çoklu vektör temsilleri ve farklı metin uzunluklarında bağlamı anlama gibi görevlerde kullanılır. Bağlamı anlamadaki başarısı ve çok yönlü yapısı ile dikkat çeker.

Modeller, Hugging Face Transformers kütüphanesiyle entegre edilmiştir. Her bir model için eğitim ve test verilerinden temsiller üretilmiş ve .npy formatında kaydedilmiştir. Bu yöntem, yeniden kullanım durumlarında işlemleri hızlandırmıştır.

3.3 Algoritmalar ve Başarı Ölçümü

Beş farklı dil modelinden üretilen temsiller; SVM, RF ve MLP olmak üzere üç farklı makine öğrenimi algoritması kullanılarak sınıflandırılmıştır:

Destek Vektör Makineleri (SVM), sınıflandırma problemlerinde yaygın olarak kullanılan bir yöntemdir [9]. Bu çalışmada, SVM modeli doğrusal olmayan sınıflandırmalar için kernel fonksiyonlarıyla desteklenmiştir. SVM'nin avantajı, farklı sınıflar arasındaki marjini maksimize eden karar sınırları oluşturmasıdır. Model, her bir dil modelinden üretilen embeddinglerle eğitilmiş ve sınıflandırma problemlerinde yüksek doğruluk ve F1 skorları elde etmiştir.

Random Forest, karar ağaçlarının topluluk öğrenme yaklaşımı ile birleştirilmesinden oluşan güçlü bir algoritmadır [10]. RF, karar ağaçlarını rastgele seçilmiş alt veri kümelerinde eğitir ve sonrasında tahminleri birleştirir. Bu yöntem hem doğruluğu artırır hem de aşırı öğrenmeyi önler. Çalışmada RF modeli, özellikle daha dengeli veri setlerinde sınıf tahminlerinde başarılı olmuştur.

Çok Katmanlı Algılayıcı (MLP), yapay sinir ağlarının temel bir biçimidir. MLP, en az bir gizli katmana sahip olup, doğrusal olmayan ilişkileri modelleyebilme kapasitesine sahiptir [11]. Bu çalışmada MLP, embeddingleri işlemek için kullanılmış ve optimize edilmiştir. Sinir ağı yapısı, her dil modelinin ürettiği yüksek boyutlu temsillerden öğrenerek, metin sınıflandırmada etkili bir performans göstermiştir. Ancak, diğer algoritmalara göre daha uzun bir eğitim süresine ihtiyaç duymuştur.

Her bir algoritma, eğitim verilerinden üretilen temsillerle eğitilmiş ve test veri setinde tahmin gerçekleştirilmiştir. Performans ölçütü olarak F1 skoru ve doğruluk (accuracy) kullanılmıştır. F1 skoru, sınıflandırma modelinin dengeli performansını değerlendirirken, doğruluk ise genel başarıyı ölçmüştür. Bu iki metrik, özellikle dengesiz sınıf dağılımına sahip veri setlerinde bir arada kullanılarak modellerin güçlü ve zayıf yönlerini belirlemede etkili olmuştur.

3.4 Ensemble Yöntemleri

Temsil tabanlı birleştirme için aynı dil modelinden gelen üç farklı algoritmanın tahminleri birleştirilerek bir *representation ensemble* oluşturulmuştur. Çoğunluk oylamasıyla sonuç belirlenmiştir. Olasılık eşitliği durumunda, modellerin tahmin olasılıkları dikkate alınarak en yüksek olasılığa sahip sınıf seçilmiştir. Algoritma tabanlı birleştirme için aynı algoritmadan gelen beş farklı dil modelinin tahminleri birleştirilerek bir *algorithm ensemble* oluşturulmuştur. Genel ensemble için ise tüm sonuçlar birleştirilmiştir.

4. DENEYSEL ÇALIŞMA VE SONUÇLAR

4.1 Turkish Product Reviews ile Çalışmalar

Veri setine ait rastgele beş gözlem aşağıdaki tabloda verilmiştir.

	sentence	sentiment
109690	ürün harika süper hiç bir başka ürüne değişemem değiştiremem kullanmaya devam edicam inşallah aynı kalitede devam eder devem edeceğinede inanıyorum teşekkürler fairy teşekkürler hepsi burada com uygun fiatta sattığınız için inşallah firmada biraz daha indirim yapar böylece herkes faydalanır	1
62867	ürün beklediğimden güzel ve uygun hepsiburadaya tesekkürler.	1
14297	çok güzel bir takım aldığımın ertesi gün kargom geldi.çok hızlı çok güzel takım çok beğendim.keşke servis ve salata tabaklarıda olsa..	1
222865	çok çabuk ısınıyor. ve arada donma oluyor. ama alınabilir bir ürün.	0
147305	sipariş verdikten sonraki gün geldi ben iki adet aldım sonra zam geldi 3 adet almadığıma pişman oldum tavsiye ederim. çil çil altın	1

Veri setindeki tüm satırlar içerisinden 2500 pozitif (1) ve 2500 negatif (0) gözlem rastgele seçilmiştir. Bu verilerin %20'si test setinde kullanılmıştır.

	Train Set	Test Set
1	2000	500
0	2000	500
Toplam	4000	1000

Train ve test kümelerindeki her bir cümlenin temsili oluşturularak kaydedilmiştir ve modeller train kümesinin temsilleri ile eğitilmiştir. Ardından test temsilleri üzerinde tahmin yapılmıştır. Modellerin tekil başarıları aşağıdaki tabloda gösterilmiştir.

	F1	Accuracy
BERT + SVM	0.865966	0.866
BERT + RF	0.831579	0.832
BERT + MLP	0.848996	0.849
MINILM + SVM	0.737895	0.738
MINILM + RF	0.686473	0.687
MINILM + MLP	0.729983	0.730
GTE + SVM	0.727205	0.728
GTE + RF	0.696081	0.697
GTE + MLP	0.713225	0.714
JINA + SVM	0.887998	0.888
JINA + RF	0.867911	0.868
JINA + MLP	0.882948	0.883
BGE + SVM	0.885978	0.886
BGE + RF	0.869000	0.869
BGE + MLP	0.858925	0.859

BERT modeli, Türkçe diline özel olarak eğitildiği için özellikle bağlam duyarlılığı yüksek görevlerde güçlü bir performans sergilemiştir. SVM ile kombinasyonu en iyi sonuçları üretirken, MLP de başarılı bir alternatif olarak öne çıkmıştır. Bununla birlikte, RF algoritması, BERT'in ürettiği yüksek boyutlu temsilleri işlemekte diğer algoritmalara kıyasla daha sınırlı bir performans göstermiştir.

MiniLM modeli, daha kompakt bir yapıya sahip olduğundan, düşük boyutlu temsiller üretmektedir. Bu durum, işlem süresini kısaltsa da bağlam duyarlılığını kısıtlayarak performansı olumsuz yönde etkileyebilmektedir. SVM, MiniLM ile elde edilen sonuçlarda diğer algoritmalara göre daha iyi performans sergilemiş olsa da genel başarı düzeyi diğer modellerin gerisinde kalmıştır. Bu, MiniLM'in daha az karmaşık görevlerde daha etkili olabileceğini göstermektedir.

GTE modeli, büyük ölçekli ve çoklu aşamalı kontrastif öğrenim ile eğitilmesine rağmen, bağlama duyarlılığı gerektiren görevlerde beklenen performansı tam olarak sağlayamamıştır. Bu model, SVM ile birlikte kullanıldığında en iyi sonuçları elde etmiş ancak MiniLM gibi diğer modellere kıyasla daha düşük bir başarı göstermiştir. Bu durum, modelin eğitim veri seti ile Turkish Product Reviews veri seti arasındaki bağlam farklılıklarından kaynaklanıyor olabilir.

Jina modeli, uzun metinlerde bağlamı etkili bir şekilde anlamlandırabilen güçlü bir model olarak öne çıkmıştır. Özellikle SVM ile kombinasyonu en iyi sonucu vermiştir. Bu, modelin genel amaçlı metin gömme yeteneklerinin ve bağlam anlama kapasitesinin yüksek olduğunu göstermektedir. Benzer şekilde, BGE modeli de bağlam anlama ve yoğun bilgi erişimi açısından başarılı sonuçlar sunmuştur. BGE'nin SVM ile kombinasyonu, Jina modeline yakın sonuçlar üretmiş ve bu iki model, genel performans açısından liderlik etmiştir.

Makine öğrenimi algoritmaları arasında, SVM genel olarak tüm modellerde en yüksek performansı sağlamıştır. Bunun nedeni, SVM'nin temsil boyutlarından bağımsız olarak sınıf ayırıcı bir hiper düzlem oluşturabilmesi ve özellikle bağlama duyarlı temsillerde etkili çalışabilmesidir. MLP ise, doğrusal olmayan ilişkileri öğrenme kapasitesiyle genel olarak iyi bir performans göstermiştir ancak eğitim süresinin uzun olması bir dezavantaj olarak öne çıkmaktadır. RF algoritması ise karar ağaçları kullanarak sınıflandırma yapmasına rağmen, bağlam duyarlılığı yüksek embeddinglerle yeterince iyi sonuçlar üretememiştir.

Temsillerin birleştirilmesiyle oluşan ensemble başarılar aşağıdaki tabloda gösterilmiştir.

	F1	**Accuracy**
BERT	0.862939	0.863
MINILM	0.757938	0.758
GTE	0.740155	0.741
JINA	0.891993	0.892
BGE	0.881995	0.882

BERT modeli, temsil birleştirme yönteminde güçlü bir performans sergilemiştir. Bu sonuç, BERT’in Türkçe dilinde eğitilmiş bağlama duyarlı temsiller üreterek ensemble yöntemine sağlam bir temel sağladığını göstermektedir.

MINILM modeli, daha kompakt bir yapıya ve düşük boyutlu temsillere sahip olmasına rağmen, ensemble yöntemi ile bireysel performansına göre bir iyileşme göstermiştir. Ancak, bu değerler BERT, Jina ve BGE modellerinin gerisinde kalmıştır.

GTE modeli, bireysel tahmin sonuçlarına paralel olarak ensemble yöntemiyle de orta seviyede bir başarı sağlamıştır. Büyük boyutlu ve çok aşamalı kontrastif öğrenimle eğitilmiş olmasına rağmen, GTE'nin bu veri setinde bağlamı tam anlamıyla yakalayamaması, genel performansını etkilemiştir.

JINA modeli, ensemble yöntemlerinde en iyi performansı sergileyerek liderliği sağlamıştır. Bu sonuç, modelin farklı algoritmaların tahminlerini birleştirerek dahi bağlama duyarlılığını koruduğunu ve güçlü temsiller ürettiğini göstermektedir.

BGE modeli de ensemble yönteminde oldukça başarılı bir performans göstermiştir. BGE'nin başarı düzeyi, bağlama duyarlılığı yüksek temsiller üretebilme kapasitesinden kaynaklanmaktadır. Jina ile kıyaslandığında, performans farkı oldukça azdır, bu da BGE'nin sınıflandırma görevlerinde Jina kadar etkili bir alternatif olduğunu göstermektedir.

Algoritmaların birleştirilmesiyle oluşan ensemble başarılar aşağıdaki tabloda gösterilmiştir.

	F1	**Accuracy**
SVM	0.884999	0.885
RF	0.863877	0.864
MLP	0.879969	0.88

SVM, algoritma bazlı ensemble yönteminde en yüksek performansı sergileyerek liderlik etmiştir. SVM'nin başarısı, sınıf ayırma yeteneğinin yüksek olması ve dil modellerinden gelen temsillerin karmaşıklığına duyarlı olmamasından kaynaklanmaktadır. SVM, özellikle yüksek boyutlu ve karmaşık temsilleri işleyebilme kabiliyetiyle ön plana çıkmıştır. Bu durum, SVM'nin yalnızca bireysel tahminlerde değil, ensemble tahminlerde de üstün bir performans sergilediğini göstermektedir.

RF, ensemble yöntemiyle makul bir performans sergilemiş, ancak SVM ve MLP'nin gerisinde kalmıştır. RF'nin başarısı, birden fazla karar ağacının tahminlerini birleştirerek oluşturduğu sağlamlık ve varyans azaltma yeteneğine dayanır. Ancak, dil modellerinden gelen karmaşık temsillerin işlenmesi sırasında RF'nin sınırlı kaldığı gözlemlenmektedir.

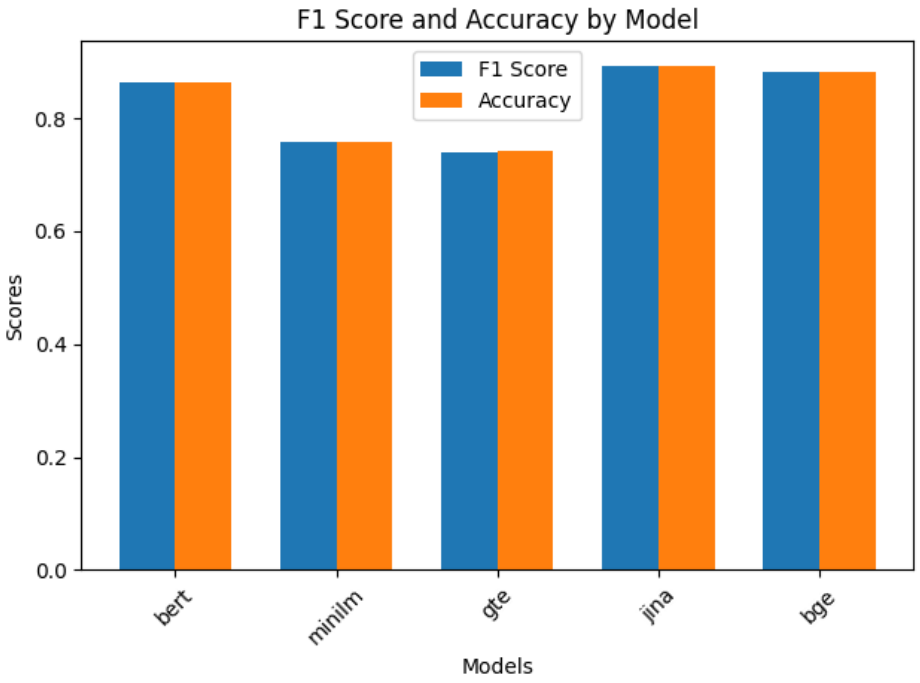
MLP, SVM'nin ardından ikinci sırada yer almış ve oldukça yüksek bir performans sergilemiştir. MLP'nin başarısı, doğrusal olmayan ilişkileri modelleyebilme kapasitesinden kaynaklanmaktadır. Çok katmanlı yapısıyla, dil modellerinden gelen temsillerdeki karmaşıklıkları öğrenmede etkili olmuştur. Ancak, MLP'nin eğitim süresi ve hesaplama maliyeti, SVM gibi daha hafif algoritmalara kıyasla bir dezavantaj oluşturabilir.

Tüm model sonuçlarının birleştirilmesiyle oluşan skor aşağıdaki tabloda verilmiştir.

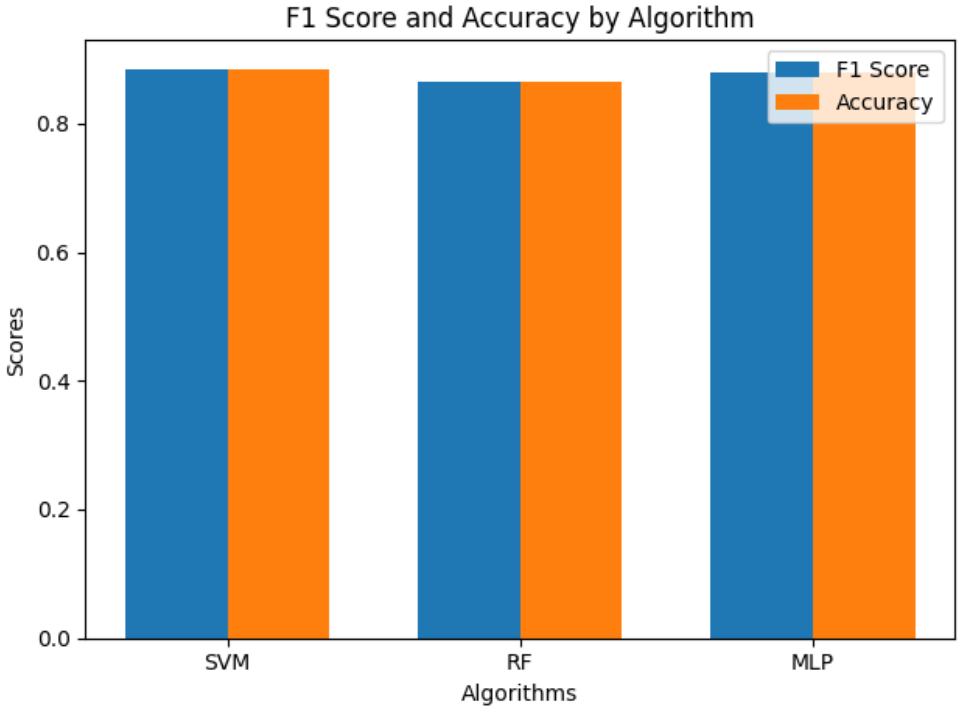
	F1	**Accuracy**
Final Ensemble	0.891989	0.892

Final ensemble yöntemi, F1 skoru 0.891 ve doğruluk değeri 0.892 ile diğer tüm bireysel modellerin ve algoritmaların yanı sıra temsil ve algoritma bazlı ensemble yöntemlerini de geride bırakmıştır. Bu yüksek performans, farklı model ve algoritmaların güçlü yönlerinin topluca değerlendirilmesinden kaynaklanmaktadır.

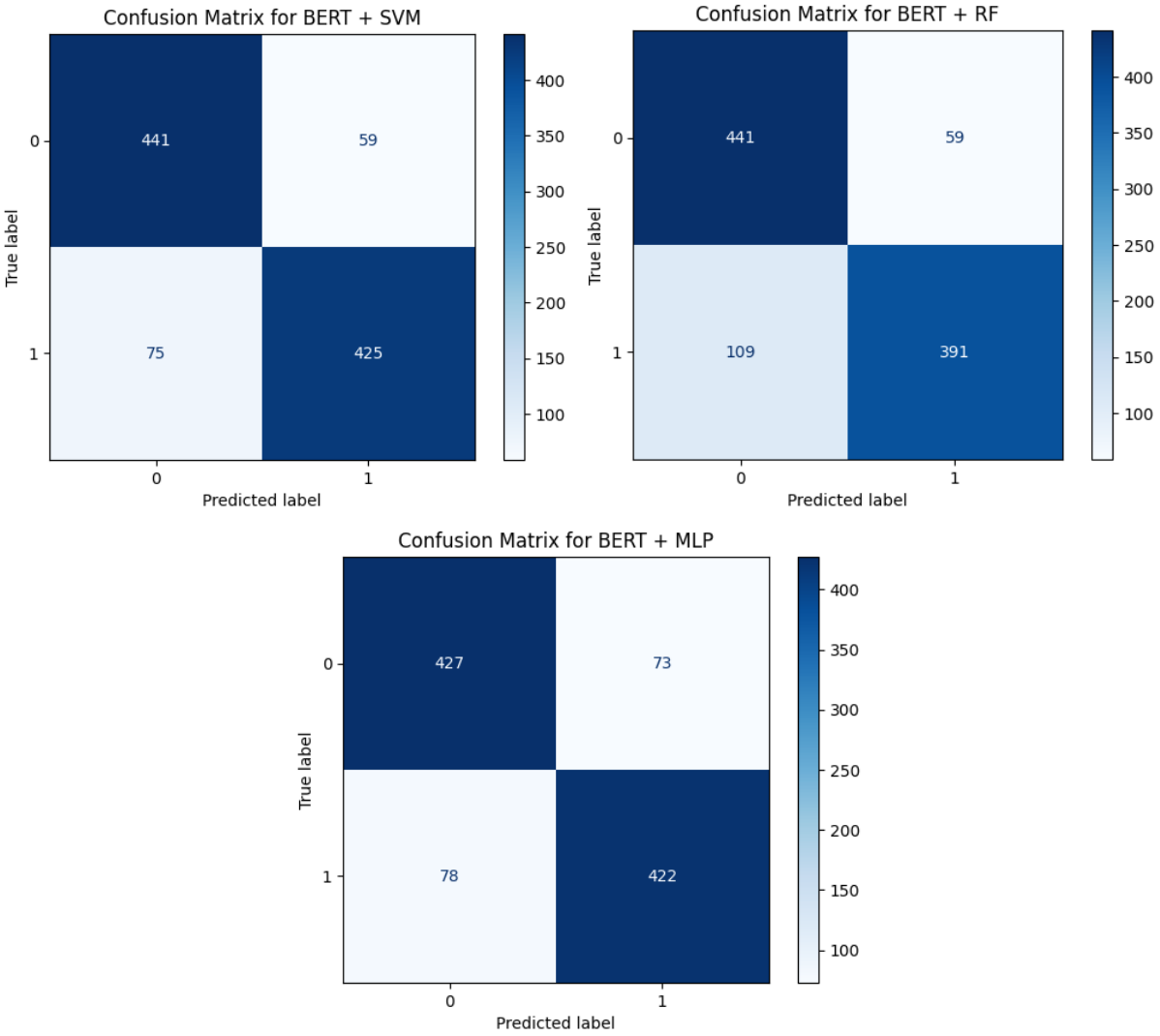
Dil modellerinin başarı skorlarını içeren bar grafiği aşağıdaki gibidir:



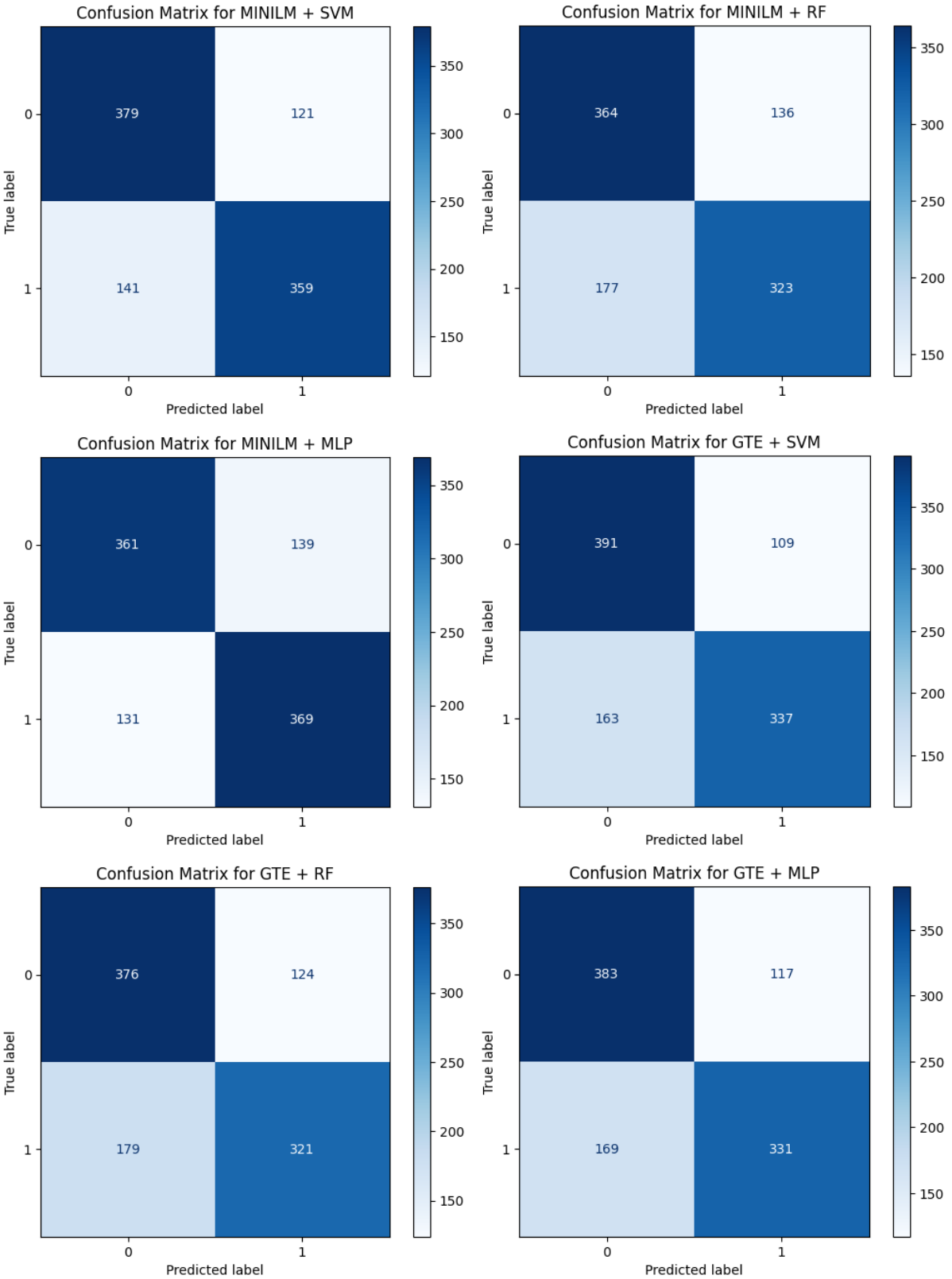
Algoritmaların başarı skorlarını içeren bar grafiği aşağıdaki gibidir:

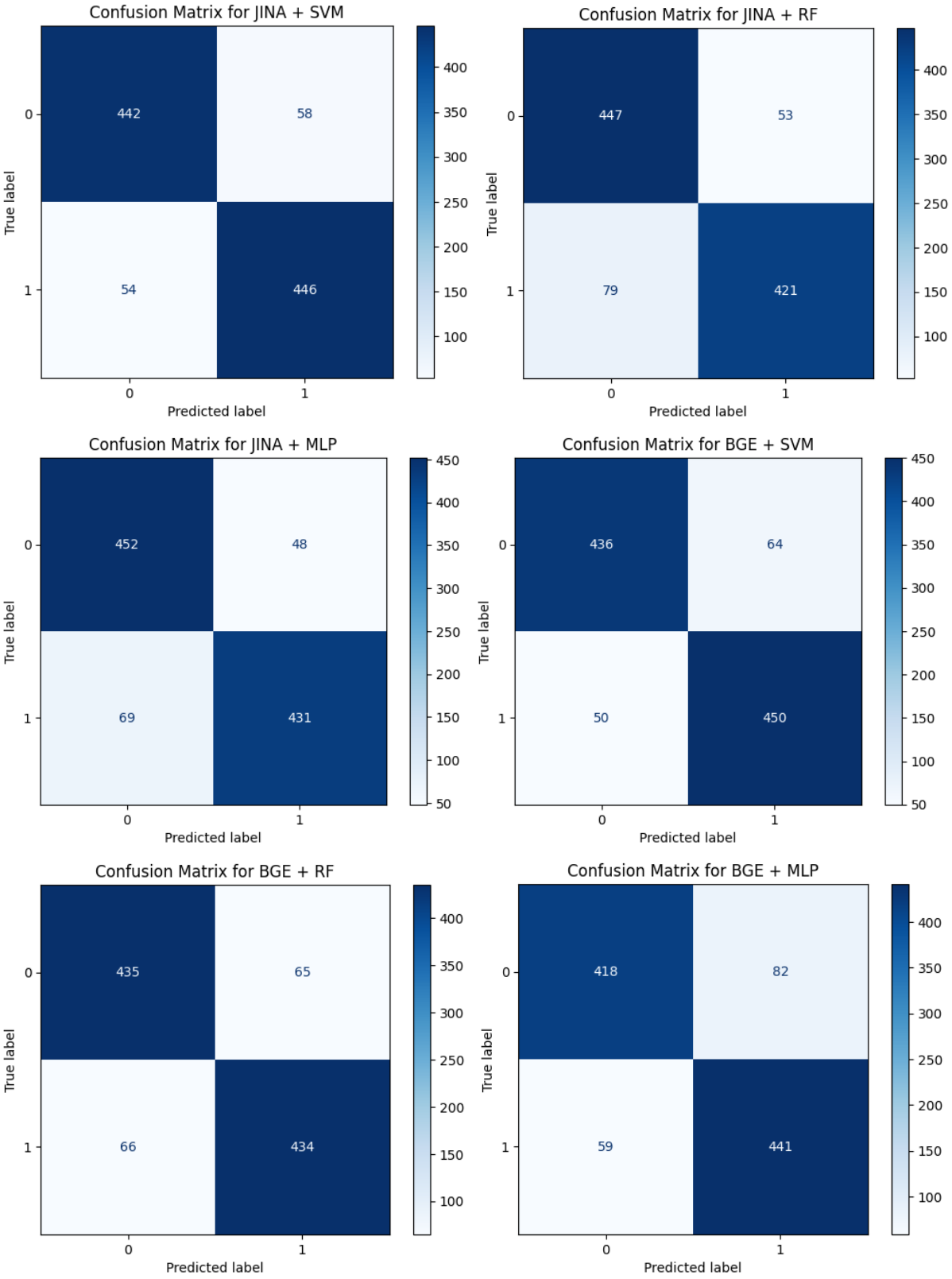


15 sonucun karmaşıklık matrisleri aşağıda verilmiştir.



Bu üç karmaşıklık matrisi, **BERT** modelinden elde edilen temsillerle üç farklı algoritmanın performansını değerlendirmek için kullanılmaktadır. SVM, pozitif (1) ve negatif (0) sınıflarını oldukça dengeli bir şekilde tahmin etmiştir. Yanlış negatif sayısının yanlış pozitiflerden biraz daha yüksek olduğu görülmektedir. Bu durum, modelin pozitif sınıfı kaçırmaya biraz daha eğilimli olduğunu gösterebilir. RF algoritması, doğru negatif tahminlerde SVM ile aynı başarıyı sağlamış olsa da yanlış negatif tahmin sayısında belirgin bir artış bulunmaktadır. Pozitif sınıfı tahmin etmede daha fazla hata yaptığı görülmektedir. MLP algoritması, doğru pozitif tahminlerde SVM'ye oldukça yakın bir performans göstermiştir. Yanlış pozitiflerin artması, MLP'nin negatif sınıfları yanlış pozitif olarak sınıflandırmaya biraz daha eğilimli olduğunu göstermektedir.


Confusion Matrix for MINILM + SVM
379
121
141
359
Confusion Matrix for MINILM + RF
364
136
177
323
Confusion Matrix for MINILM + MLP
361
139
131
369
Confusion Matrix for GTE + SVM
391
109
163
337
Confusion Matrix for GTE + RF
376
124
179
321
Confusion Matrix for GTE + MLP
383
117
169
331
True label
Predicted label
0
1
150
200
250
300
350


Confusion Matrix for JINA + SVM
442
58
54
446
True label
Predicted label
Confusion Matrix for JINA + RF
447
53
79
421
Confusion Matrix for JINA + MLP
452
48
69
431
Confusion Matrix for BGE + SVM
436
64
50
450
Confusion Matrix for BGE + RF
435
65
66
434
Confusion Matrix for BGE + MLP
418
82
59
441

4.1 Interpress News Category TR ile Çalışmalar

Veri setine ait rastgele seçilen bir gözlem aşağıdaki tabloda verilmiştir.

	Category	ID	PublishDateTime	Category Code	Content	Title
89029	magazin	15262458	2012-05-15T00:00:00Z	7	I MUHTEŞEM Yüzyıl sezon finaline doğru yaklaşırken dizinin Muhteşem Süleyman ı ünlü oyuncu Halit Ergenç eşi Bergüzar Korelle tatil alışverişi yaparken görüntülendi. Sezon finaliyle birlikte tatile çıkacak olan çift, birlikte yorgunluk atacak.	TATİL HAZIRLIĞI

Veri setindeki tüm satırlar içerisinden 10000 gözlem rastgele seçilmiştir. Bu verilerin %20'si test setinde kullanılmıştır.

	Train Set	Test Set
0	542	136
1	537	134
2	538	135
3	492	123
4	438	109
5	244	61
6	377	94
7	475	119
8	484	121
9	196	49
10	357	89
11	578	145
12	589	147
13	513	128
14	656	164
15	418	105
16	566	141
Toplam	8000	2000

Bir öncekinde olduğu gibi train ve test kümelerindeki her bir cümlenin temsili oluşturularak kaydedilmiştir ve modeller train kümesinin temsilleri ile eğitilmiştir. Ardından test temsilleri üzerinde tahmin yapılmıştır. Modellerin tekil başarıları aşağıdaki tabloda gösterilmiştir.

	F1	Accuracy
BERT + SVM	0.463191	0.4775
BERT + RF	0.419171	0.4350
BERT + MLP	0.440212	0.4380
MINILM + SVM	0.330297	0.3380
MINILM + RF	0.243147	0.2570
MINILM + MLP	0.303271	0.3035
GTE + SVM	0.389519	0.3975
GTE + RF	0.283247	0.2970
GTE + MLP	0.340231	0.3440
JINA + SVM	0.508183	0.5170
JINA + RF	0.389146	0.4030
JINA + MLP	0.443005	0.4440
BGE + SVM	0.464968	0.4790
BGE + RF	0.386000	0.4015
BGE + MLP	0.413068	0.4130

BERT modeli, bağlama duyarlı temsiller üretme kapasitesiyle bu veri setinde güçlü bir performans sergilemiştir. SVM ile kombinasyonu BERT'in en iyi sonucunu sağlamıştır. RF ve MLP algoritmaları ile elde edilen performanslar ise nispeten daha düşük kalmıştır. RF'nin doğrusal olmayan kararlar üretme kapasitesinin sınırlı olması ve MLP'nin çok katmanlı yapısına rağmen karmaşık sınıflar arasında yeterince ayrım yapamaması bu düşüşün olası nedenlerindendir.

MiniLM modeli, daha kompakt bir yapıya ve düşük boyutlu temsillere sahip olduğu için bu karmaşık veri setinde performans açısından zorlanmıştır. Özellikle RF algoritmasıyla kombinasyonunda elde edilen F1 skoru diğer kombinasyonlara kıyasla oldukça düşüktür. Bu, MiniLM'in daha büyük ölçekli dil modelleriyle karşılaştırıldığında bağlama duyarlılığının sınırlı olduğunu ve karmaşık sınıflar için yeterli temsiller üretemediğini göstermektedir. SVM, MiniLM ile diğer algoritmalara kıyasla daha iyi sonuçlar sağlamış olsa da genel performans, diğer modellerin gerisinde kalmıştır.

GTE modeli, geniş kapsamlı ve çok aşamalı kontrastif öğrenimle eğitilmiş olmasına rağmen, bu veri setinde beklenen performansı tam olarak sağlayamamıştır. SVM ile kombinasyonu, modelin en iyi sonucunu üretmiş ancak bu değerler BERT, Jina ve BGE modellerine kıyasla düşük kalmıştır. RF ve MLP algoritmalarıyla kombinasyonunda ise performans daha da düşerek modelin sınıf ayırma kapasitesinin sınırlı olduğunu ortaya koymuştur.

Jina modeli, bu veri setinde en yüksek performansı sergileyen model olmuştur. SVM ile kombinasyonu, F1 skoru açısından en iyi sonuçları sağlamıştır. Bu sonuç, Jina'nın bağlama duyarlılığı yüksek temsiller üretebilme kapasitesini ve SVM'nin bu temsilleri etkili bir şekilde kullanabilmesini göstermektedir. RF ve MLP ile kombinasyonlarında ise performans bir miktar düşmüş olsa da, Jina'nın genel olarak diğer modellere kıyasla daha etkili olduğu açıktır.

BGE modeli, Jina'dan sonra en iyi performansı sergileyen modeldir. SVM ile kombinasyonu, güçlü bir sonuç elde etmiştir. Bu durum, BGE'nin bağlama duyarlılığını iyi bir şekilde yakalayabildiğini ve SVM'nin bu temsilleri etkili bir şekilde işlediğini göstermektedir. Ancak, RF ve MLP algoritmalarıyla kombinasyonunda performans düşüşü gözlemlenmiştir. Bu durum, BGE'nin etkili temsiller üretmesine rağmen, algoritma seçiminde SVM'nin diğer seçeneklere kıyasla daha avantajlı olduğunu ortaya koymaktadır.

Temsillerin birleştirilmesiyle oluşan ensemble başarılar aşağıdaki tabloda gösterilmiştir.

	F1	**Accuracy**
BERT	0.485303	0.497
MINILM	0.359738	0.366
GTE	0.409978	0.42
JINA	0.494634	0.505
BGE	0.460108	0.473

BERT temsilleriyle elde edilen ensemble sonuçları güçlü bir performans sergilemiştir. Bu sonuç, BERT'in bağlama duyarlı temsiller üreterek farklı algoritmaların bu temsillerden etkin bir şekilde faydalanmasını sağladığını göstermektedir. Farklı algoritmaların birleştirilmesiyle de daha iyi bir sonuç oluşturmuştur.

MiniLM, ensemble yönteminde düşük performans sergilemiş ve diğer modellere kıyasla en alt sırada yer almıştır. Ensemble yöntemi, MiniLM'in bağlama duyarlılığı eksikliğini tamamen telafi edememiştir. Bu sonuç, MiniLM'in daha az karmaşık veya daha az sınıf içeren veri setlerinde daha etkili olabileceğini işaret etmektedir.

GTE modeli, ensemble yöntemiyle orta seviyede bir başarı elde etmiş ve BERT'in gerisinde, ancak MiniLM'in üzerinde bir performans sergilemiştir. Bu durum, GTE'nin çok aşamalı kontrastif öğrenimle eğitilmesine rağmen bağlam farklılıklarına tam anlamıyla adapte olamadığını göstermektedir. Ancak, ensemble yöntemi, GTE'nin bireysel algoritmalardan elde ettiği sonuçlara göre bir iyileşme sağlamış ve farklı algoritmaların tahminlerinin birleştirilmesiyle daha dengeli bir sonuç elde edilmiştir.

JINA, ensemble yöntemiyle en yüksek başarıyı sergileyen model olmuştur. Bu sonuç, Jina'nın bağlama duyarlılığı yüksek temsiller üreterek farklı algoritmaların bu temsilleri verimli bir şekilde kullanmasını sağladığını göstermektedir. Jina'nın başarısı, özellikle karmaşık sınıflar arasında doğru tahminler yapabilme kapasitesini ve bu temsillerin ensemble yönteminde daha kararlı sonuçlar ürettiğini ortaya koymaktadır.

BGE temsilleri, Jina'nın ardından en iyi ensemble performansını sağlamıştır. Bu sonuç, BGE'nin bağlama duyarlılığı yüksek temsiller üretebildiğini ve bu temsillerin farklı algoritmalarla birleştirilmesinde etkili olduğunu göstermektedir. Ancak, Jina ile kıyaslandığında performans farkı, BGE'nin karmaşık sınıflar arasında daha fazla hata yaptığını düşündürmektedir. Bu, BGE'nin bağlam zenginliği gerektiren bu tür sınıflandırma görevlerinde Jina'nın gerisinde kalmasına neden olabilir.

Algoritmaların birleştirilmesiyle oluşan ensemble başarılar aşağıdaki tabloda gösterilmiştir.

	F1	**Accuracy**
SVM	0.488679	0.502
RF	0.408421	0.43
MLP	0.483303	0.49

SVM algoritması algoritma bazlı ensemble yöntemlerinde en yüksek performansı göstermiştir. Bu sonuç, SVM'nin özellikle bağlama duyarlılığı yüksek temsillerle etkili bir şekilde çalıştığını ve sınıflar arasında net ayrımlar yapabildiğini göstermektedir. Ensemble yöntemi, SVM'nin tek

başına elde ettiği performansı bir adım daha ileriye taşımış ve algoritmanın genel doğruluğunu artırmıştır.

RF algoritması, algoritma birleştirme yöntemiyle diğer algoritmalara kıyasla daha düşük bir performans sergilemiş ve sınıflandırma görevinde sınırlı bir etki göstermiştir. Bu durum, RF'nin bağlama duyarlılığı gerektiren karmaşık embeddinglerle sınıf ayrımında zorlanabileceğini göstermektedir.

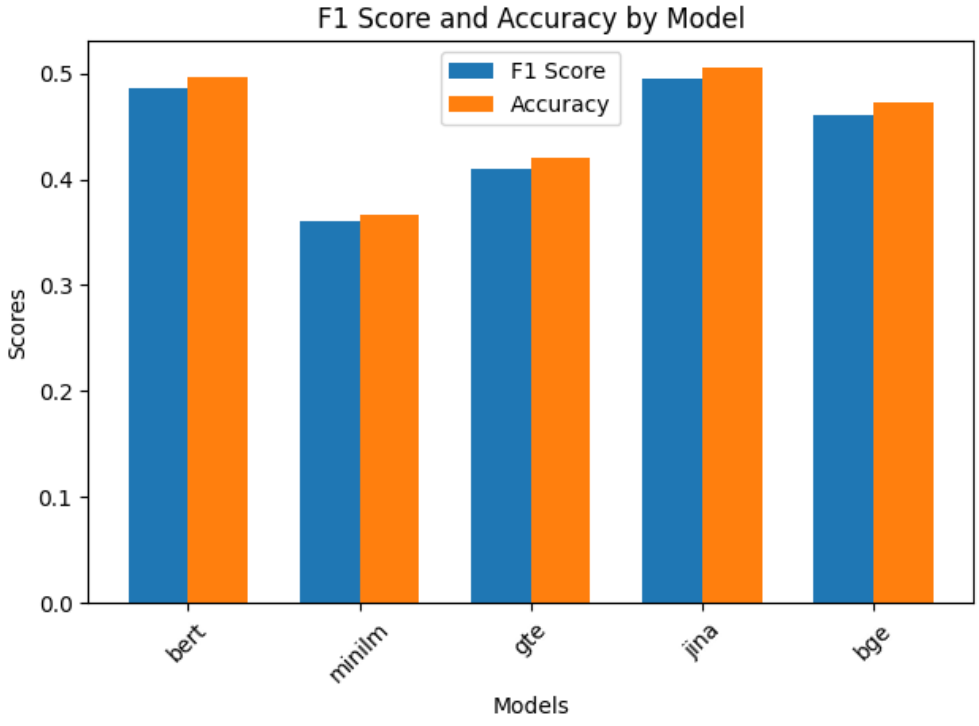
MLP algoritması, SVM'ye yakın bir performans sergilemiş, ancak bir adım geride kalmıştır. Çok katmanlı bir yapıya sahip olan MLP, doğrusal olmayan ilişkileri öğrenme kapasitesine sahip olsa da, veri setindeki karmaşık bağlamları SVM kadar etkili bir şekilde değerlendirememiştir.

Tüm model sonuçlarının birleştirilmesiyle oluşan skor aşağıdaki tabloda verilmiştir.

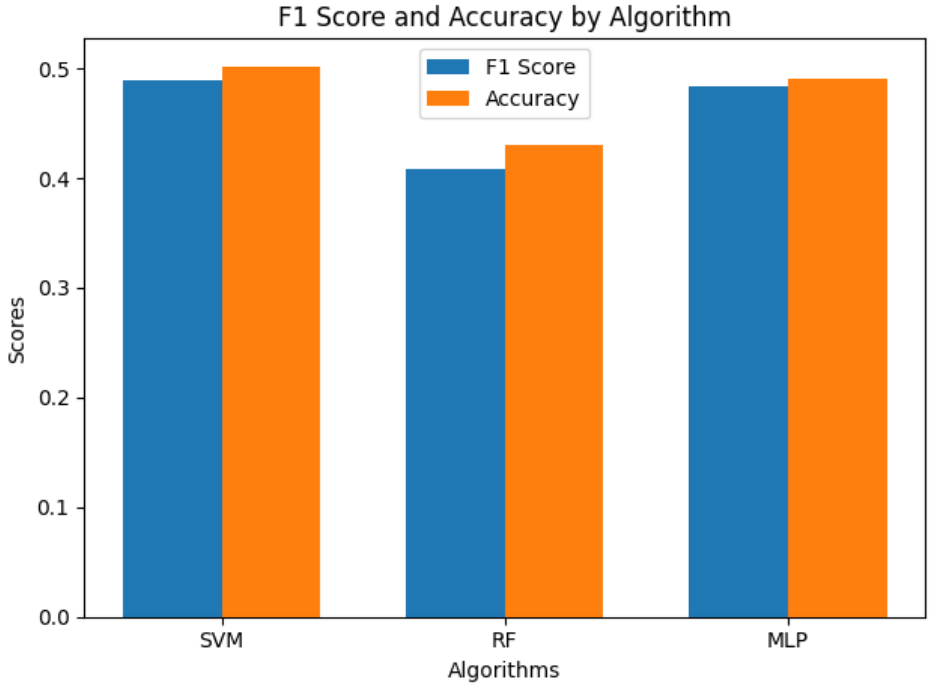
	F1	**Accuracy**
Final Ensemble	0.4943503742970672	0.51

Final ensemble yöntemiyle elde edilen F1 skoru veri setinin karmaşıklığı ve sınıf dağılımındaki dengesizlik göz önüne alındığında, oldukça makul bir performans sergilemiştir. Bu sonuçlar, ensemble yönteminin hem temsil hem de algoritma düzeyindeki bireysel eksiklikleri dengelemeyi başardığını ve genel olarak daha güvenilir bir sınıflandırma sistemi oluşturduğunu göstermektedir. Ancak, bu başarı birleştirilen modellerin bireysel performanslarının üstüne yalnızca sınırlı bir ölçüde çıkmıştır, bu da bazı modellerin veya algoritmaların sonuçları dengeleme konusunda daha fazla katkıda bulunduğunu düşündürmektedir.

Dil modellerinin başarı skorlarını içeren bar grafiği aşağıdaki gibidir:



Algoritmaların başarı skorlarını içeren bar grafiği aşağıdaki gibidir:

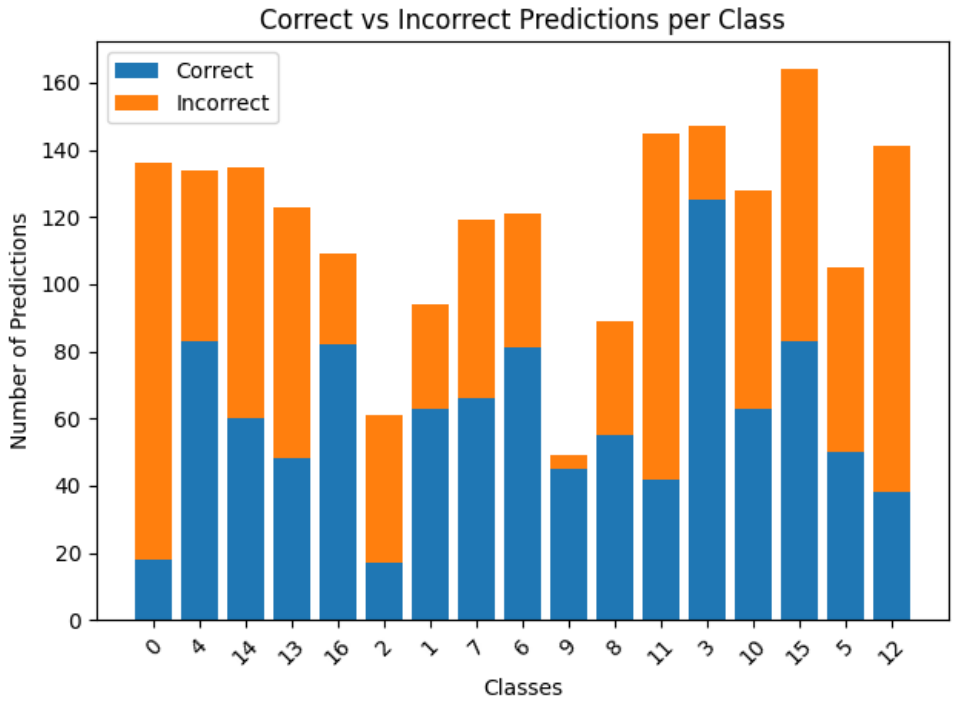


Aşağıdaki grafik, ilgili veri setinde final ensemble modelinin her bir sınıf için F1 skorlarını görselleştirmektedir. Grafik, sınıflar arasındaki performans farklılıklarını net bir şekilde ortaya koyarak modelin hangi sınıflarda daha iyi veya daha zayıf performans gösterdiğini göstermektedir.

Yüksek F1 skorları, modelin ilgili sınıflarda dengeli bir performans sergilediğini gösterirken, düşük F1 skorları modelin sınıf tahminlerinde daha fazla hata yaptığını işaret eder. Örneğin, 9 numaralı sınıfın 0.94 ile en yüksek F1 skoruna sahip olması, modelin bu sınıfta çok başarılı

olduğunu gösterirken, 0 numaralı sınıfın 0.17 gibi düşük bir F1 skoru alması, modelin bu sınıfta ciddi performans sorunları yaşadığını ifade eder.

Aşağıdaki grafikte de benzer şekilde modelin her sınıf için doğru (Correct) ve yanlış (Incorrect) tahminlerini karşılaştırmalı olarak göstermektedir.



Sonuç olarak, bu çalışma, iki farklı Türkçe veri seti üzerinde çeşitli dil modelleri ve makine öğrenimi algoritmalarını kullanarak metin sınıflandırma görevlerinde temsillerin ve kararların birleştirilmesinin etkisini değerlendirmiştir.

Elde edilen sonuçlar, temsillerin ve algoritmaların birleştirilmesinin sınıflandırma performansına olumlu katkı sağladığını göstermiştir. Özellikle bağlama duyarlılığı yüksek temsiller (örneğin JINA ve BGE modelleri) ve SVM algoritması, bireysel ve ensemble performanslarında en başarılı yöntemler olarak öne çıkmıştır. Ancak, veri setlerindeki sınıf dengesizliği ve bazı temsillerin bağlam karmaşıklığını yeterince yansıtamaması gibi faktörler, model performansında belirli sınırlamalara yol açmıştır.

4. KAYNAKLAR

[1] https://huggingface.co/dbmdz/bert-base-turkish-uncased
[2] https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2
[3] https://huggingface.co/thenlper/gte-large
[4] https://huggingface.co/jinaai/jina-embeddings-v3
[5] https://huggingface.co/BAAI/bge-m3
[6] https://huggingface.co/datasets/fthbrmnby/turkish_product_reviews
[7] https://huggingface.co/datasets/yavuzkomecoglu/interpress_news_category_tr
[8] Z. Li, X. Zhang, Y. Zhang, D. Long, P. Xie, and M. Zhang, Towards general text embeddings with multi-stage contrastive learning, 2023. arXiv: 2308 . 03281[cs.CL]. [Online]. Available: https://arxiv.org/abs/2308.03281.
[9] https://www.geeksforgeeks.org/support-vector-machine-algorithm/
[10] https://www.ibm.com/topics/random-forest
[11] https://www.datacamp.com/tutorial/multilayer-perceptrons-in-machine-learning