

# Yoğun ve Seyrek Temsil Modelleri ile Ensemble Yöntemlerinin Bilgi Geri Getirmedeki Rolü: Başarı Oranlarının Derinlemesine Analizi

Muhammed Onur ULU - 24501049  
Bilgisayar Mühendisliği A.B.D., Fen Bilimleri Enstitüsü  
Yıldız Teknik Üniversitesi  
<https://github.com/sweonurulu/ensemble>

**Özet**—Bu çalışmada, WikiRAG-TR veri kümesi kullanılarak bilgi geri getirme performansı çeşitli temsil yöntemleriyle değerlendirilmiştir. Dense vektör temsil yöntemleri (ör. e5, Jina), kelime tabanlı ve BERT-tokenized kelime eşleştirme (BM25) yaklaşımlarıyla Top-1 ve Top-5 başarı oranları ölçülmüştür. Ayrıca, temsil yöntemlerinin birleştirilmesiyle ensemble yöntemlerinin doğruluğa etkisi incelenmiştir. Sonuçlar, Weighted Average Ensemble yönteminin bilgi geri getirme başarısını artırmada diğer yöntemlere kıyasla üstün olduğunu göstermektedir.

**Index Terms**—Information Retrieval, Dense Vector Representations, BM25, BERT-tokenized Matching, Weighted Average Ensemble, Model Evaluation

## I. GİRİŞ

Bilgi geri getirme sistemleri, kullanıcı sorgularına en doğru sonuçları sunabilmek için çeşitli temsil yöntemlerini ve modelleri kullanır. Son yıllarda, dense vektör temsilleri, BERT-tokenized eşleştirme ve BM25 gibi yöntemler, bilgi geri getirme süreçlerinde önemli bir rol oynamaktadır. Ancak, bu yöntemlerin etkinlikleri veri kümesine ve sorgu türüne bağlı olarak farklılık gösterebilir.

Bu çalışmada, WikiRAG-TR veri kümesi kullanılarak bilgi geri getirme sistemlerinde kullanılan farklı temsil yöntemleri ve bu yöntemlerin ensemble yaklaşımları ile birleştirilmesi incelenmiştir. Average Ensemble, Weighted Average Ensemble ve Majority Vote Ensemble yöntemleri karşılaştırılmış ve Top-1 ile Top-5 başarı oranları değerlendirilmiştir. Sonuçlar, özellikle Weighted Average Ensemble yönteminin performansı artırmada etkili olduğunu göstermektedir.

## II. YÖNTEM

### A. Veri Hazırlama

Bu çalışmada, WikiRAG-TR veri kümesi kullanılmıştır. Veri kümesi toplamda 5999 satır ve 8 sütundan oluşmaktadır. Tablo I veri kümesindeki sütunların açıklamalarını sunmaktadır.

Veri kümesindeki `ctx_split_points` değişkenine göre metinler `chunk0`, `chunk1`, ..., `chunk5` sütunlarına ayrılmıştır. Ancak bazı satırlarda toplam chunk sayısı 6'dan az olduğu için, eksik chunk sütunları NaN (boş) değerlerle doldurulmuştur.

Çalışma için, chunk sayısı 3 veya daha fazla olan satırlar rastgele seçilmiş ve bu örneklerden toplam 1000 satır kullanılmıştır.

Tablo I  
WIKIRAG-TR VERİ KÜMESİ SÜTUNLARI

Sütun Adı	Açıklama
<code>id</code>	Her bir soruya özgü benzersiz kimlik
<code>question</code>	Kullanıcı tarafından sorulan soru
<code>answer</code>	Sorunun doğru cevabı
<code>context</code>	Soru ile ilgili metin (chunk'lar burada bulunur)
<code>is_negative_response</code>	Cevabın negatif bir cevap olup olmadığını belirten bayrak (0/1)
<code>number_of_articles</code>	İlgili makale sayısı
<code>ctx_split_points</code>	Chunk'ların ayrıldığı metin karakter konumları
<code>correct_intro_idx</code>	Doğru cevabın bulunduğu chunk'ın indeksi

### B. Model Tanımı

Bu çalışmada, bilgi geri getirme sistemleri için 7 adet dense vector representation modeli ve 2 adet sparse representation modeli kullanılmıştır. Öncelikle, tüm chunk metinleri için gömme vektörleri (embedding) oluşturulmuş ve bu vektörler FAISS (Facebook AI Similarity Search) veritabanına kaydedilmiştir. FAISS, yüksek boyutlu vektör verilerinin hızlı ve etkili bir şekilde aranmasını sağlayan bir kütüphanedir. Bu çalışmada, FAISS veritabanı, chunk gömme vektörlerini depolamak ve sorgulara en uygun chunk'ları getirmek amacıyla kullanılmıştır.

Kullanılan modeller ve temsil yöntemleri aşağıda listelenmiştir:

#### 1) Dense Representations:

- nomic-ai/nomic-embed-text-v1**: Genel amaçlı gömme vektör oluşturmak için kullanılan bir model. Farklı dillerde ve metin türlerinde kullanılabilen geniş kapsamlı bir modeldir.
- hkunlp/instructor-large**: Çeşitli görevlerde metin temsilleri ve bilgi geri getirme için optimize edilmiş büyük ölçekli bir model.
- sentence-transformers/all-roberta-large-v1**: Metin temsilleri ve metin eşleştirme görevleri için geliştirilen büyük bir RoBERTa tabanlı model.
- Thaweewat/jina-embedding-v3-m2v-1024**: Özelleştirilmiş bir Jina modeli olup, metin gömme vektörleri oluşturmak için kullanılır.
- dbmdz/bert-base-turkish-cased**: Türkçe dilinde metin işleme için eğitilmiş, büyük/küçük harf duyarlılığı olan bir BERT tabanlı model.
- intfloat/multilingual-e5-base**: Çok dilli bilgi geri getirme ve metin temsili görevlerinde kullanılabilen etkili

ve hafif bir model.

- **BAAI/bge-m3**: Büyük ölçekli, çok dilli metin temsili ve bilgi geri getirme görevleri için geliştirilmiş bir model.

## 2) Sparse Representations:

- **TF-IDF (Term Frequency-Inverse Document Frequency)**: Metinlerdeki kelime frekanslarını ve bu kelimelerin dokümanlar arasındaki ayırt ediciliğini dikkate alan klasik bir bilgi geri getirme temsil yöntemi.
- **BERT-tokenized TF-IDF**: BERT modeliyle elde edilen tokenlar üzerinden TF-IDF tabanlı sparse temsili oluşturan bir yöntem. Bu yöntem, BERT'in tokenizasyon yeteneklerini TF-IDF'in temsil gücüyle birleştirir.
- **BM25 (Best Matching 25)**: Belirli bir sorgu için doküman sıralaması yapan ve metinler arasındaki benzerlikleri ölçmek için kullanılan popüler bir bilgi geri getirme algoritması. TF-IDF'in geliştirilmiş bir versiyonu olarak kabul edilir.
- **BERT-tokenized BM25**: BERT modeliyle tokenize edilen metinler üzerinden BM25 tabanlı sıralama yapan bir yöntem. Bu yaklaşım, BERT'in güçlü metin işleme kabiliyetlerini BM25'in etkili sıralama mekanizmasıyla birleştirir.

3) *Ensemble Yöntemleri*: Bu çalışmada, çeşitli temsil yöntemlerinden elde edilen sonuçları birleştirmek için üç farklı ensemble yöntemi kullanılmıştır:

- **Equal Weight Ensemble**: Bu yöntemde, tüm modellere eşit ağırlık verilerek tahminler birleştirilir. Her modelin ilk *Top-k* tahminleri alınır ve tahmin edilen her sonuca eşit ağırlık atanır. Rank'e (tahmin sırası) dayalı olarak ağırlık azaltımı uygulanabilir. Elde edilen tahminler arasında oy sayıları hesaplanır ve en fazla oyu alan sonuçlar sıralanarak *Top-k* sonuç olarak belirlenir. Bu yöntem, modeller arasında herhangi bir ayırım yapmaksızın tahminlerin eşit katkısını varsayar.
- **Weighted Vote Ensemble**: Bu yöntemde, modellerin doğruluk oranlarına (*Top-1* ve *Top-5*) dayalı olarak ağırlıklar hesaplanır. Daha yüksek doğruluk oranına sahip modeller, tahminlerinde daha fazla ağırlık taşır. Her modelin ilk *Top-k* tahmini, modelin ağırlıkları ve tahminin sırası (rank'i) dikkate alınarak birleştirilir. Ağırlıklı oylar toplandıktan sonra, en yüksek puanı alan sonuçlar sıralanır ve *Top-k* sonuç olarak belirlenir. Bu yöntem, daha başarılı modellerin etkisini artırarak daha kesin sonuçlar elde etmeyi amaçlar.
- **Majority Vote Ensemble**: Bu yöntem, her bir modelin tahmin ettiği sonuçları çoğunluk esasına göre birleştirir. Her modelin ilk *Top-k* tahminleri alınır ve bu tahminler birleştirilir. Daha sonra, en sık tahmin edilen sonuçlar sıralanarak *Top-k* genel sonuç olarak belirlenir. Son olarak, doğru sonuçların *Top-1* ve *Top-5* içerisinde olup olmadığına göre doğruluk oranları hesaplanır. Bu yöntem, farklı modellerin tahmin güçlerini bir araya getirerek daha güçlü bir sonuç elde etmeyi amaçlar.

Bu yöntemler, temsil yöntemlerinden elde edilen Top-1 ve Top-5 doğruluk oranlarını artırmak ve bilgi geri getirme

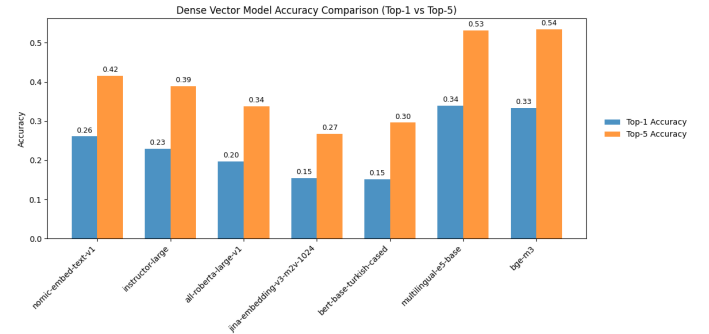
performansını optimize etmek için kullanılmıştır.

## III. BAŞARI SONUÇLARI

Tablo II Dense modellerin bireysel Top-1 ve Top-5 başarı oranlarını göstermektedir. Dense modeller arasında BAAI/bge-m3 en yüksek doğruluk oranlarına ulaşmıştır.

Tablo II  
DENSE MODELLERİN TOP-1 VE TOP-5 BAŞARI ORANLARI

Model İsmi	Top-1 Accuracy	Top-5 Accuracy
nomic-ai/nomic-embed-text-v1	0.2610	0.4160
hkunlp/instructor-large	0.2290	0.3890
sentence-transformers/all-roberta-large-v1	0.1970	0.3380
Thaweewat/jina-embedding-v3-m2v-1024	0.1540	0.2670
dbmdz/bert-base-turkish-cased	0.1520	0.2960
intfloat/multilingual-e5-base	0.3390	0.5310
BAAI/bge-m3	0.3330	0.5350

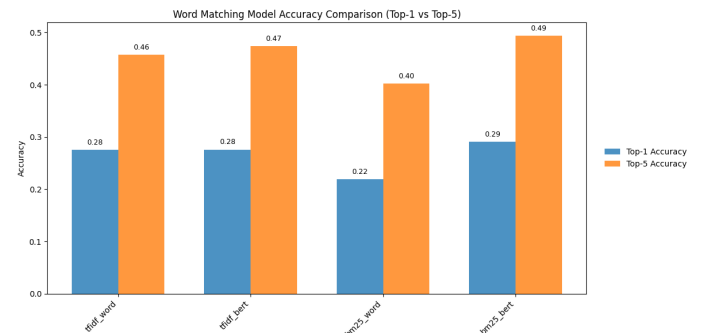


Şekil 1. Dense Modellerin Top-1 ve Top-5 Başarı Oranlarının Grafiği

Tablo III Sparse modellerin bireysel Top-1 ve Top-5 başarı oranlarını göstermektedir. Dense modeller arasında BAAI/bge-m3 en yüksek doğruluk oranlarına ulaşmıştır.

Tablo III  
SPARSE MODELLERİN TOP-1 VE TOP-5 BAŞARI ORANLARI

Model İsmi	Top-1 Accuracy	Top-5 Accuracy
tfidf_word	0.2750	0.4570
tfidf_bert	0.2760	0.4740
bm25_word	0.2190	0.4020
bm25_bert	0.2910	0.4940

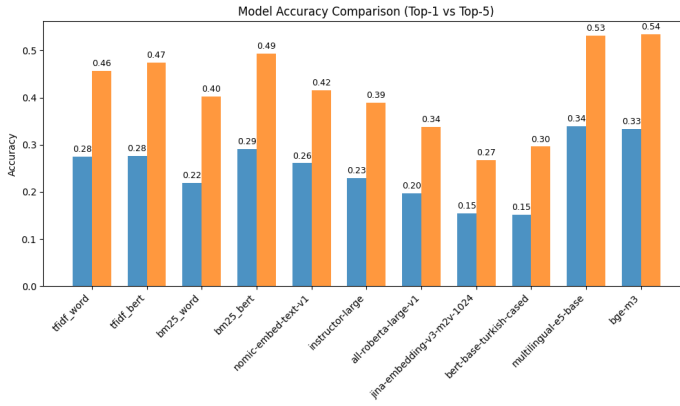


Şekil 2. Sparse Modellerin Top-1 ve Top-5 Başarı Oranlarının Grafiği

Tablo IV tüm modellerin bireysel Top-1 ve Top-5 başarı oranlarını göstermektedir. Dense modeller arasında BAAI/bge-m3, Sparse modeller arasında ise bm25\_bert en yüksek doğruluk oranlarına ulaşmıştır.

Tablo IV  
TÜM MODELLERİN TOP-1 VE TOP-5 BAŞARI ORANLARI

Model İsmi	Top-1 Accuracy	Top-5 Accuracy
<b>Sparse Modeller</b>		
tfidf_word	0.2750	0.4570
tfidf_bert	0.2760	0.4740
bm25_word	0.2190	0.4020
bm25_bert	0.2910	0.4940
<b>Dense Modeller</b>		
nomlc-ai/nomic-embed-text-v1	0.2610	0.4160
hkunlp/instructor-large	0.2290	0.3890
sentence-transformers/all-roberta-large-v1	0.1970	0.3380
Thaweevat/jina-embedding-v3-m2v-1024	0.1540	0.2670
dbmdz/bert-base-turkish-cased	0.1520	0.2960
intfloat/multilingual-e5-base	0.3390	0.5310
BAAI/bge-m3	0.3330	0.5350



Şekil 3. Tüm Modellerin Top-1 ve Top-5 Başarı Oranlarının Grafiği

Tablo IX Top-1 ve Top-5 başarı oranlarını göstermektedir. Equal Weight Ensemble yöntemi, tüm modellerin tahminlerini eşit ağırlıkla birleştirerek sonuçlar üretmiştir. Weighted Vote Ensemble yöntemi, modellerin doğruluk oranlarına dayalı olarak ağırlıklandırılmış oylama yapmıştır ve en yüksek Top-5 doğruluk oranına ulaşmıştır. Modellerin ağırlıkları Tablo X içerisinde verilmiştir.

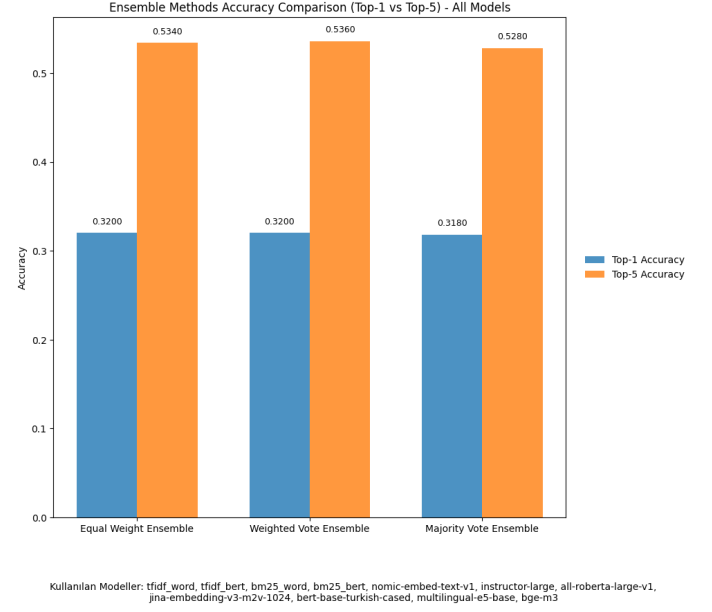
Tablo V  
ENSEMBLE YÖNTEMLERİNİN TOP-1 VE TOP-5 BAŞARI ORANLARI

Yöntem	Top-1 Accuracy	Top-5 Accuracy
Equal Weight Ensemble	0.3200	0.5340
Weighted Vote Ensemble	0.3200	0.5360
Majority Vote Ensemble	0.3180	0.5280

Ensemble Yöntemlerinin Top-1 ve Top-5 Başarı Oranlarının Grafiği

Tablo VI  
MODELLERİN TOP-1 VE TOP-5 DOĞRULUKLARINA DAYALI AĞIRLIKLARI

Model İsmi	Top-1 Ağırlığı	Top-5 Ağırlığı
tfidf_word	0.1009	0.0994
tfidf_bert	0.1012	0.1031
bm25_word	0.0803	0.0874
bm25_bert	0.1067	0.1074
nomlc-embed-text-v1	0.0957	0.0905
instructor-large	0.0840	0.0846
all-roberta-large-v1	0.0723	0.0735
jina-embedding-v3-m2v-1024	0.0565	0.0581
bert-base-turkish-cased	0.0558	0.0644
multilingual-e5-base	0.1244	0.1155
bge-m3	0.1222	0.1163



Şekil 4. Ensemble Yöntemlerinin Top-1 ve Top-5 Başarı Oranlarının Grafiği

Tablo IX Top-1 ve Top-5 başarı oranlarını göstermektedir. Equal Weight Ensemble yöntemi, en iyi 3 dense model (nomic-embed-text-v1, multilingual-e5-base, ve bge-m3) ve BM25 modellerinin (bm25\_word ve bm25\_bert) tahminlerini eşit ağırlıkla birleştirerek sonuçlar üretmiştir. Weighted Vote Ensemble yöntemi, modellerin doğruluk oranlarına dayalı olarak ağırlıklandırılmış oylama yapmıştır ve en yüksek Top-1 doğruluk oranına ulaşmıştır. Modellerin ağırlıkları Tablo X içerisinde verilmiştir.

Tablo VII  
ENSEMBLE YÖNTEMLERİNİN TOP-1 VE TOP-5 BAŞARI ORANLARI  
(FİLTERLENMİŞ MODELLER)

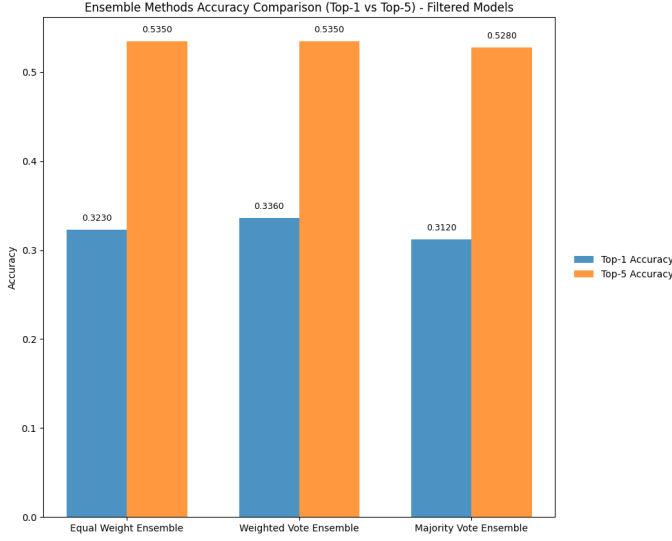
Yöntem	Top-1 Accuracy	Top-5 Accuracy
Equal Weight Ensemble	0.3230	0.5350
Weighted Vote Ensemble	0.3360	0.5350
Majority Vote Ensemble	0.3120	0.5280

Ensemble Yöntemlerinin Top-1 ve Top-5 Başarı Oranlarının Grafiği (Filtrelenmiş Modeller)

Tablo IX Top-1 ve Top-5 başarı oranlarını göstermektedir. Equal Weight Ensemble yöntemi, en iyi

Tablo VIII  
FILTRELENMİŞ MODELLERİN TOP-1 VE TOP-5 DOĞRULUKLARINA  
DAYALI AĞIRLIKLARI

Model İsmi	Top-1 Ağırlığı	Top-5 Ağırlığı
bm25_word	0.1518	0.1690
bm25_bert	0.2017	0.2077
nomic-embed-text-v1	0.1809	0.1749
multilingual-e5-base	0.2349	0.2233
bge-m3	0.2308	0.2250



Kullanılan Modeller: bm25\_word, bm25\_bert, nomic-embed-text-v1, multilingual-e5-base, bge-m3

Şekil 5. Ensemble Yöntemlerinin Top-1 ve Top-5 Başarı Oranlarının Grafiği (Filtrelenmiş Modeller)

5 model (bm25\_word, bm25\_bert, tfidf\_bert, multilingual-e5-base, ve bge-m3) tahminlerini eşit ağırlıkla birleştirerek sonuçlar üretmiştir. Weighted Vote Ensemble yöntemi, modellerin doğruluk oranlarına dayalı olarak ağırlıklandırılmış oylama yapmıştır ve en yüksek Top-1 ve Top-5 doğruluk oranlarına ulaşmıştır. Modellerin ağırlıkları Tablo X içerisinde verilmiştir.

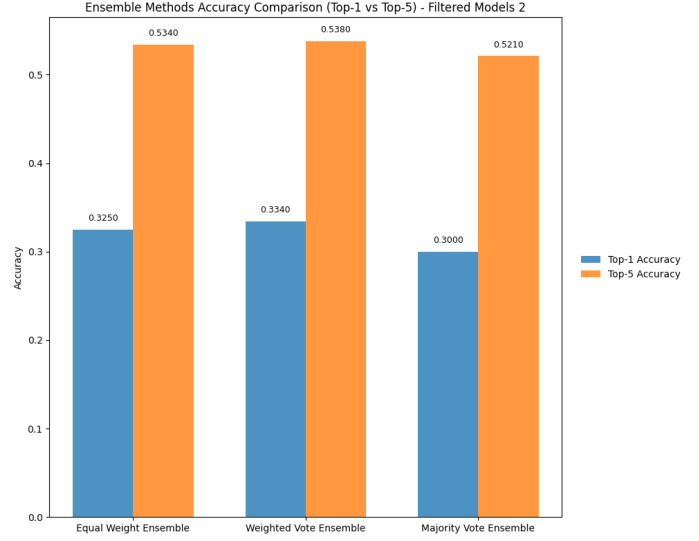
Tablo IX  
ENSEMBLE YÖNTEMLERİNİN TOP-1 VE TOP-5 BAŞARI ORANLARI  
(FILTRELENMİŞ MODELLER)

Yöntem	Top-1 Accuracy	Top-5 Accuracy
Equal Weight Ensemble	0.3250	0.5340
Weighted Vote Ensemble	0.3340	0.5380
Majority Vote Ensemble	0.3000	0.5210

Ensemble Yöntemlerinin Top-1 ve Top-5 Başarı Oranlarının Grafiği (Filtrelenmiş Modeller)

Tablo X  
FILTRELENMİŞ MODELLERİN TOP-1 VE TOP-5 DOĞRULUKLARINA  
DAYALI AĞIRLIKLARI

Model İsmi	Top-1 Ağırlığı	Top-5 Ağırlığı
bm25_word	0.1502	0.1650
bm25_bert	0.1996	0.2028
tfidf_bert	0.1893	0.1946
multilingual-e5-base	0.2325	0.2180
bge-m3	0.2284	0.2196



Kullanılan Modeller: bm25\_word, bm25\_bert, tfidf\_bert, multilingual-e5-base, bge-m3

Şekil 6. Ensemble Yöntemlerinin Top-1 ve Top-5 Başarı Oranlarının Grafiği (Filtrelenmiş Modeller)

Tablo XI Top-1 ve Top-5 başarı oranlarını göstermektedir. Equal Weight Ensemble yöntemi, yalnızca dense modeller (nomic-embed-text-v1, instructor-large, all-roberta-large-v1, jina-embedding-v3-m2v-1024, bert-base-turkish-cased, multilingual-e5-base, ve bge-m3) kullanılarak tahminleri eşit ağırlıkla birleştirmiştir. Weighted Vote Ensemble yöntemi, modellerin doğruluk oranlarına dayalı olarak ağırlıklandırılmış oylama yapmıştır ve en yüksek Top-1 doğruluk oranına ulaşmıştır. Modellerin ağırlıkları Tablo XII içerisinde verilmiştir.

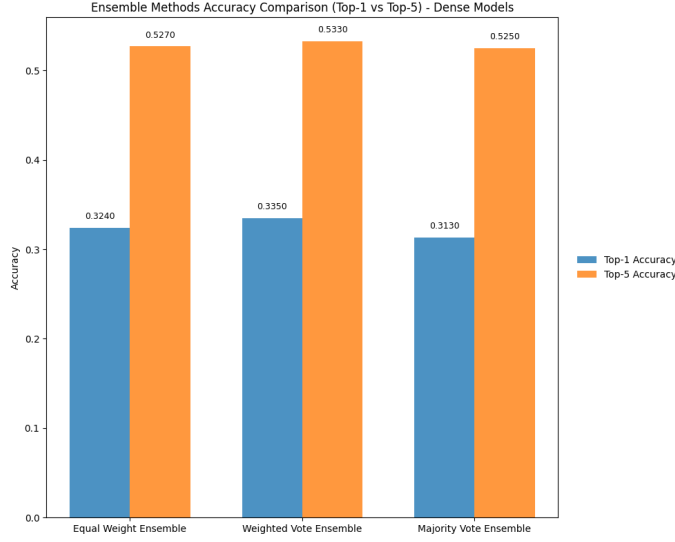
Tablo XI  
ENSEMBLE YÖNTEMLERİNİN TOP-1 VE TOP-5 BAŞARI ORANLARI  
(DENSE MODELLER)

Yöntem	Top-1 Accuracy	Top-5 Accuracy
Equal Weight Ensemble	0.3240	0.5270
Weighted Vote Ensemble	0.3350	0.5330
Majority Vote Ensemble	0.3130	0.5250

Ensemble Yöntemlerinin Top-1 ve Top-5 Başarı Oranlarının Grafiği (Dense Modeller)

Tablo XII  
DENSE MODELLERİN TOP-1 VE TOP-5 DOĞRULUKLARINA DAYALI AĞIRLIKLARI

Model İsmi	Top-1 Ağırlığı	Top-5 Ağırlığı
nomic-embed-text-v1	0.1568	0.1501
instructor-large	0.1375	0.1403
all-roberta-large-v1	0.1183	0.1219
jina-embedding-v3-m2v-1024	0.0925	0.0963
bert-base-turkish-cased	0.0913	0.1068
multilingual-e5-base	0.2036	0.1916
bge-m3	0.2000	0.1930



Kullanılan Modeller: nomic-embed-text-v1, instructor-large, all-roberta-large-v1, jina-embedding-v3-m2v-1024, bert-base-turkish-cased, multilingual-e5-base, bge-m3

Şekil 7. Ensemble Yöntemlerinin Top-1 ve Top-5 Başarı Oranlarının Grafiği (Dense Modeller)

Tablo XIII Top-1 ve Top-5 başarı oranlarını göstermektedir. Equal Weight Ensemble yöntemi, yalnızca Sparse modeller (tfidf\_word, tfidf\_bert, bm25\_word, ve bm25\_bert) kullanılarak tahminleri eşit ağırlıkla birleştirmiştir. Weighted Vote Ensemble yöntemi, modellerin doğruluk oranlarına dayalı olarak ağırlıklandırılmış oylama yapmıştır ve en yüksek Top-1 ve Top-5 doğruluk oranlarına ulaşmıştır. Modellerin ağırlıkları Tablo XIV içerisinde verilmiştir.

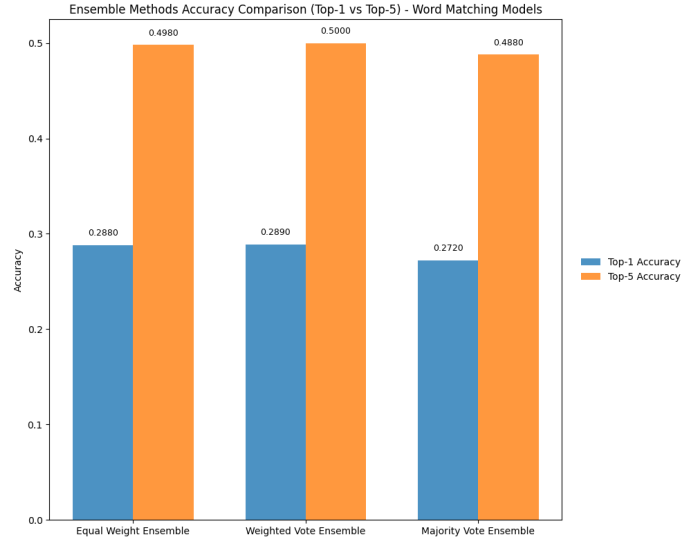
Tablo XIII  
ENSEMBLE YÖNTEMLERİNİN TOP-1 VE TOP-5 BAŞARI ORANLARI (SPARSE MODELLER)

Yöntem	Top-1 Accuracy	Top-5 Accuracy
Equal Weight Ensemble	0.2880	0.4980
Weighted Vote Ensemble	0.2890	0.5000
Majority Vote Ensemble	0.2720	0.4880

Ensemble Yöntemlerinin Top-1 ve Top-5 Başarı Oranlarının Grafiği (Sparse Modeller)

Tablo XIV  
SPARSE MODELLERİN TOP-1 VE TOP-5 DOĞRULUKLARINA DAYALI AĞIRLIKLARI

Model İsmi	Top-1 Ağırlığı	Top-5 Ağırlığı
tfidf_word	0.2592	0.2501
tfidf_bert	0.2601	0.2594
bm25_word	0.2064	0.2200
bm25_bert	0.2743	0.2704



Kullanılan Modeller: tfidf\_word, tfidf\_bert, bm25\_word, bm25\_bert

Şekil 8. Ensemble Yöntemlerinin Top-1 ve Top-5 Başarı Oranlarının Grafiği (Sparse Modeller)

## TEŞEKKÜR

Bu çalışma, Yıldız Teknik Üniversitesi Bilgisayar Mühendisliği Anabilim Dalı Kollektif Öğrenme dersi kapsamında gerçekleştirilmiştir.