BLM5109 3. ÖDEV

RETRIEVAL ENSEMBLE

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

1. GİRİŞ

Bu çalışma, doğal dil işleme (NLP) alanında farklı temsil modellerinin soru-cevap görevindeki performanslarını değerlendirmek ve bu modellerin sonuçlarını birleştirerek daha başarılı cevaplar üretebilmek amacıyla gerçekleştirilmiştir. Bu amaç doğrultusunda, Türkçe dilinde soru-cevap çiftlerinden oluşan iki ayrı veri kümesi kullanılarak beş farklı modelden embedding'ler oluşturulmuş ve her model için Top1 ve Top5 cevaplar belirlenmiştir. Ayrıca, bu sonuçları birleştirmek için farklı ensemble yöntemleri uygulanmış ve her bir yöntemin başarı oranı incelenmiştir. Çalışma, modellerin ve ensemble yöntemlerinin performanslarını grafiklerle karşılaştırarak, toplu öğrenme yöntemlerinin bireysel modellerden daha etkili sonuçlar üretebileceğini göstermeyi amaçlamaktadır.

2. VERİ SETLERİ

Bu çalışmada kullanılan ilk veri kümesi, **Turkish Instructions Dataset** adlı Türkçe dilinde hazırlanmış bir soru-cevap veri kümesidir [1]. Veri kümesi, çeşitli talimatlar ve bu talimatlara verilen yanıtları içeren bir yapıya sahiptir. Veri kümesi aşağıdaki özelliklere sahiptir:

- **Talimat (instruction):** Kullanıcıdan gelen soru veya komut.
- **Çıktı (output):** Verilen talimata karşılık oluşturulmuş yanıt.
- **Giriş (giriş):** Talimatı destekleyen ek bir bilgi (bazı satırlarda bulunur).

Veri kümesi, toplamda 51,563 örnek içermektedir. Ancak, bu çalışmanın kapsamını ve hesaplama maliyetlerini optimize etmek amacıyla, rastgele örnekleme yöntemiyle veri kümesinden 2000 soru-cevap çifti seçilmiştir. Verinin ilk 5 satırı aşağıdaki gibidir.

	instruction	output
16540	Aşağıdaki cümleyi orijinal biçimine geri yükleyin: "Bu şimdiye kadar gördüğüm en iyi şey"	Bundan daha iyi bir şey görmedim.
33335	İşyerinde çeşitlilik sorunlarını ele almak için yaratıcı fikirler üretin	İşyerinde çeşitliliğe değinmeye yönelik fikirler, azınlıklar için bir çalışan kaynak grubu oluşturmayı, yöneticilere ve personele çeşitlilik ve kapsayıcılık farkındalığı eğitimi vermeyi, kimlik ve kültür etrafında ekip oluşturma faaliyetlerine ev sahipliği yapmayı, insanları kültürel beceri geliştirme atölyelerine dahil etmeyi ve etkinlikler düzenlemeyi içerebilir. ve çeşitliliği kutlayan etkinlikler.
34956	Yunan mitolojisindeki 'Aşil' karakterini kısaca tanımlayın.	Aşil, Yunan mitolojisinde Truva Savaşı'nın bir kahramanıydı. Nereid Thetis ile Myrmidonların kralı Peleus'un oğluydu. Yenilmez bir zırh giyen Yunan ordusunun en büyük savaşçısıydı. Hem arkadaşları hem de düşmanları ondan korkuyordu ve sonunda savunmasız topuğuna bir okla öldürüldü. Cesareti ve cesareti ile hatırlanır.
5733	Bir arının yaptığı üç şeyi söyleyin.	Arılar önemli tozlayıcılardır ve ekosistemde önemli bir rol oynarlar. Bal yapmak, nektar ve polen toplamak, çiçekleri tozlamak gibi pek çok önemli görevleri vardır. Sosyal böceklerdir ve bal peteği hücrelerini, kraliçe arıyı ve larvalarını oluşturmak için bir kovanda birlikte çalışırlar. Arılar ayrıca bal peteklerini yapmak için kendi balmumunu yaratırlar. Son olarak, arılar, yiyecek kaynaklarının yeri hakkında bilgi aktarmak için kendi "dans etme" biçimlerini kullanarak diğer arılarla iletişim kurabilirler.
26934	Aşağıdaki molekülün fonksiyonel grubunu belirleyiniz. CH3-CH2-COOH	CH3-CH2-COOH molekülünün fonksiyonel grubu bir karboksilik asittir.

Veri ön işleme aşamasında, talimat ve giriş bilgileri birleştirilerek tek bir talimat sütunu oluşturulmuş ve gereksiz sütunlar temizlenmiştir. Ayrıca, tüm talimat ve yanıt metinlerinden gereksiz boşluklar kaldırılmıştır. Bu işlemler sonucunda, analiz için temiz ve düzenli bir veri seti hazırlanmıştır.

Çalışmada kullanılan diğer veri seti, 2023–2024 Yapay Zeka lisans dersinde öğrencilerin katkılarıyla oluşturulmuş ve toplamda 16.730 soru-cevap çiftinden oluşan özel bir veri setidir [2]. Veri seti, aşağıdaki sütunlardan oluşmaktadır:

- **Soru:** Kullanıcı tarafından yöneltilen talimat veya soru.
- **Makine Cevabı:** Yapay zeka modeli tarafından oluşturulan cevap.
- **İnsan Cevabı:** İnsan tarafından yazılmış cevap.
- **Tercih:** Makine cevabının mı yoksa insan cevabının mı seçileceğini belirten bir değer (1, 2 veya 3).

Bu veri setinde, tercih sütunundaki değerler aşağıdaki şekilde değerlendirilmiştir:

- **1:** İnsan cevabının seçilmesi gerektiğini ifade eder.
- **2:** Makine cevabının seçilmesi gerektiğini ifade eder.
- **3:** Her iki sonucun da başarılı olduğunu ifade eder. Çalışmada '3' için insan cevabı tercih edilmiştir.

Bu kurallar doğrultusunda, toplamda 2.000 soru-cevap çifti seçilmiştir. Seçim sürecinde, her bir soru için tercih sütunu dikkate alınarak ilgili cevap (insan veya makine) veri kümesine dahil edilmiştir. Veriye ait ilk üç satır aşağıdaki gibidir.

	instruction	output
1061	Ekmeğin etten daha çok protein içermesinin sebebi nedir	Ekmek karbonhidrat zengini bir yiyecektir fakat protein içermez ve et protein açısından daha zengindir.
12545	İlaçsız tedavi var mıdır?	Evet, ilaçsız tedavi mümkündür. Aslında, birçok sağlık sorunu için ilaç kullanmadan tedavi olmak mümkündür. İlaçsız tedavinin birçok farklı yöntemi vardır ve her yöntemin kendine özgü avantajları ve dezavantajları vardır. İlaçsız tedavinin bazı yaygın türleri şunlardır: Akupunktur: Akupunktur, vücudun belirli noktalarına ince iğneler batırarak yapılan bir Çin tıbbı yöntemidir. Ağrı, kaygı ve depresyon gibi çeşitli sağlık sorunlarını tedavi etmek için kullanılabilir. Masaj: Masaj, kasları gevşetmek, stresi azaltmak ve ağrıyı hafifletmek için kullanılabilir. Yoga: Yoga, stresi azaltmak, zihinsel ve fiziksel sağlığı iyileştirmek için kullanılabilir. Meditasyon: Meditasyon, stresi azaltmak, kaygıyı yönetmek ve zihinsel netliği iyileştirmek için kullanılabilir. Diyet ve beslenme: Sağlıklı bir diyet ve beslenme, birçok sağlık sorununu önlemeye ve tedavi etmeye yardımcı olabilir. Egzersiz: Düzenli egzersiz yapmak, birçok sağlık sorununu önlemeye ve tedavi etmeye yardımcı olabilir. Fizik tedavi: Fizik tedavi, yaralanmalardan veya hastalıklardan sonra iyileşmeyi hızlandırmak için kullanılabilir. İlaçsız tedavinin sizin için uygun olup olmadığını doktorunuzla konuşmanız önemlidir. Doktorunuz, sağlık sorununuzun ciddiyetini ve sizin için en iyi tedavi seçeneklerini değerlendirebilir. İlaçsız tedavinin bazı potansiyel faydaları şunlardır: Yan etki riski daha düşüktür. Daha doğal bir tedavi yöntemidir. Bazı durumlarda ilaçlardan daha etkili olabilir. İlaçsız tedavinin bazı potansiyel dezavantajları şunlardır: Her zaman etkili olmayabilir. İlaçlardan daha uzun sürebilir. Daha pahalı olabilir.
7472	Einstein 20 sene daha yaşasa ne olurdu?	Einstein, 20 yıl daha yaşasaydı, bilim dünyasına ve insanlığa pek çok katkıda bulunma potansiyeline sahip olabilirdi. Bu süre içinde, genel görelilik ve kuantum mekaniği gibi alanlardaki çalışmalarını derinleştirebilir ve belki de yeni teoriler geliştirebilirdi. Ayrıca, uzay-zamanın doğası, kara delikler, karanlık madde ve enerji gibi bilinmeyen alanlarda daha fazla araştırma yapabilirdi. Bununla birlikte, Einstein'ın ilgisinin sadece bilimsel keşiflerle sınırlı olmayabileceğini düşünmek mantıklı olurdu. Örneğin, siyasi ve sosyal meselelerle ilgili daha fazla aktiflik gösterebilir, barış ve insan hakları için çalışabilir veya eğitim alanında reformlar önerip bilimsel bilginin yayılmasını teşvik edebilirdi. Sonuç olarak, Einstein'ın 20 yıl daha yaşaması, bilim, sosyal adalet ve insanlık için pek çok potansiyel fayda sağlayabilirdi, ancak kesin olarak ne yapacağı belirsizdir, çünkü bu tür spekülasyonlar spekülatif ve belirsizdir.

3. KULLANILAN MODELLER

Bu çalışmada aşağıdaki temsil modelleri kullanılmıştır:

1. **Cosmos-ColBERT**: ytu-ce-cosmos/turkish-colbert
2. **MiniLM**: sentence-transformers/all-MiniLM-L12-v2
3. **BGE**: BAAI/bge-m3
4. **Jina Embeddings**: jinaai/jina-embeddings-v3
5. **GTE-Large**: thenlper/gte-large

Cosmos-ColBERT [3], Yıldız Teknik Üniversitesi tarafından Türkçe doğal dil işleme görevleri için optimize edilmiş bir modeldir. ColBERT (Contextualized Late Interaction over BERT) mimarisine dayanır ve Türkçe veri üzerinde ince ayar yapılmıştır. Model, özellikle çift aşamalı doğrulama ve bilgi erişimi görevlerinde başarılıdır. ColBERT'in temel özelliği, metinlerin bağlamsal temsilini ince taneli bir şekilde karşılaştırmasına olanak tanıyan geç etkileşim (late interaction) mekanizmasıdır. Bu mekanizma, metin parçalarını bağımsız olarak işlemesine rağmen, bağlamsal bilginin etkili bir şekilde kullanılmasını sağlar.

MiniLM [4], 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.

BGE [5], 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.

Jina [6], 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.

GTE [7], 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.

4. ENSEMBLE YÖNTEMLER

Ensemble yöntemler, birden fazla modelin çıktısını birleştirerek daha güçlü ve genel geçer sonuçlar elde etmeyi amaçlar. Bu çalışmada üç farklı ensemble yöntemi kullanılmıştır: Majority Voting, Borda Count ve Mean Ensemble. Her bir yöntem aşağıda detaylandırılmıştır.

4.1 Majority Voting

Majority Voting, ensemble yöntemlerinin en temel ve yaygın kullanılanlarından biridir. Bu yöntemde, her modelin birinci tercihi (Top1) bir oy olarak değerlendirilir. Çoğunluğun oyunu alan cümle **ensemble Top1** olarak seçilir. **Top5** için ise tüm modellerin Top5 sonuçları bir araya getirilir ve en sık tekrar eden 5 cümle **ensemble Top5** olarak belirlenir. Eşitlik durumunda, benzerlik skorları kullanılarak en yüksek puana sahip cümle seçilir.

4.2 Borda Count

Borda Count, sıralama bilgilerini dikkate alan daha gelişmiş bir ensemble yöntemidir. Her modelin sıralama sonucu, **Borda puanlama sistemi** ile değerlendirilir. Bu yöntemde:

- **Top1 için:** Her modelin Top1 sonucu puan alır ve tüm modellerin puanları toplanır. En yüksek puanı alan cümle **ensemble Top1** olur.
- **Top5 için:** Her modelin Top5 sonuçlarına sıralama bazlı puanlar atanır (örneğin, 1. sıraya 5 puan, 2. sıraya 4 puan vb.). Tüm modellerin puanları birleştirilir ve toplam puanı en yüksek olan 5 cümle **ensemble Top5** olarak seçilir.

4.3 Mean Ensemble

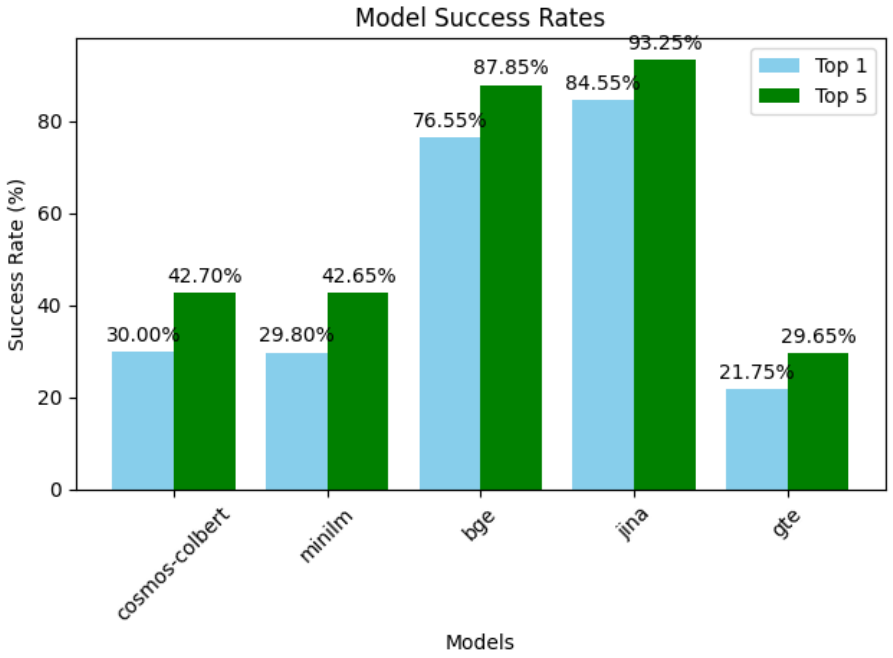
Mean Ensemble, benzerlik skorlarının ortalamalarını kullanarak ensemble sonuçlarını belirler. Bu yöntemde:

- **Top1 için:** Her modelin Top1 sonucu için benzerlik skorları alınır ve bu skorların ortalaması hesaplanır. En yüksek ortalama skora sahip cümle **ensemble Top1** olarak seçilir.
- **Top5 için:** Her modelin Top5 sonuçlarında yer alan cümleler bir araya getirilir. Aynı cümle birden fazla modelde bulunuyorsa, benzerlik skorlarının ortalaması alınır. Ortalaması en yüksek olan 5 cümle **ensemble Top5** olarak belirlenir.

5. DENEYSEL SONUÇLAR

5.1 İlk Veri Seti İçin Sonuçlar

Modellerin tekil başarıları aşağıdaki gibidir.



Cosmos-ColBERT modeli, bağlamsal bilgiyi yakalama konusunda sınırlı bir başarı göstermiştir. Top1 başarı oranı %30 iken, Top5 başarı oranı %42.70 olarak kaydedilmiştir. Bu sonuçlar, modelin doğru cevabı Top5 içine dahil etme ihtimalinin daha yüksek olduğunu ancak Top1 tahminlerinde zayıf kaldığını göstermektedir. Bu durum, modelin daha çok bağlam bağımlı bilgilerde sınırlı performans göstermesiyle açıklanabilir.

MiniLM modeli de benzer şekilde sınırlı bir performans sergilemiştir. Top1 başarı oranı %29.80, Top5 başarı oranı ise %42.65 olarak hesaplanmıştır. MiniLM, hafif bir model olduğundan, doğruluk oranının düşük kalması beklenebilir. Bu modelin hesaplama maliyetinin düşük olması, performans kaybının bir bedeli olarak değerlendirilebilir.

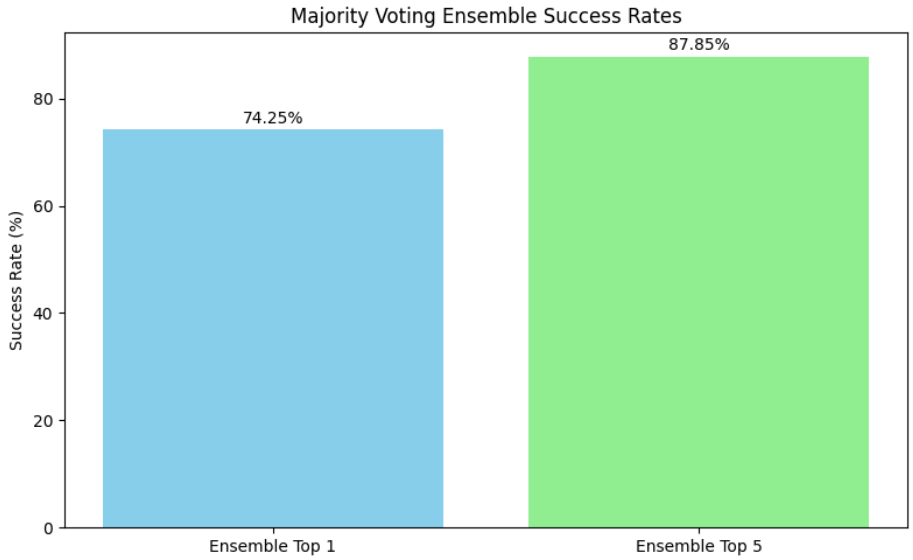
BGE modeli, bağlamsal anlamları yakalama konusunda oldukça başarılı sonuçlar sunmuştur. Top1 başarı oranı %76.55, Top5 başarı oranı ise %87.85 olarak hesaplanmıştır. Bu, modelin semantik uyumu güçlü bir şekilde temsil edebildiğini ve doğru cevabı Top5 içerisine dahil etme konusunda güvenilir olduğunu göstermektedir.

Jina Embeddings, hem Top1 hem de Top5 başarı oranlarında en iyi performansı sergilemiştir. Top1 başarı oranı %84.55, Top5 başarı oranı ise %93.25 olarak ölçülmüştür. Bu sonuçlar, Jina'nın bağlamsal bilgiyi etkili bir şekilde yakalayabildiğini ve doğru cevabı yüksek doğrulukla tahmin edebildiğini göstermektedir. Model, bu tür semantik eşleştirme görevleri için en güvenilir seçenek olarak dikkat çekmektedir.

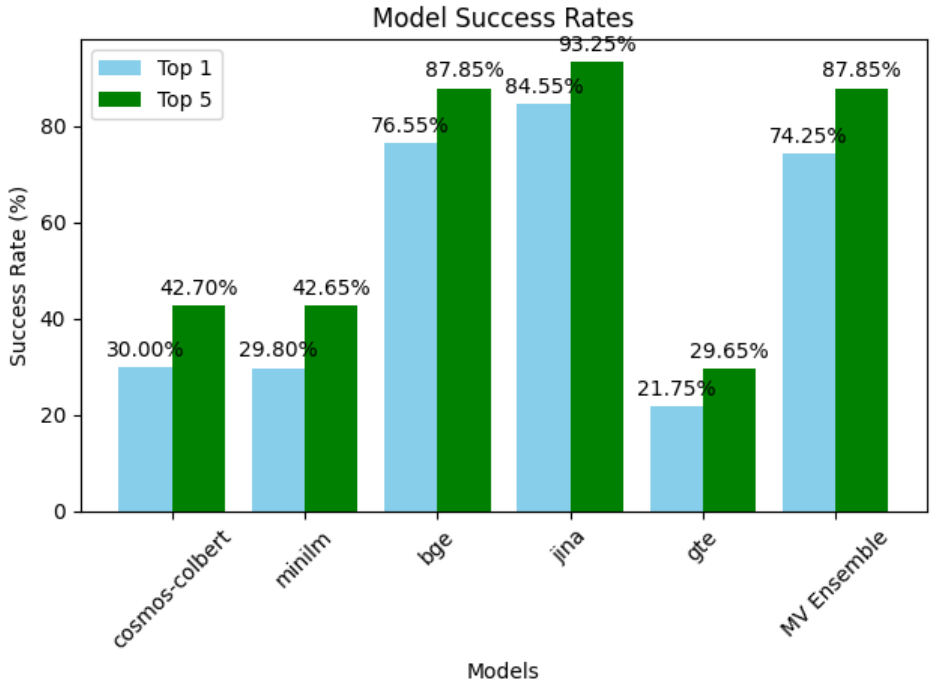
GTE-Large modeli ise diğer modellere kıyasla en düşük başarı oranlarına sahiptir. Top1 başarı oranı %21.75, Top5 başarı oranı ise %29.65 olarak belirlenmiştir. Bu sonuçlar, modelin bağlamsal bilgiyi yakalama yeteneğinin zayıf olduğunu ve doğru cevabı belirlemede sınırlı bir performans sunduğunu ortaya koymaktadır.

Genel olarak, Jina Embeddings ve BGE modelleri, semantik uyum ve bağlamsal anlam çıkarma konusunda en başarılı sonuçları vermiştir. Cosmos-ColBERT ve MiniLM modelleri orta düzey performans gösterirken, GTE-Large modeli bu görev için uygun bir seçenek olarak değerlendirilmemiştir.

Majority Voting metodunu kullanarak yapılan ensemble sonucundaki performans aşağıdaki gibidir.

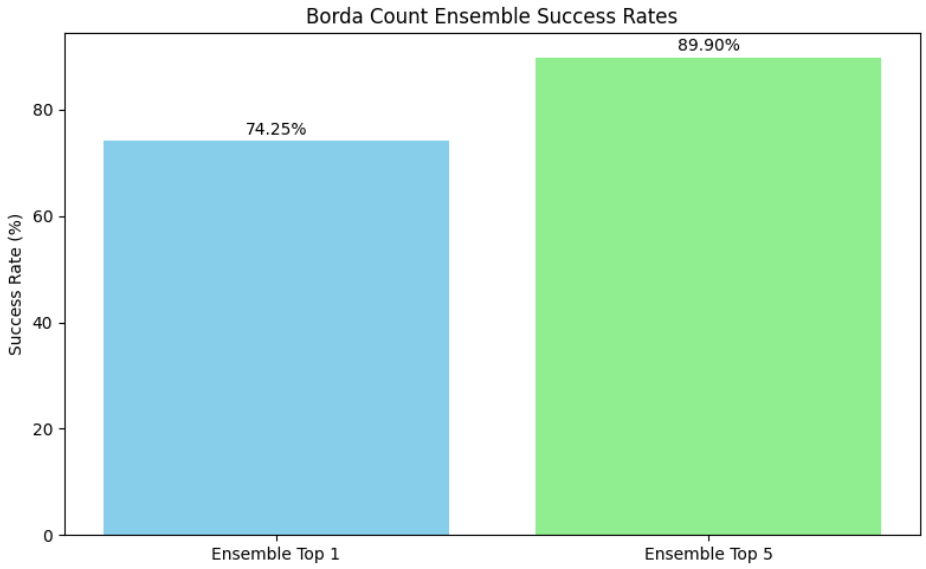


Aşağıdaki grafikte, Majority Voting (MV Ensemble) yöntemi, bireysel modellerin performansını birleştirerek elde edilen Top1 ve Top5 başarı oranlarını göstermektedir.

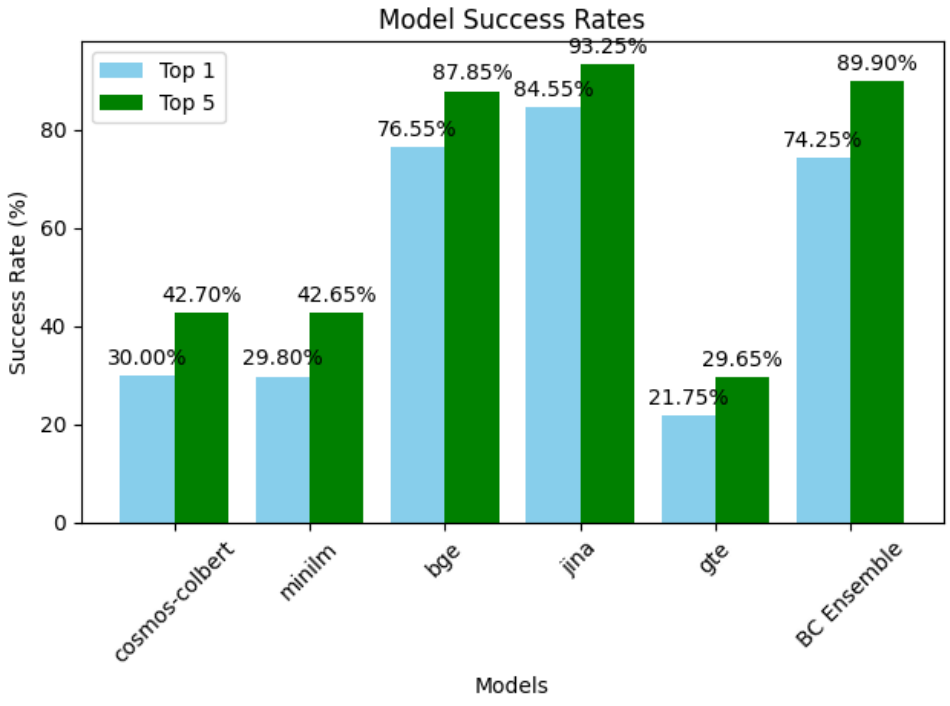


MV Ensemble yöntemi, bireysel modellerin güçlü yönlerini birleştirerek Top1 için %74.25, Top5 için ise %87.85 başarı oranı elde etmiştir. Bu sonuçlar, ensemble yöntemiyle bireysel modellerin performansının üzerinde bir başarı sağlandığını göstermektedir. Özellikle Cosmos-ColBERT, MiniLM ve GTE gibi düşük performanslı modellerin etkisinin dengelendiği ve BGE ile Jina Embeddings gibi yüksek performanslı modellerin sonuçlarına yaklaşıldığı gözlemlenmiştir.

Borda Count metodunu kullanarak yapılan ensemble sonucundaki performans aşağıdaki gibidir.

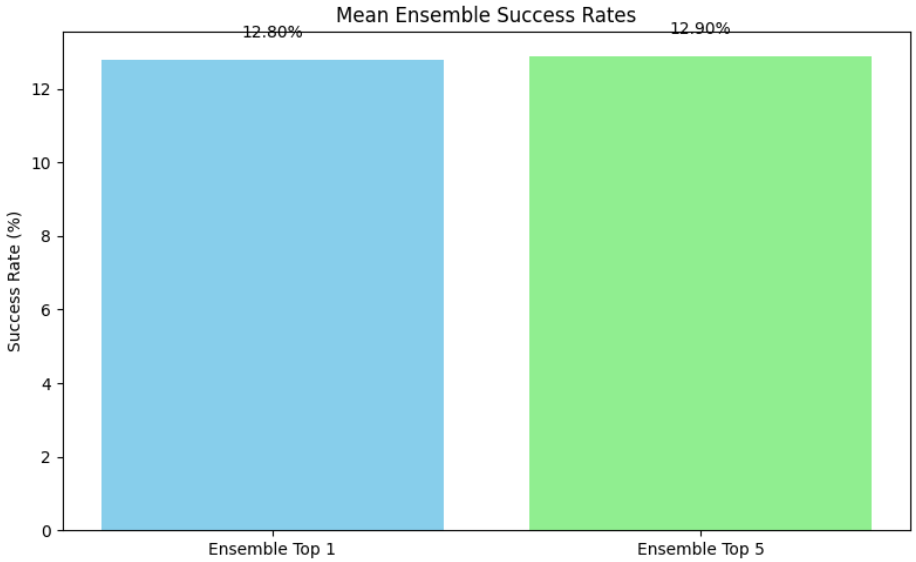


Aşağıdaki grafikte Borda Count (BC Ensemble) yöntemi, bireysel modellerin performansını birleştirerek elde edilen Top1 ve Top5 başarı oranlarını göstermektedir.

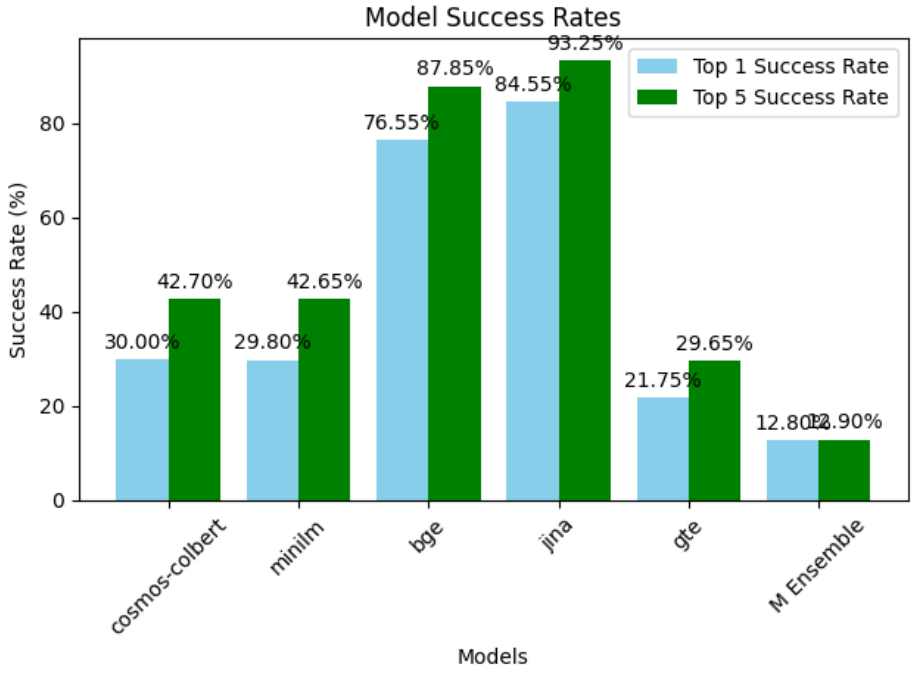


Top1 için %74.25, Top5 için ise %89.90 başarı oranı elde edilmiştir. BC Ensemble yöntemi, yalnızca bireysel modellerin Top1 ve Top5 seçimlerini değil, aynı zamanda bu seçimlerin sıralamalarını da dikkate alarak daha dengeli bir sonuç üretmiştir. Bu durum, özellikle farklı modellerin katkılarının birleştirildiği ve semantik anlam çıkarımının optimize edildiği Top5 başarısında açıkça görülmektedir. BC Ensemble, bireysel modellerden daha yüksek performans göstermiş ve Jina ile BGE gibi güçlü modellerin sonuçlarına oldukça yaklaşmıştır. Bu da Borda Count'un sıralama bilgilerini etkin bir şekilde kullanan etkili bir ensemble yöntemi olduğunu göstermektedir.

Mean Ensemble metodunu kullanarak yapılan ensemble sonucundaki performans aşağıdaki gibidir.

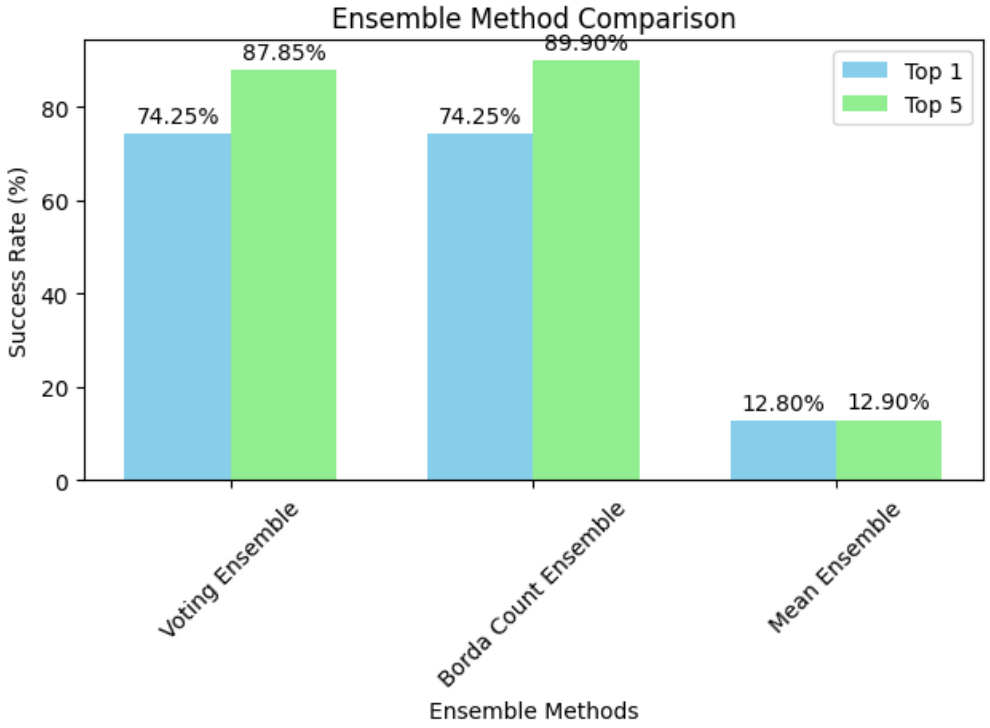


Aşağıdaki grafikte Mean Ensemble (M Ensemble) yöntemi, bireysel modellerin performansını birleştirerek elde edilen Top1 ve Top5 başarı oranlarını göstermektedir.



Mean Ensemble, Top1 için %12.80, Top5 için ise %12.90 başarı oranına ulaşmıştır. Bu oranlar, diğer ensemble yöntemlerine kıyasla oldukça düşük bir performansa işaret etmektedir. Bu durum, Mean Ensemble yönteminin özellikle bireysel modellerin skorlarının birbirinden çok farklı olduğu durumlarda etkili olamayabileceğini göstermektedir. Bununla birlikte, yöntemin Top5 başarısındaki göreceli artış, modellerin skorlarının daha geniş bir yelpazede toplandığı durumlarda performans gösterebileceğini ifade etmektedir. Ancak genel olarak Mean Ensemble, bu çalışmada diğer ensemble yöntemlerine kıyasla daha düşük bir başarı sergilemiştir.

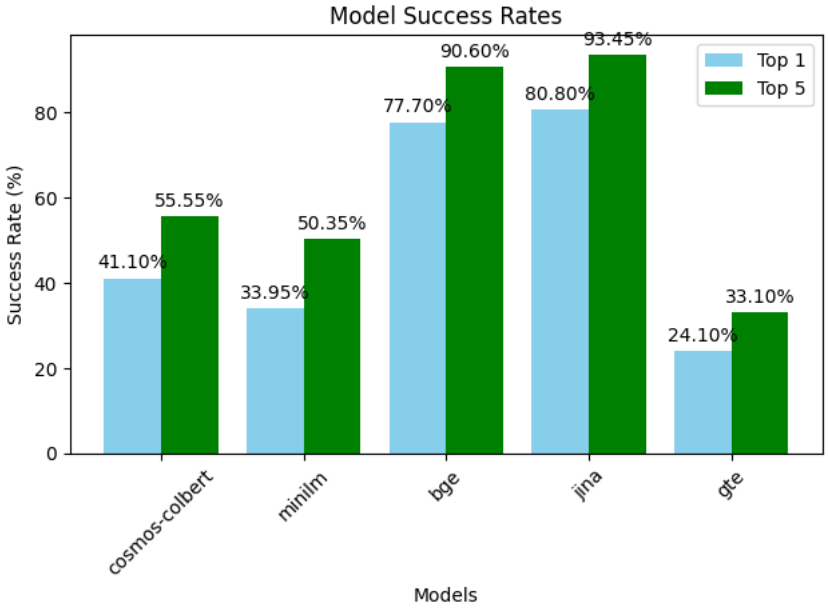
Grafikte üç farklı ensemble yöntemi olan Majority Voting, Borda Count ve Mean Ensemble yöntemlerinin Top1 ve Top5 başarı oranları karşılaştırılmaktadır.



Majority Voting ve Borda Count yöntemleri benzer Top1 başarı oranları (%74.25) elde ederken, Top5 başarı oranlarında Borda Count yöntemi (%89.90) ile öne çıkmıştır. Bu durum, Borda Count'un sıralama bilgilerini dikkate alarak daha dengeli ve doğru sonuçlar üretebildiğini göstermektedir. Mean Ensemble yöntemi ise hem Top1 (%12.80) hem de Top5 (%12.90) başarı oranlarında diğer yöntemlerin oldukça gerisinde kalmıştır. Bu sonuç, Mean Ensemble yönteminin düşük performanslı modellerin etkisini dengelemekte yetersiz kaldığını ve güçlü modellerin etkisini tam anlamıyla yansıtamadığını ortaya koymaktadır. Genel olarak, Majority Voting ve Borda Count yöntemleri daha etkili sonuçlar sunarken, Mean Ensemble yöntemi bu veri seti ve görev için uygun bir seçenek olmamıştır.

5.2 İkinci Veri Seti İçin Sonuçlar

Modellerin tekil başarıları aşağıdaki gibidir.



Cosmos-ColBERT, Top1 için %41.10 ve Top5 için %55.55 başarı oranı ile orta düzey bir performans sergilemiştir. Bu model, bağlamsal ve semantik bilgiyi yakalamada yeterli olsa da diğer modellerle karşılaştırıldığında, daha karmaşık semantik ilişkileri çözmede sınırlı kalmıştır. Top5 başarısının Top1'e göre daha yüksek olması, modelin doğru cevabı daha geniş bir sonuç kümesi içinde bulma yeteneğini yansıtmaktadır.

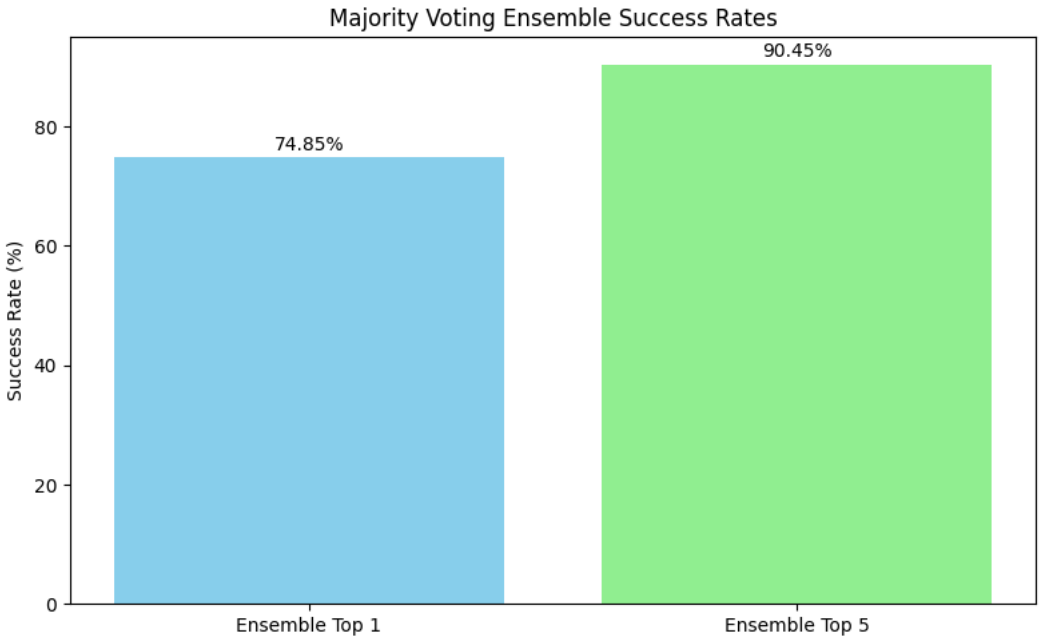
MiniLM, Top1 için %33.95 ve Top5 için %50.35 başarı oranına sahiptir. Hafif bir model olması nedeniyle daha düşük doğruluk oranları göstermesi beklenebilir. Bu model, bağlam bağımlı bilgiyi yakalamakta sınırlı bir performans sergilemiştir. Bununla birlikte, Top5 oranının Top1'e kıyasla daha yüksek olması, daha geniş bir olasılık aralığında doğru cevaba ulaşma kabiliyetini ortaya koymaktadır.

BGE modeli, Top1 için %77.70 ve Top5 için %90.60 başarı oranları ile güçlü bir performans sergilemiştir. Semantik ilişkileri yakalama ve bağlamsal anlam çıkarma yeteneği, bu modelin yüksek doğruluk oranlarına ulaşmasını sağlamıştır. BGE'nin geniş ve detaylı bir veri temsili sunması, özellikle Top5 başarı oranında kendini göstermiştir.

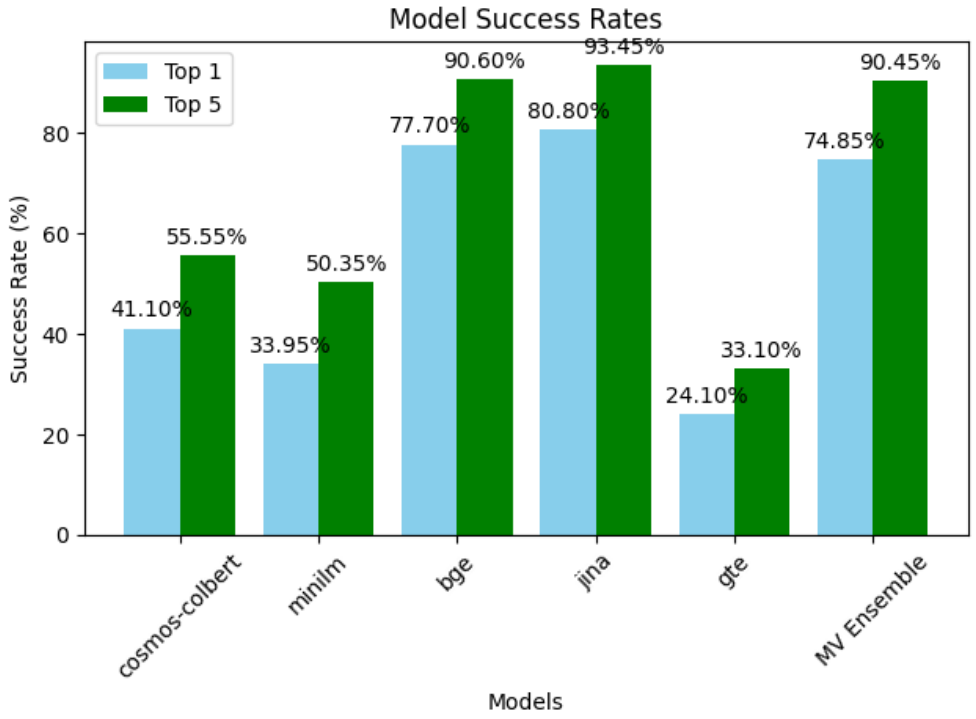
Jina Embeddings, Top1 için %80.80 ve Top5 için %93.45 başarı oranları ile tüm modeller arasında en iyi performansı göstermiştir. Bu model, bağlamsal ve semantik bilgiyi etkili bir şekilde işleyerek doğru cevabı bulmada yüksek bir doğruluk oranına sahiptir. Özellikle Top5 başarısının %93.45 olması, modelin cevap kümesinde doğru cevabı bulmadaki güvenilirliğini açıkça ortaya koymaktadır.

GTE-Large, Top1 için %24.10 ve Top5 için %33.10 başarı oranları ile diğer modellere kıyasla en düşük performansı sergilemiştir. Bu model, bağlam bağımlı ve semantik bilgiyi yakalama konusunda sınırlı bir başarı göstermiştir. GTE'nin düşük doğruluk oranları, daha karmaşık ve bağlama duyarlı görevlerde yetersiz kaldığını göstermektedir.

Majority Voting metodunu kullanarak yapılan ensemble sonucundaki performans aşağıdaki gibidir.

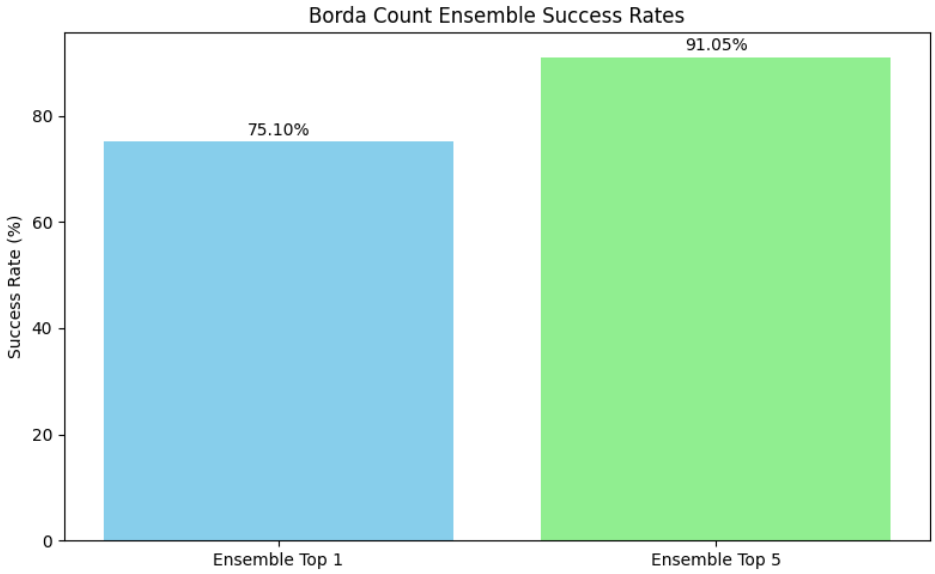


Aşağıdaki grafikte, Majority Voting (MV Ensemble) yöntemi, bireysel modellerin performansını birleştirerek elde edilen Top1 ve Top5 başarı oranlarını göstermektedir.

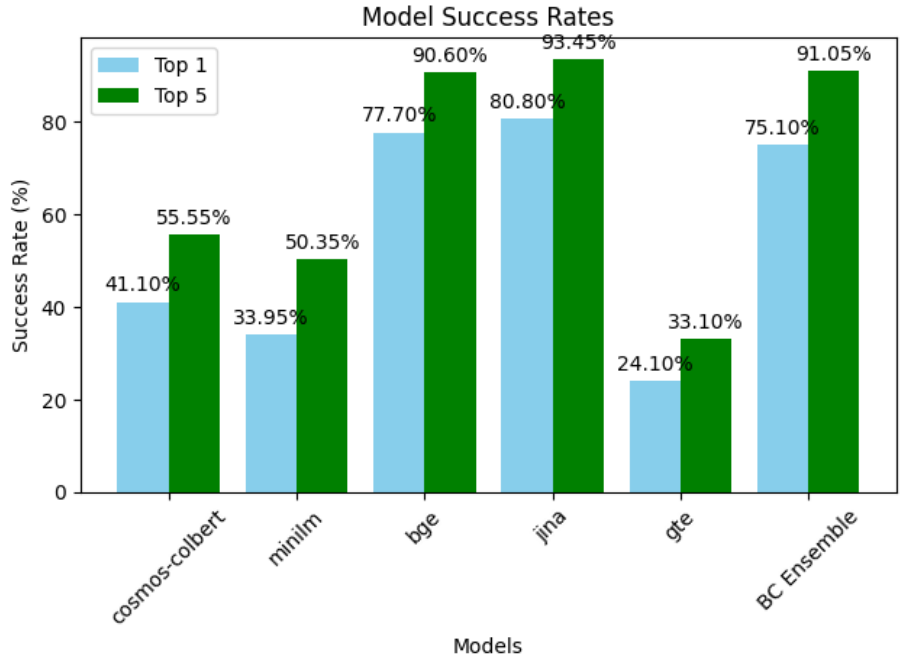


Top1 için %74.85 ve Top5 için %90.45 başarı oranı ile MV Ensemble, bireysel modellerin performansını belirgin şekilde artırmıştır. Özellikle GTE, Cosmos-ColBERT ve MiniLM gibi daha düşük performanslı modellerin sonuçlarının dengelendiği ve BGE ile Jina Embeddings gibi güçlü modellerin katkılarının öne çıktığı gözlemlenmektedir. Top5 başarısının %90.45 gibi yüksek bir seviyeye ulaşması, bu yöntemin doğru cevabı geniş bir sonuç kümesinde güvenilir şekilde yakaladığını göstermektedir. Majority Voting, bireysel modellerin güçlü yönlerini harmanlayarak güvenilir ve dengeli bir sonuç sunmuş, bu nedenle etkili bir ensemble yöntemi olarak öne çıkmıştır.

Borda Count metodunu kullanarak yapılan ensemble sonucundaki performans aşağıdaki gibidir.

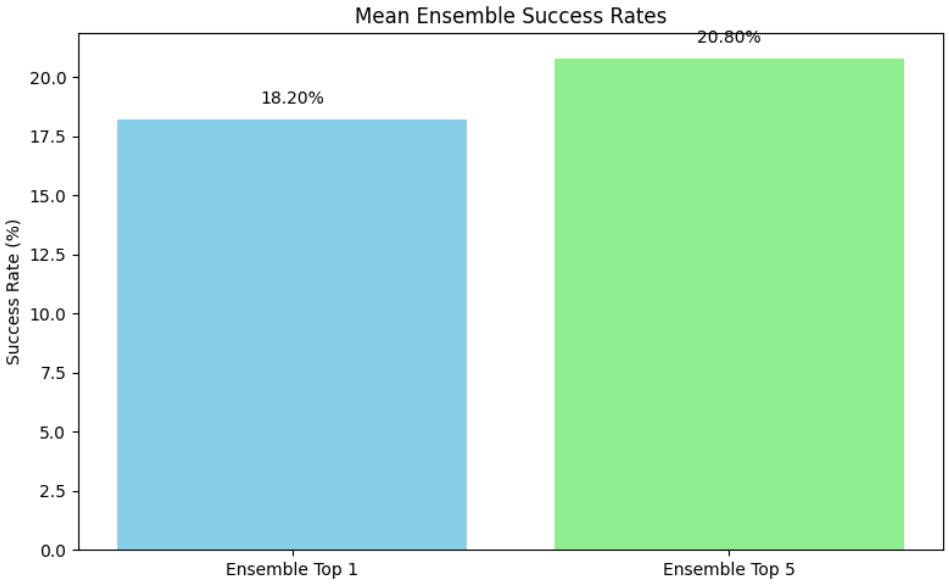


Aşağıdaki grafikte Borda Count (BC Ensemble) yöntemi, bireysel modellerin performansını birleştirerek elde edilen Top1 ve Top5 başarı oranlarını göstermektedir.

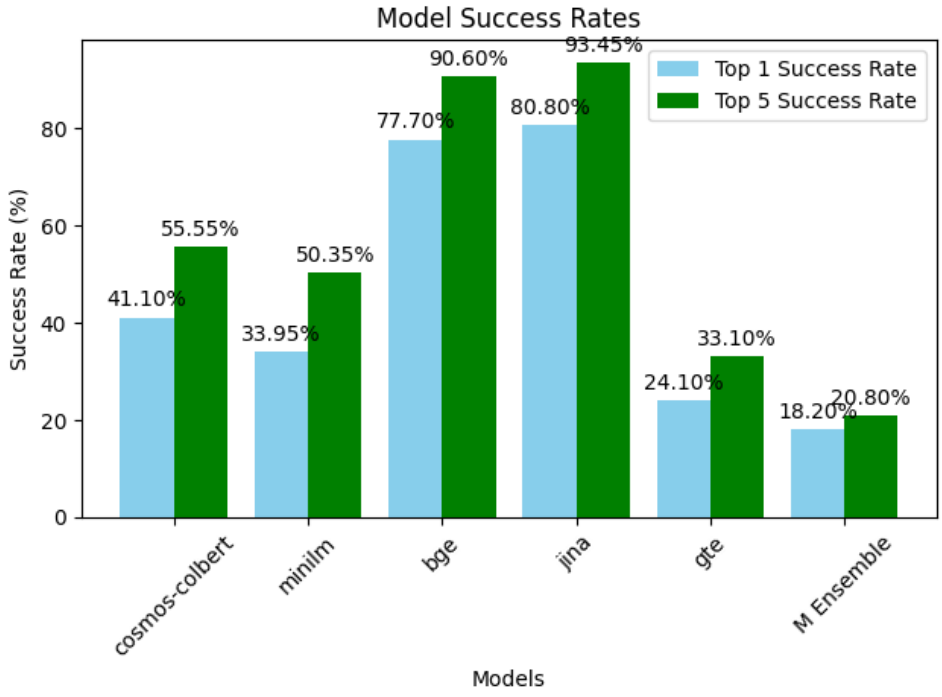


BC Ensemble yöntemi, Top1 için %75.10 ve Top5 için %91.05 başarı oranı ile güçlü bir performans sergilemiştir. Bu yöntemin, bireysel modellerin sıralamalarını dikkate alarak hem düşük performanslı modellerin etkisini dengelediği hem de güçlü modellerin katkılarını optimize ettiği görülmektedir. Özellikle Top5 başarısının %91.05 olması, yöntemin doğru cevabı daha geniş bir sonuç kümesi içerisinde güvenilir bir şekilde yakaladığını göstermektedir. BC Ensemble, sıralama bilgilerini etkili bir şekilde kullanarak hem bireysel modellerden hem de Majority Voting gibi diğer ensemble yöntemlerinden daha iyi bir genel başarı sağlamıştır.

Mean Ensemble metodunu kullanarak yapılan ensemble sonucundaki performans aşağıdaki gibidir.

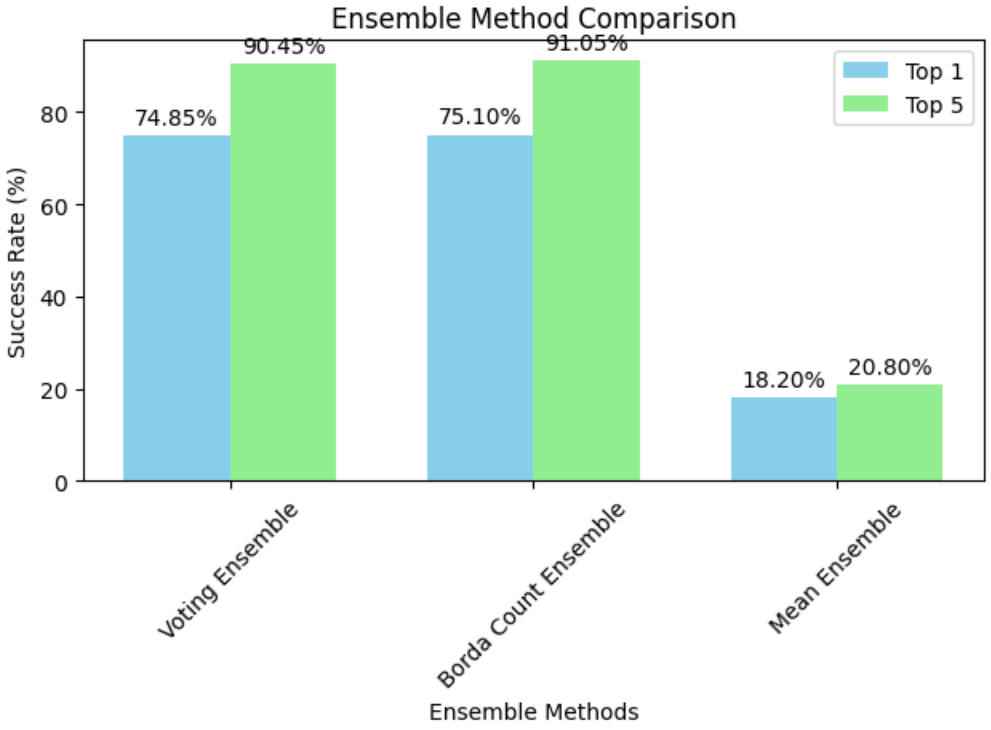


Aşağıdaki grafikte Mean Ensemble (M Ensemble) yöntemi, bireysel modellerin performansını birleştirerek elde edilen Top1 ve Top5 başarı oranlarını göstermektedir.



Mean Ensemble, Top1 için %18.20 ve Top5 için %20.80 başarı oranları ile diğer yöntemlere kıyasla oldukça düşük bir performans sergilemiştir. Bu düşük başarı oranları, yöntemin skorları ortalama alarak güçlü ve zayıf modellerin etkisini dengelemek yerine zayıf modellerin etkisini artırmış olabileceğini göstermektedir. Mean Ensemble'ın başarısızlığı, özellikle düşük performanslı modellerin (örneğin GTE) ağırlığının yüksek olmasıyla ilişkilendirilebilir. Bu durum, yöntemin bağlamsal ve semantik bilgiyi doğru bir şekilde temsil etme kabiliyetinin sınırlı olduğunu ve bu veri seti için uygun bir seçim olmadığını ortaya koymaktadır.

Grafikte üç farklı ensemble yöntemi olan Majority Voting, Borda Count ve Mean Ensemble yöntemlerinin Top1 ve Top5 başarı oranları karşılaştırılmaktadır.



Majority Voting ve Borda Count yöntemleri, Top1 başarı oranlarında birbirine yakın sonuçlar elde etmiştir (%74.85 ve %75.10). Ancak, Top5 başarı oranında Borda Count yöntemi (%91.05) ile en yüksek performansı göstermiştir. Bu, Borda Count'un modellerin sıralamalarını dikkate alarak daha dengeli ve başarılı sonuçlar üretebildiğini ortaya koymaktadır. Öte yandan, Mean Ensemble yöntemi, hem Top1 (%18.20) hem de Top5 (%20.80) başarı oranlarında diğer yöntemlerin oldukça gerisinde kalmıştır. Bu durum, Mean Ensemble'ın güçlü ve zayıf modellerin etkisini dengelemede yetersiz kaldığını ve düşük performanslı modellerin sonuçlarını optimize edemediğini göstermektedir. Genel olarak, Borda Count yöntemi, ensemble yöntemleri arasında en yüksek başarıyı sağlayarak öne çıkmıştır.

6. KAYNAKLAR

[1] https://huggingface.co/datasets/merve/turkish_instructions
[2] https://docs.google.com/spreadsheets/d/1NCMx8QCK4qJzMhFAwMwl9dIH88fyYYvr/edit?gid=1608624003#gid=1608624003
[3] https://huggingface.co/ytu-ce-cosmos/turkish-colbert
[4] https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2
[5] https://huggingface.co/BAAI/bge-m3
[6] https://huggingface.co/jinaai/jina-embeddings-v3
[7] https://huggingface.co/thenlper/gte-large