Yıldız Teknik Üniversitesi Veri Bilimi ve Büyük Veri Yüksek Lisans Programı

2024-2025 Güz Dönemi

Retrieval Ensemble

Müşerref Özkan

235B7053

1. Giriş

Bu proje kapsamında, metin tabanlı sorgu-cevap eşleştirme süreçlerini iyileştirmek amacıyla farklı embedding yöntemleri ve ensemble birleştirme stratejileri incelenmiştir. 51.563 satır ve dört sütundan oluşan bir veri seti kullanılarak, çeşitli gömme modelleri üzerinde sorgu ve cevaplar arasında anlam benzerliği hesaplanmıştır. Bu süreçte cosine similarity kullanılarak benzerlik matrisleri oluşturulmuş ve Top1 ile Top5 doğruluk oranları temel alınarak modellerin bireysel başarıları değerlendirilmiştir. Ayrıca, bireysel embedding yöntemlerinin kararlarını birleştirmek amacıyla average ensemble, weighted average ensemble ve max voting gibi yöntemler uygulanmış ve bu yöntemlerin performansları karşılaştırılmıştır. Projenin sonunda, en yüksek bireysel başarıyı jina-embeddings-v3 modeli gösterirken, weighted average ensemble yöntemi kararların birleştirilmesinde en yüksek başarıyı sağlamıştır. Bu çalışma, metin gömme yöntemlerinin doğruluğunu ve bu yöntemlerin birleştirme stratejileri ile nasıl daha etkili hale getirilebileceğini ortaya koymaktadır.

2. Veri Seti ve Özellikler

Veri seti, 51.563 satır ve 4 sütundan oluşmaktadır. “Unnamed: 0” sütunu indeks bilgilerini, “talimat” sütunu talimat bilgilerini, “çıktı” sütunu ise bu talimatlara karşılık verilen çıktı bilgilerini içermektedir. “giriş” sütunu ise “talimat” sütununun devamı niteliğindedir ve talimatlara ek bilgiler sağlamak amacıyla veri setine dahil edilmiştir, ancak her satırda bu sütun doldurulmuş durumda değildir.

Proje kapsamında veri setinden 2.000 satır örnek seçilirken bu durum dikkate alınmış ve dengeli bir veri seçimi yapılmıştır. Seçilen verilerin 1.000’inde “giriş” sütunu doluyken, diğer 1.000'inde bu sütun boş değerlerden oluşmaktadır. Veri seçiminin ardından veri seti üzerinde çeşitli düzenlemeler yapılmıştır.

Bu düzenlemeler kapsamında, “talimat” ve “giriş” sütunları birleştirilmiş ve yeni bir “birleştirilmiş” sütunu oluşturulmuştur. Veri setinin son halinde yalnızca “birleştirilmiş” ve “çıktı” sütunları yer almaktadır. Bu sadeleştirme ile analiz ve modelleme aşamaları daha verimli hale getirilmiştir.

3. Temsil Yöntemleri

Projede kullanılan temsil yöntemleri aşağıda sıralanmıştır. Soru ve cevaplar için uygulanarak benzerlik matrisi hesabı için kullanılmıştır.

- **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.

4. Ensemble Yöntemleri

Bireysel temsil yöntemlerinin kararlarını birleştirmek için farklı yöntemler uygulanıp başarıları değerlendirilmiştir. Bu kapsamda kullanılan birleştirme yöntemleri aşağıda verilmiştir.

Skor Tabanlı Ensemble Yöntemleri:

1. **Average Ensemble:** Tüm modellerin benzerlik skorlarının ortalaması alınır.
2. **Weighted Average Ensemble:** Modeller ağırlıklandırılır ve ağırlıklı ortalama hesaplanır. Proje kapsamında bireysel temsil çıktıları incelenerek ağırlıklar [1,1,1,2,2] olarak belirlenmiştir.
3. **Max Voting:** Her sorgu için en yüksek benzerlik skorunu veren model seçilir.

5. Uygulama

5.1. Benzerlik Matrisinin Hesaplanması

Veri seti düzenlenerek embedding yöntemleri uygulanır. Sorgu embedding'leri ve cevap embedding'leri arasındaki ilişkiyi ölçmek amacıyla benzerlik matrisi oluşturulur. Proje kapsamında benzerlik matrisi için cosine matrix yöntemi kullanılmıştır. Matrisin her bir hücresi, bir sorgu ile bir cevap arasındaki benzerlik skorunu içerir.

$$\textit{cosine similarity} = \frac{u.v}{\|u\|.\|v\|}$$

5.2. Top1 ve Top5 Belirlenmesi

Oluşturulan cosine matrix üzerinden sorgular için en yüksek benzerlik skoruna sahip (Top1) ve en yüksek ilk beş benzerlik skoruna sahip (Top5) cevaplar seçilir. Top1 ve Top5 sonuçları embedding yöntemlerinin başarılarını ölçmek için kullanılır.

5.3. Doğruluk Hesabı

Yöntemlerin başarılarını değerlendirmek için bir doğruluk hesaplaması yapılır. Yöntem, başarı değerlendirmesi için seçilen Top1 değerinin cevap ile eşleşmesini kontrol eder. Top1, cevap ile eşleşiyorsa doğru kabul edilir. Top5 yönteminin başarı değerlendirmesi için ise cevabın, Top5 yanıtları içinde olma durumunu kontrol eder. Cevap, Top5 içinde bulunuyorsa doğru kabul edilir.

5.4. Benzerliklerin Birleştirilmesi

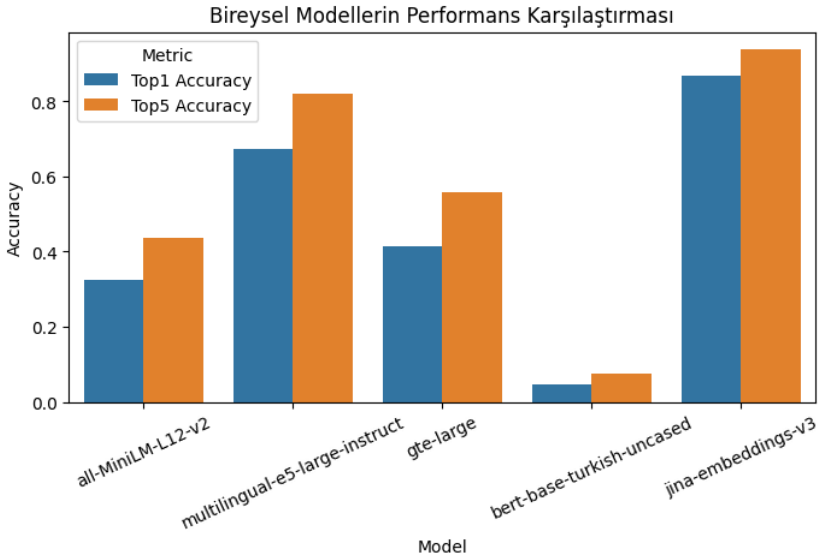
Bireysel embedding yöntemlerinin kararları, yukarıda belirtildiği şekilde average ensemble, weighted average ensemble ve max voting yöntemleri ile birleştirilmiştir. Weighted average ensemble yöntemi için ağırlıklar, bireysel temsil yöntemlerinin başarı oranlarına göre optimize edilmiştir. Bu süreçte, en yüksek performansı gösteren jina-embeddings-v3 ve multilingual-e5-large-instruct modellerinin daha baskın bir katkı sağlaması hedeflenmiştir. Farklı değerlerin test edilmesi sonucunda, ağırlıklar [1,10,1,1,10] olarak uygun bulunmuştur.

6. Sonuçlar

6.1. Bireysel Embedding Başarısı

Bireysel embedding başarıları değerlendirildiğinde jina-embeddings-v3 modeli en yüksek başarıyı gösterirken, bert-base-turkish-uncased modeli en düşük başarıyı göstermiştir. Yüksek başarı gösteren ikinci yöntemin ise multilingual-e5-large-instruct olduğu gözlemlenmiştir. Bu iki başarılı yöntemin başarısı diğer modellerin başarısına göre oldukça yüksektir.

Tüm modeller için Top5 başarısı, Top1 başarısından yüksektir. Bu durumda Top1 yanıtının gerçek cevapla eşleşmediği durumlarda Top5 cevapları arasında bulunmaktadır. jina-embeddings-v3 için Top1 ve Top5 doğrulukları arasındaki fark azdır. Bu durumda yöntemin doğru cevabı bulmakta daha tutarlı olduğu söylenebilir. multilingual-e5-large-instruc yöntemi içinse Top5 ve Top1 başarısı arasındaki fark fazladır. Bu durumda modelin doğru cevabı Top5 içinde bulmakta oldukça başarılı olduğunu, ancak en doğru cevabı Top1 olarak seçmede nispeten zorlandığı söylenebilir.



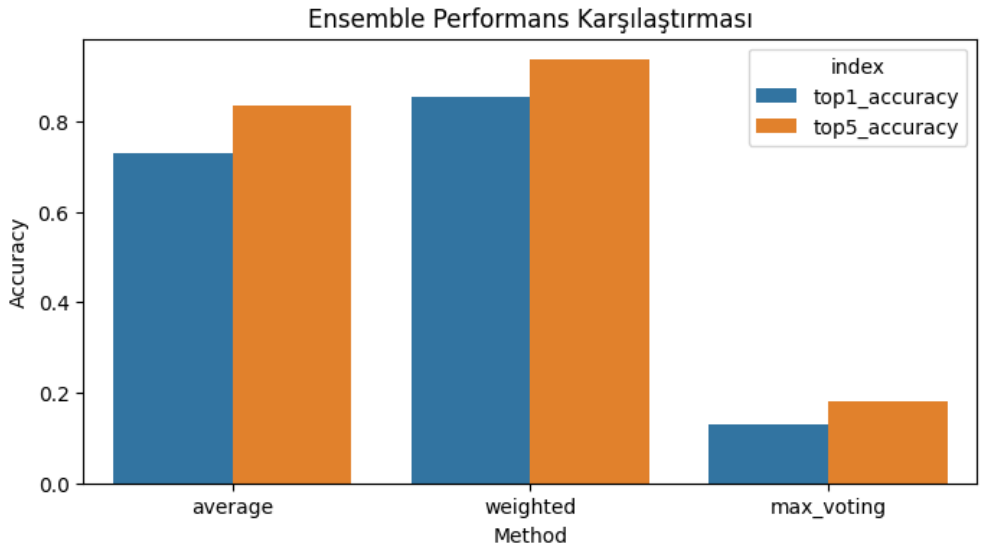
6.1. Ensemble Başarısı

Kararların birleştirilme yöntemlerinden weighted average en yüksek başarıyı gösterirken, max voting yöntemi en düşük başarıyı göstermiştir.

Average ensemble yöntemi tüm modelleri eşit ağırlıkta değerlendirerek dengeli bir sonuç üretir. Özellikle Top5 başarısı açısından iyi olduğu söylenebilir.

Weighted average yöntemi, daha yüksek performanslı modellere daha fazla ağırlık verildiğinde etkili bir sonuç üretir. Top 5 başarısı oldukça yüksektir. Bu durum yöntemin bireysel başarısı yüksek olan modellerden daha fazla etkilendiğini göstermektedir.

Max voting yöntemi, her modelin en yüksek skorunu değerlendirdiği için yanlış yüksek skorların etkisiyle başarısız olabilir.Bu sonuçlar, max voting yönteminin veri setinde etkisiz olduğunu ve doğru cevabı seçmekte zorlandığını gösteriyor.


Ensemble Performans Karşılaştırması
index
top1_accuracy
top5_accuracy
Accuracy
0.0
0.2
0.4
0.6
0.8
average
weighted
max_voting
Method