TÜRKÇE SAĞLIK SORUNLARI İÇİN HİBRİT SINIFLANDIRMA VE LLM TABANLI ÜRETİMSEL ASİSTAN

Semih Bekdaş

ÖZET

Bu çalışmada, Türkçe sağlık alanındaki hasta–doktor etkileşimlerini desteklemek amacıyla iki bileşenli bir sistem geliştirilmiştir: (i) hasta sorularının metin içeriğinden hareketle doktor uzmanlık alanını otomatik tahmin eden çok sınıflı bir sınıflandırma modülü ve (ii) kullanıcılara tıbbi bilgilendirme sunan üretimsel bir yapay zeka asistanı. Uzmanlık alanı tahmini problemi 16 sınıflı metin sınıflandırma görevi olarak ele alınmıştır. Bu amaçla, ham hâliyle 150.105 kayıt içeren soru–cevap verisi üzerinde temizlik ve filtreleme adımları uygulanmış; 16 branşa indirgenerek sınıf dengesizliğini azaltmak için dengelenmiş 88.816 örnekten oluşan bir alt küme oluşturulmuştur. Eğitim/doğrulama ayrımı 41.493/7.323 olacak şekilde yapılmış, değerlendirme aşamasında 17.888 örnekten oluşan test kümesi kullanılmıştır.

Sınıflandırma modülünde TF-IDF tabanlı klasik yöntemler (Logistic Regression, Linear SVM) ile transformer tabanlı modeller (BERTurk, XLM-RoBERTa) karşılaştırılmıştır. En iyi klasik yöntem Linear SVM ile test kümesinde Accuracy=0.6467 ve Macro F1=0.6481 elde edilirken, transformer tabanlı yaklaşımlar içinde BERTurk testte Accuracy=0.6705 ve Macro F1=0.6882 ile en yüksek performansı sağlamış; XLM-RoBERTa ise Macro F1=0.6434 seviyesinde kalmıştır.

Üretimsel asistan modülünde Meta Llama-3.1-8B-Instruct modeli, bellek verimliliği sağlayan 4-bit nicemleme altında Unsloth altyapısı kullanılarak LoRA (r=16) yöntemiyle ince ayar yapılmıştır. Bu aşamada 503196 örnek içeren Türkçe tıbbi soru–cevap verisi temizlenmiş ve 20 branşta dengelenmiş 60.000 örnek üzerinden (54.000 eğitim / 6.000 doğrulama) eğitim gerçekleştirilmiş; doğrulama kaybı 2.27'den 2.12 seviyelerine düşmüştür. Elde edilen sonuçlar, Türkçe odaklı transformer modelinin klasik yöntemlere kıyasla sınıflandırma başarısını artırdığını ve LoRA tabanlı büyük dil modeli ince ayarının Türkçe tıbbi bilgilendirme üretimi için uygulanabilir bir yaklaşım sunduğunu göstermektedir.

GİRİŞ

Sağlık alanında kullanıcıların şikâyetlerini serbest metin biçiminde ilettiği çevrim içi platformlar; ön değerlendirme, randevu planlama ve doğru uzmana yönlendirme gibi süreçlerde yaygın biçimde kullanılmaktadır. Bu senaryolarda kritik ihtiyaçlardan biri, kullanıcının yazdığı metinden hareketle başvurunun hangi tıbbi uzmanlık alanına (branşa) yönlendirilmesi gerektiğinin otomatik olarak belirlenmesidir. Nitekim literatürde hasta metinlerinden uzmanlık alanı tespiti (specialty detection) probleminin, tele-tıp/uzaktan danışmanlık gibi ortamlarda karar destek bileşeni olarak ele alındığı ve çok sınıflı bir sınıflandırma görevi olarak modellendiği çalışmalar bulunmaktadır [1].

Son yıllarda kullanıcı beklentisi yalnızca "yönlendirme" ile sınırlı kalmamış; aynı zamanda kullanıcıya güvenli sınırlar içinde bilgilendirici ve anlaşılır yanıt üretebilen üretimsel sistemlere ihtiyaç artmıştır. Bu bağlamda problem iki tamamlayıcı bileşenle değerlendirilebilir: (i) serbest metin sorunun uygun uzmanlık alanına yönlendirilmesi (çok sınıflı metin sınıflandırma) ve (ii) tıbbi riskleri gözeterek genel bilgilendirme sağlayan yanıt üretimi (generative QA / diyalog). Bu iki bileşenin birlikte ele alınması, uçtan uca "yönlendirme + bilgilendirme" akışını mümkün kılar.

Metin sınıflandırma görevleri için geleneksel yöntemler ve derin öğrenme yaklaşımları literatürde birlikte kullanılmaktadır. Geleneksel tarafta TF-IDF temsilleri üzerinde Lojistik Regresyon ve doğrusal SVM gibi modeller, güçlü ve maliyet-etkin bir başlangıç (baseline) sunar. Derin öğrenme tarafında CNN/RNN tabanlı yaklaşımlar denenmiş olsa da, transformer mimarilerinin yaygınlaşmasıyla bağlamsal dil temsilleri pek çok sınıflandırma probleminde belirgin başarı artışı sağlamıştır. BERT gibi ön-eğitimli transformer modelleri ve ince ayar (fine-tuning) paradigması, metin sınıflandırma dâhil çok sayıda görevde standart yaklaşım hâline gelmiştir [2].

Türkçe özelinde ise eklemeli (aglutinatif) yapı ve yüksek morfolojik çeşitlilik nedeniyle yüzey biçim çeşitliliği artmakta; kullanıcı metinlerinde görülen yazım hataları, kısaltmalar ve konuşma dili kullanımı da veri gürültüsünü yükseltebilmektedir. Ayrıca Türkçe için görev-özgü, yüksek kaliteli ve büyük ölçekli etiketli veri kaynaklarının görece sınırlı olması, model performansını ve genellenebilirliği doğrudan etkileyen bir faktör olarak raporlanmaktadır [4]. Bu nedenle Türkçe odaklı (monolingual) ön-eğitimli modellerin (ör. BERTurk) Türkçe metinlerde avantaj sağlayabildiği; buna karşılık çok dilli modellerin (ör. XLM-RoBERTa) geniş dil kapsamı sayesinde farklı ifade biçimlerinde daha dayanıklı temsiller öğrenebildiği vurgulanmaktadır [3], [5].

Üretimsel tarafta instruction-tuned büyük dil modellerinin (LLM) yaygınlaşmasıyla sağlık alanında “bilgilendirici asistan” senaryoları güçlenmiştir. Güncel bir temel model olan Llama 3.1-8B-Instruct, alan uyarlaması için güçlü bir başlangıç noktası sunmaktadır [6]. Bununla birlikte Hugging Face ekosisteminde Türkçe tıbbi içerik ve soru-cevap odaklı çeşitli açık modellerin (örn. Doktor-Llama türevleri, Türkçe tıbbi GPT-2 varyantları, Gemma tabanlı tıbbi uyarlamalar) paylaşıldığı görülmektedir [7]–[9]. Ancak LLM’lerin alan verisine uyarlanmasında tam ince ayar yüksek maliyetli olabildiğinden, pratikte LoRA gibi parametre-verimli yöntemler ve düşük-bit nicemleme altında ince ayar yaklaşımları öne çıkmaktadır [10], [11].

Bu çalışmada, Llama 3.1–8B-Instruct modelinin parametre-verimli ince ayarı için Unsloth altyapısı kullanılmış; ayrıca elde edilen modelin yerel ortamda çalıştırılması ve uygulama entegrasyonu için Ollama üzerinden servis edilebilir bir kurulum hedeflenmiştir [12], [13].

Bu çalışmanın amacı, Türkçe sağlık soru metinleri üzerinde (i) branş yönlendirmesi yapan çok sınıflı bir sınıflandırma bileşeni geliştirmek ve klasik ML ile transformer tabanlı yaklaşımları karşılaştırmalı olarak değerlendirmek; (ii) bilgilendirici yanıt üretimi için bir LLM’yi parametre-verimli biçimde Türkçe tıbbi diyalog verisine uyarlayarak üretimsel asistan bileşeni oluşturmaktır. Çalışmanın katkıları özetle: (1) Türkçe sağlık metinleri için gürültü ve kişisel bilgi azaltmayı hedefleyen veri hazırlama hattının tanımlanması, (2) TF-IDF tabanlı güçlü baseline’lar ile BERT tabanlı modellerin aynı deneysel kurgu altında çok metrikli kıyaslanması, (3) LLM tarafında LoRA ve 4-bit nicemleme ile kaynak-verimli bir fine-tuning yaklaşımının uygulanması ve iki bileşenli uçtan uca senaryonun raporlanmasıdır.

VERİ SETİ

Bu çalışmada iki ayrı hedef bulunduğundan (branş sınıflandırma + üretimsel yanıt üretimi), veri hazırlama süreci de iki farklı veri seti üzerinden yürütülmüştür: (i) branş sınıflandırma için

alibayram/doktorsitesi [14], (ii) LLM tabanlı üretimsel asistanın ince ayarı için **kayrab/patient-doctor-qa-tr-167732** [15]. Her iki veri seti de Hugging Face üzerinde herkese açık biçimde yayımlanmıştır; ancak **alibayram/doktorsitesi** veri seti sayfasında erişimin "gated" olduğu ve indirmenin/okumanın belirli koşullara tabi tutulabildiği belirtilmektedir.

1) Branş sınıflandırma veri seti: alibayram/doktorsitesi

Kaynak ve erişim durumu: Veri seti, Doktorsitesi platformundan derlenmiş Türkçe hasta–doktor soru–cevap içeriklerini içermektedir ve Hugging Face üzerinde "alibayram/doktorsitesi" adıyla paylaşılmıştır[14].

Lisans: Veri seti sayfasındaki lisans bölümünde CC BY-NC 4.0 (Creative Commons Attribution-NonCommercial 4.0) lisansı belirtilmiştir. Bu raporda etik/uygunluk açısından daha kısıtlayıcı beyan esas alınmıştır.

Örnek sayısı, alanlar ve sınıf yapısı: Veri seti; doktor unvanı, branş etiketi, soru metni ve doktor yanıtı alanlarını barındırır (örn. doctor_title, doctor_speciality, question_content, question_answer).
Veri seti toplam 187,632 satır içermekte; Train: 150.105 Test: 37.527 splitleri bulunmaktadır. Ham hâliyle etiket uzayı geniş ve belirgin biçimde dengesizdir (bazı branşlar çok yüksek örnek sayısına sahipken birçok branş düşük örnek sayısına sahiptir). Bu durum, çoğunluk sınıflara sapma riskini artırdığı için sınıf filtreleme ve dengeleme adımları uygulanmıştır.

Bu çalışmada kullanılan deneysel alt küme**:** Sınıflandırma görevinde, temiz veri üzerinde **en az 1500** örneğe sahip branşlar tutulmuş ve toplam **16 sınıf** ile çalışılmıştır. Ardından çoğunluk sınıfların baskınlığını azaltmak amacıyla sınıf başına **en fazla 4000 örnek** seçilerek dengeli bir havuz oluşturulmuştur. Bu işlem sonucunda:

- Dengelenmiş eğitim havuzu: **48.816** örnek
- Train/Validation bölünmesi: **41.493 / 7.323**
- Test (orijinal test split'inden aynı **16** sınıfa filtrelenmiş): **17.888**

Dengelenmiş dağılım:

	doctor_speciality	Örnek Sayısı
1	beyin-ve-sinir-cerrahisi	4000
2	dahiliye-ve-ic-hastaliklari-nefroloji	4000
3	ortopedi-ve-travmatoloji	4000
4	kadin-hastaliklari-ve-dogum	4000
5	uroloji	4000
6	genel-cerrahi	3850
7	kadin-hastaliklari-ve-dogum-jinekolojik-onkoloji	3771
8	fiziksel-tip-ve-rehabilitasyon	2970
9	kadin-hastaliklari-ve-dogum-ureme-endokrinolojisi-ve-infertilite	2944
10	kulak-burun-bogaz-hastaliklari	2880

	doctor_speciality	Örnek Sayısı
11	cocuk-sagligi-ve-hastaliklari	2469
12	cocuk-sagligi-ve-hastaliklari-cocuk-norolojisi	2303
13	plastik-rekonstruktif-ve-estetik-cerrahi	2179
14	psikiyatri	2116
15	dermatoloji	1714
16	dahiliye-ve-ic-hastaliklari-endokrinoloji-ve-metabolizma-hastaliklari	1620

2) Üretimsel asistan (LLM) veri seti: kayrab/patient-doctor-qa-tr-167732

Kaynak ve erişim durumu: Üretimsel asistanın ince ayarı için "kayrab/patient-doctor-qa-tr-167732" veri seti kullanılmıştır. Veri seti Türkçe hasta sorusu, yanıt ve branş etiketini birlikte içeren diyalog örnekleri sağlamaktadır. [15]

Lisans: Veri seti sayfasında lisans MIT olarak belirtilmiştir.

Örnek sayısı ve alanlar: Veri seti toplam **563.196** satır içermekte; Train**: 503.196**, Test**: 60.000** splitleri bulunmaktadır.
Temel alanlar "question/answer/speciality" yapısındadır (veri seti kartında sütunlar ve etiket çeşitliliği özetlenmektedir).

Bu çalışmada kullanılan deneysel alt küme: Llama 3.1-8B-Instruct modelini güvenli ve dengeli biçimde uyarlamak için:

- Kapsamlı PII/promo temizliği ve riskli örnek filtreleri sonrası train/test üzerinde temizlik uygulanmış,
- Daha sonra **20 branş** seçilerek **sınıf başına 3000** örnek olacak şekilde **60.000** örnekten oluşan dengeli bir alt küme oluşturulmuş,
- Train/Validation ayrımı **54.000 / 6.000** olarak yapılmış ve bu alt küme üzerinde 1 epoch SFT uygulanmıştır.

3) Ön işleme ve veri hazırlama adımları

Bu projede iki ayrı hat (pipeline) bulunduğundan, ön işleme adımları model ailesine göre farklılaştırılmıştır**.**

3.1. Ortak gürültü/etik temizlik (metin güvenliği)

Her iki veri setinde de gerçek kullanıcı metinleri bulunabildiğinden, etik riskleri azaltmak ve modelin "iletişim bilgisi/tanıtım" gibi sınıflandırma dışı ipuçlarına yaslanmasını engellemek amacıyla aşağıdaki temizlikler uygulanmıştır:

- **PII temizliği:** URL, e-posta, telefon numarası gibi desenler regex ile tespit edilip kaldırılmıştır.

- **Tanıtım/iletişim satırı temizliği:** “randevu, iletişim, tel, whatsapp, muayenehane, klinik, web sitesi…” gibi pazarlama/iletişim çağrışımlı satırlar satır-bazlı olarak silinmiştir.
- **Unicode ve boşluk normalizasyonu:** NFKC normalizasyonu, satır sonu birleştirme ve çoklu boşlukların sadeleştirilmesi yapılmıştır.
- **Tekrarların kaldırılması:** Özellikle “temizlenmiş soru metni” üzerinden duplicate kayıtlar kaldırılmıştır.

LLM ince ayarı tarafında ek olarak:

- **İlaç/doz odaklı örneklerin filtrelenmesi:** “hangi ilaç / kaç mg / reçete” gibi riskli istemleri tetikleyebilen soru kalıpları ve yanıtta doz–frekans pattern’i içeren örnekler elenmiştir (amaç: modelin ilaç/doz öneren davranışa kaymasını azaltmaktır).
- **Uzunluk eşikleri:** Aşırı kısa/aşırı uzun soru–yanıt çiftleri elenerek hem kalite hem de eğitim stabilitesi iyileştirilmiştir.

3.2. Klasik ML hattı (TF-IDF + LogReg / Linear SVM) için ön işleme

Klasik modeller, seyrek TF-IDF özelliklerine dayandığı için Türkçe’ye uygun metin normalizasyonu performansı anlamlı ölçüde etkiler. Bu nedenle:

- **Türkçe lowercasing:** İ/ı dönüşümleri gözetilerek küçük harfe çevirme yapılmıştır.
- **Sayısal ifadeler:** Sayılar “num” gibi bir belirteçle normalize edilmiştir (farklı yazımları tekleştirmek için).
- **Karakter temizliği:** Türkçe harfler dışındaki karakterler sadeleştirilmiştir.
- **Tokenization:** Kelime bazlı parçalama uygulanmıştır.
- **Stop-word çıkarımı:** Türkçe stop-word listesi ile sık geçen işlevsel kelimeler çıkarılmıştır.
- **Kök bulma (stemming):** Türkçe’nin eklemeli yapısında yüzey form çeşitliliğini azaltmak için TurkishStemmer kullanılmıştır.
- **TF-IDF vektörleştirme:** (1,2)-gram, max_features=20.000 yapılandırmasıyla vektör uzayı oluşturulmuştur.

3.3. Transformer sınıflandırma hattı (BERTurk / XLM-R) için ön işleme

Transformer modeller kendi alt-kelime (subword) tokenizasyonlarıyla bağlamsal temsil öğrendiği için, anlamı bozmamak adına minimal temizlik tercih edilmiştir:

- Tip dönüşümü (string’e çevirme), baş/son boşluk kırpma, çoklu boşluk normalizasyonu
- Tokenization doğrudan ilgili model tokenizer’ı ile (truncation + max_length=128, dinamik padding) yapılmıştır.

3.4. LLM ince ayarı (Llama 3.1 + Unsloth/LoRA) için veri formatlama

Üretimsel model eğitiminde örnekler, sohbet şablonu (chat template) biçimine dönüştürülmüştür:

- **System mesajı:** “Sağlık bilgilendirme asistanı; tanı koyamaz, ilaç/doz öneremez…” gibi güvenlik yönergeleri.
- **User mesajı:** “Branş: … / Soru: …” formatı (branş bilgisinin bağlamsal yönlendirme sağlaması için).
- **Assistant mesajı:** Temizlenmiş doktor yanıtı.

Bu yaklaşım, modeli hem Türkçe tıbbi diyalog üslubuna yaklaştırmayı hem de güvenli yanıt çerçevesini korumayı hedefler.

4) Deneylerde kullanılan veri miktarları (özet tablo)

Bileşen	**Veri seti**	**Ham boyut (HF split)**	**Temizlik / filtreleme sonrası**	**Çalışmada kullanılan son split**
Branş sınıflandırma	alibayram/ doktorsitesi	Train: 150.105 Test: 37.527	Train split: 150.105 → 90.305 (temizlenmiş) 16 sınıf filtresi + dengeleme: 48.816 (dengeli havuz)	Train/Val: 41.493 / 7.323 Test: 17.888 (16 sınıfa filtrelenmiş)
Üretimsel asistan (LLM)	kayrab/ patient-doctor-qa-tr-167732	Train: 503.196 Test: 60.000	PII/promo + kalite filtreleri sonrası havuz küçültülmüş; ardından 20 sınıfta dengeli örnekleme uygulanmıştır	60.000 (20×3000) Train/Val: 54.000 / 6.000

YÖNTEMLER

Bu çalışmada iki tamamlayıcı NLP bileşeni geliştirilmiştir:

(i) **Branş yönlendirmesi** için 16 sınıflı çok sınıflı metin sınıflandırma modeli,
(ii) **Bilgilendirici yanıt üretimi** için Türkçe tıbbi diyaloglara uyarlanmış üretimsel bir büyük dil modeli (LLM).
Karşılaştırılabilirlik amacıyla sınıflandırma bileşeninde hem geleneksel makine öğrenmesi hem de transformer tabanlı yaklaşımlar aynı deneysel kurgu altında değerlendirilmiştir. Üretimsel bileşende ise parametre-verimli fine-tuning (LoRA) ve düşük-bit nicemleme ile kaynak verimli bir uyarlama hedeflenmiştir.

1) Branş yönlendirmesi (16 sınıflı metin sınıflandırma)

1.1. Geleneksel ML (Baseline): TF-IDF + Lojistik Regresyon / Doğrusal SVM

Model seçimi gerekçesi:
TF-IDF tabanlı doğrusal sınıflandırıcılar (Logistic Regression, Linear SVM), metin sınıflandırmada hızlı eğitilen ve güçlü bir başlangıç çizgisi (baseline) sunan standart yöntemlerdir. Özellikle n-gram tabanlı temsiller, branşları ayırt eden anahtar kelime örüntülerini yakalamada etkilidir. Bu nedenle transformer tabanlı yaklaşımların gerçek katkısını göstermek için baseline olarak kullanılmıştır.

Embedding / temsil yaklaşımı:
Bu hat, bağlamsal embedding yerine seyrek (sparse) özellik temsili üretir. Metinler TF-IDF ile vektörleştirilmiş; her örnek, kelime ve ikili kelime öbeklerinin (1–2 gram) ağırlıklarını içeren yüksek boyutlu bir vektörle temsil edilmiştir.

TF-IDF ayarları (özellik çıkarımı):

- ngram_range = (1, 2)
- max_features = 20.000

Model 1: Logistic Regression (Multinomial)

- Çok sınıflı yapı için multinomial sınıflandırma tercih edilmiştir.
- Düzenleme katsayısı **C**, doğrulama kümesinde **Macro F1** ile seçilmiştir.
- Kod tarafında: solver="saga", max_iter=2000, random_state=42 kullanılmıştır.

Model 2: Linear SVM (LinearSVC)

- Yüksek boyutlu TF-IDF uzayında maksimum marjin yaklaşımı nedeniyle güçlü bir baseline sağlar.
- Düzenleme katsayısı **C**, doğrulama kümesinde **Macro F1** ile seçilmiştir.
- Kod tarafında: max_iter=5000 kullanılmıştır.

Hiperparametre seçimi (klasik modeller):
Her iki modelde de $C \in \{0.1, 0.5, 1.0, 3.0, 10.0\}$ denenmiş; en iyi doğrulama Macro-F1 değerini veren C seçilerek test değerlendirmesi yapılmıştır.

1.2. Transformer tabanlı sınıflandırma: BERTurk ve XLM-RoBERTa (Fine-tuning)

Model seçimi gerekçesi:

- **BERTurk (dbmdz/bert-base-turkish-cased)**: Türkçe odaklı (monolingual) ön-eğitimli bir BERT varyantı olduğu için Türkçe'nin morfolojik özelliklerini ve dil içi örüntülerini güçlü biçimde temsil edebilmesi beklenir.
- **XLM-RoBERTa (xlm-roberta-base)**: Çok dilli (multilingual) ön-eğitim sayesinde farklı ifade biçimlerine dayanıklıdır ve Türkçe için güçlü bir karşılaştırma noktası sunar.

Embedding / temsil yaklaşımı:
Bu modellerde temsil, TF-IDF gibi statik özelliklerden değil; transformer katmanlarının ürettiği bağlamsal (contextual) alt-kelime (subword) embedding'lerinden öğrenilir. Sınıflandırma kararı, son katman temsillerinin bir classification head tarafından 16 branşa ayrıştırılması ile üretilmiştir.

Tokenizasyon ve girdi hazırlama:

- truncation = True
- max_length = 128
- Dinamik padding için DataCollatorWithPadding kullanılmıştır.
  Transformer modellerde anlamı bozmamak için metin temizliği minimal tutulmuş, tokenizasyon doğrudan ilgili modelin tokenizer'ı ile yapılmıştır.

Ortak fine-tuning ayarları (BERTurk ve XLM-R):

- learning_rate = 2e-5
- num_train_epochs = 3
- per_device_train_batch_size = 16
- per_device_eval_batch_size = 16
- weight_decay = 0.01
- Değerlendirme stratejisi: eval_strategy="steps", eval_steps=400
- Model kaydetme: save_strategy="steps", save_steps=400, save_total_limit=1
- En iyi modeli seçme: load_best_model_at_end=True,
  metric_for_best_model="macro_f1", greater_is_better=True
- Donanım hızlandırma: GPU varsa fp16=True (aksi durumda kapalı)

Not (metrik seçimi):

Model seçimi için Macro-F1 kullanılması, sınıflar arası örnek sayısı farklılıklarının sonuçları yanıltmasını azaltmak ve azınlık sınıflardaki hataları daha adil yansıtmak amacıyla tercih edilmiştir.

2) Bilgilendirici yanıt üretimi (LLM): Llama 3.1-8B + Unsloth + LoRA (4-bit)

Bu bölümde amaç, kullanıcıların sağlık sorularına doktor yerine geçmeden, güvenli sınırlar içinde bilgilendirici yanıt üretebilen bir asistan bileşeni oluşturmaktır. Üretimsel model, sınıflandırmadan bağımsız bir ikinci modül olarak ele alınmış ve Türkçe tıbbi diyalog verisi üzerinde parametre-verimli biçimde uyarlanmıştır.

2.1. Model seçimi ve gerekçe

- Temel model: **Meta-Llama-3.1-8B-Instruct**
- Fine-tuning çerçevesi: **Unsloth** (bellek verimliliği ve hız avantajı)
- Yaklaşım: **LoRA (Low-Rank Adaptation)**
  Tam fine-tuning yüksek maliyetli olduğundan, modelin tüm ağırlıklarını güncellemek

yerine belirli projeksiyon katmanlarında düşük-rank adaptörler eğitilerek kaynak kullanımı azaltılmıştır.

2.2. Nicemleme (quantization) ve bellek verimliliği

Model, eğitim sırasında bellek kullanımını azaltmak amacıyla **4-bit nicemleme** ile yüklenmiştir (load_in_4bit=True). Böylece 8B ölçeğindeki bir modelle sınırlı GPU kaynaklarında fine-tuning yapılabilir hale getirilmiştir.

2.3. Eğitim verisi ve sohbet şablonu

- Kullanılan veri: **kayrab/patient-doctor-qa-tr-167732**
- Temizlik sonrası, **20 branşta dengelenmiş 60.000 örnek** seçilmiştir.
- Eğitim/doğrulama ayrımı: **54.000 / 6.000**
- Örnekler, Llama 3.1 sohbet şablonuna uygun biçimde system/user/assistant rollerine dönüştürülmüş; kullanıcı mesajında "Branş: … / Soru: …" yapısı kullanılmıştır.
- Güvenlik kısıtları, sistem mesajında açıkça tanımlanmıştır (tanı koymama, ilaç/doz önermeme, kişisel veri istememe vb.).

2.4. LoRA hedef katmanları ve ayarlar

LoRA adaptörleri, transformer bloğundaki temel projeksiyon katmanlarına uygulanmıştır:

- q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj

LoRA hiperparametreleri:

- r = 16
- lora_alpha = 16
- lora_dropout = 0
- bias = "none"
- Gradient checkpointing: use_gradient_checkpointing="unsloth"

2.5. SFT (Supervised Fine-Tuning) hiperparametreleri

Eğitim, TRL kütüphanesinin SFTTrainer yapısı ile gerçekleştirilmiştir:

- num_train_epochs = 1
- per_device_train_batch_size = 16
- gradient_accumulation_steps = 4 (efektif batch ≈ 64)
- per_device_eval_batch_size = 8
- learning_rate = 1e-4
- warmup_ratio = 0.03
- lr_scheduler_type = "cosine"
- Optimizer: adamw_8bit
- weight_decay = 0.01
- max_grad_norm = 1.0
- Logging/eval/save:

    - logging_steps = 10
    - eval_steps = 250
    - save_steps = 250
    - save_total_limit = 2
- Sayısal format:
    - bf16 = True, fp16 = False
- Uzunluk bazlı gruplayarak verim:
    - group_by_length = True

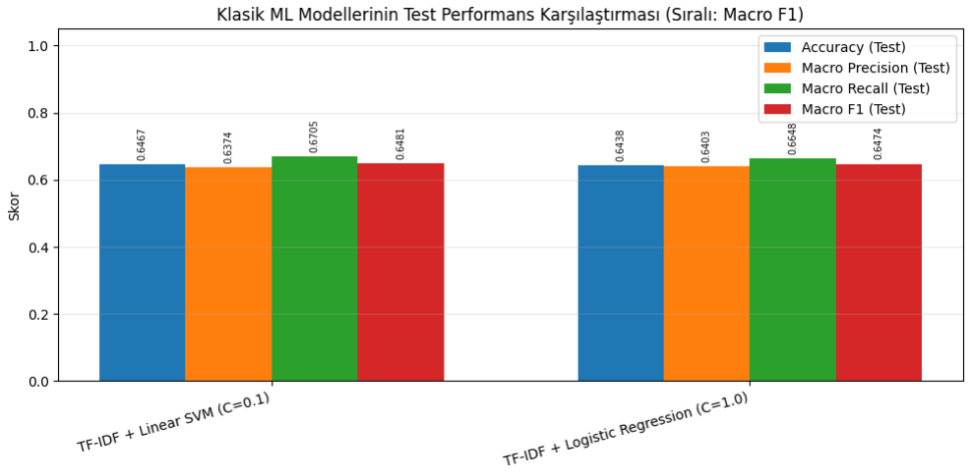
Eğitim hedefi (loss uygulama):
Yanıt kalitesini doğrudan optimize etmek için eğitim kaybı yalnızca assistant yanıtı üzerinde hesaplanacak şekilde ayarlanmıştır (train_on_responses_only). Bu sayede modelin kullanıcı girdisini kopyalama eğilimi azaltılarak, hedeflenen kısım olan yanıt üretimi daha etkin biçimde öğrenilmiştir.

DENEYSEL SONUÇLAR

Bu bölümde, Türkçe sağlık soru metinlerinden uzmanlık alanı (branş) tahmini yapılan 16 sınıflı metin sınıflandırma görevi ile üretimsel yanıt modülünün (LLM fine-tuning) deneysel çıktıları nicel ve görsel olarak sunulmaktadır. Sınıflandırma görevinde **Accuracy**, **Macro Precision**, **Macro Recall** ve **Macro F1** metrikleri raporlanmıştır. Macro ortalamalar, sınıflar arası örnek sayısı farklılıklarında azınlık sınıfların performansını daha adil yansıttığı için özellikle tercih edilmiştir. Deneylerde eğitim/doğrulama/test ayrımı sırasıyla **41.493 / 7.323 / 17.888** örnek olacak şekilde kullanılmıştır.

1) Klasik ML (TF-IDF) Sonuçları

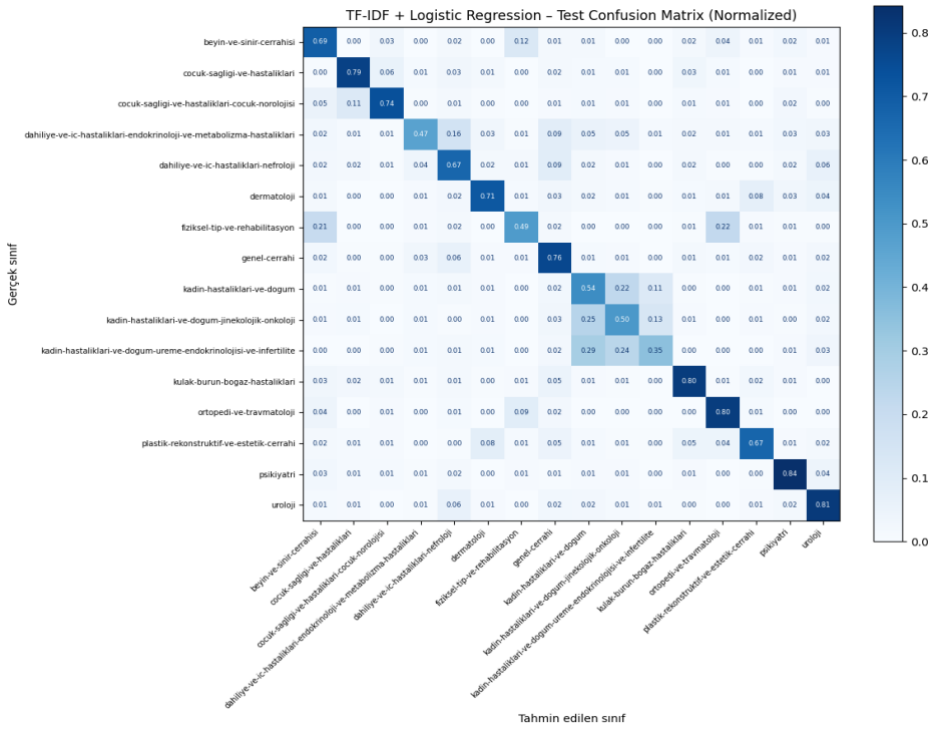
Klasik yaklaşımlarda metinler TF-IDF ile vektörleştirilmiş, ardından **Logistic Regression** ve **Linear SVM (LinearSVC)** ile sınıflandırma yapılmıştır. Her iki model için düzenleme katsayısı **C**, doğrulama kümesinde **Macro F1** performansına göre seçilmiştir. Logistic Regression için en iyi değer **C=1.0** (Val Macro F1=0.6497), Linear SVM için ise **C=0.1** (Val Macro F1=0.6545) bulunmuştur.



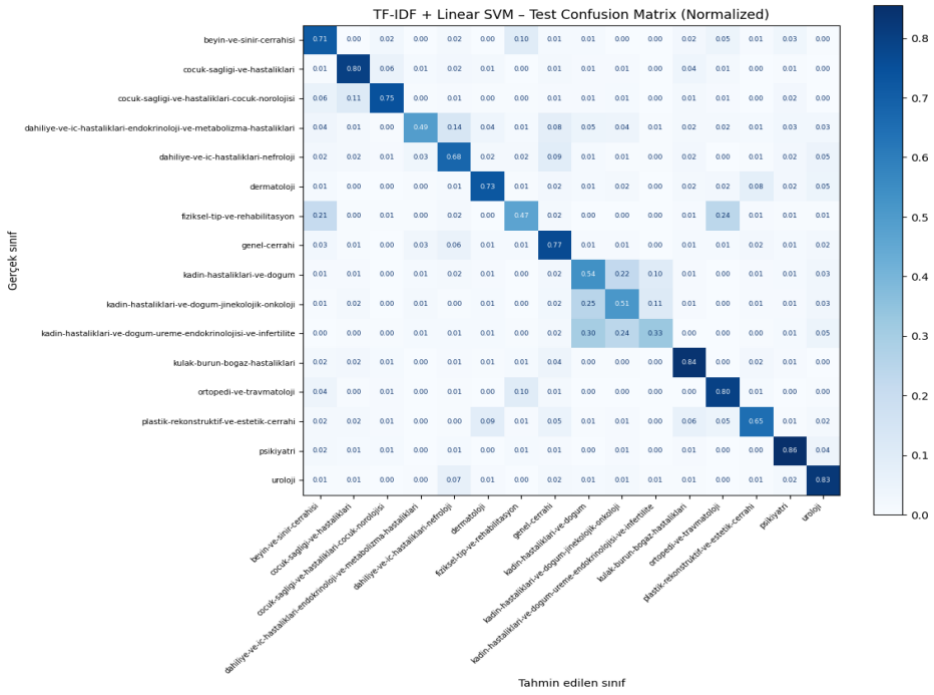
Şekil 1.1 Klasik ML modellerinin test metrik karşılaştırması (Accuracy/Macro P/Macro R/Macro F1).

Mertik karşılaştırması incelendiğinde, iki klasik modelin birbirine yakın performans gösterdiği; **Linear SVM'in** özellikle Macro metriklerde **Logistic Regression'a göre küçük bir farkla** üstünlük sağladığı görülmektedir.

Klasik modellerin hatalarının dağılımını incelemek için test kümesinde **normalize edilmiş confusion matrix** analizi yapılmıştır. Bu görseller, bazı branş çiftlerinde (semptom/terminoloji yakınlığı nedeniyle) karışmaların yoğunlaştığını ve hataların rastgele dağılmadığını göstermektedir.



Şekil 1.2 TF-IDF + Logistic Regression için normalize edilmiş test confusion matrix.



Şekil 1.3 TF-IDF + Linear SVM için normalize edilmiş test confusion matrix.

Confusion matrix'ler birlikte değerlendirildiğinde, iki klasik modelin benzer hata örüntülerine sahip olduğu görülmektedir. Özellikle alt uzmanlıkların birbirine yakın olduğu branş kümelerinde (ör. kadın-doğum alt kırılımları gibi) ve kas-iskelet şikâyetlerinin kesiştiği branşlarda (ör. fiziksel tıp/rehabilitasyon ile ortopedi) çapraz hatalar daha belirgin olabilmektedir. Bu bulgu, yalnızca model kapasitesinin değil, aynı zamanda etiketlerin pratikte ayrıştırılabilirliği ve veri içindeki terminolojik örtüşmenin de performans üst sınırını etkilediğini göstermektedir.

2) Transformer (Fine-tuning) Sonuçları

Transformer tabanlı modellerde metinler ilgili model tokenizer'ı ile tokenize edilmiş ve bağlamsal temsiller üzerinden ince ayar (fine-tuning) uygulanmıştır. **BERTurk (dbmdz/bert-base-turkish-cased)** ve **XLM-RoBERTa (xlm-roberta-base)** modelleri aynı görev için eğitilmiş; eğitim **3 epoch** sürmüş ve değerlendirmeler **400 adımda bir** gerçekleştirilmiştir. En iyi model seçimi için temel ölçüt **Macro F1** olarak belirlenmiştir.

Transformer Modelleri Performansı (Validation/Test)

Model	Val Loss	Val Acc	Val Macro P	Val Macro R	Val Macro F1	Test Loss	Test Acc	Test Macro P	Test Macro R	Test Macro F1
BERTurk	0.9264	0.6859	0.6917	0.6896	0.6889	0.9293	0.6705	0.6817	0.7074	0.6882
XLM-R	0.9803	0.6546	0.6580	0.6588	0.6552	1.0114	0.6289	0.6362	0.6623	0.6434

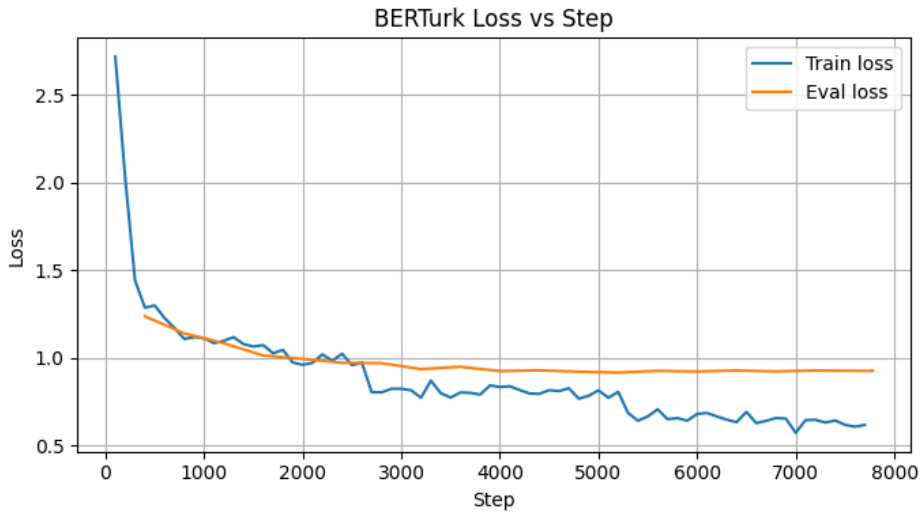
BERTurk modeli doğrulamada **Macro F1=0.6889**, testte ise **Macro F1=0.6882** ile en yüksek performansı sağlamıştır. XLM-R modeli doğrulamada orta düzey sonuç üretmesine rağmen testte düşüş göstermiş ve test **Macro F1=0.6434** olarak ölçülmüştür. Bu bulgu, Türkçeye özel eğitilmiş monolingual modellerin (BERTurk) bu görevde çok dilli modele (XLM-R) göre daha avantajlı olabildiğine işaret etmektedir.

2.1 Eğitim eğrileri (Loss / Accuracy)

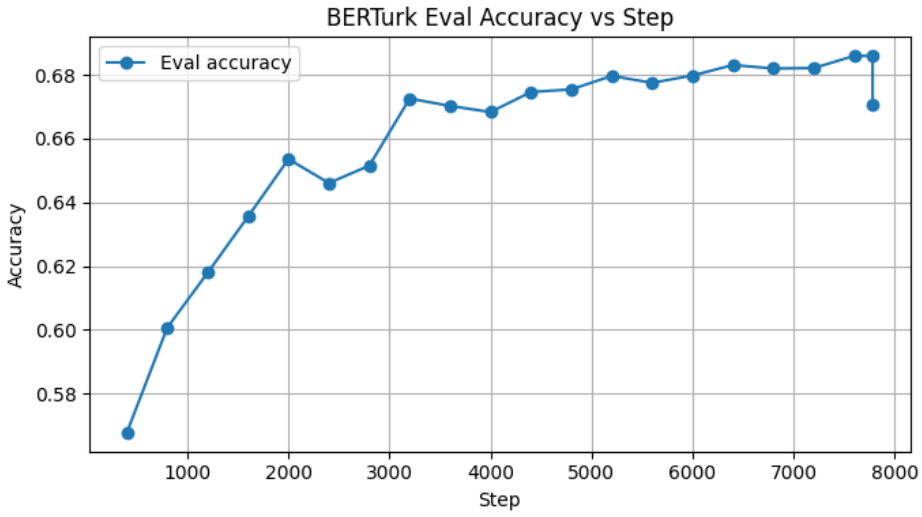
BERTurk tarafında değerlendirme kaybının adımlar boyunca düzenli olarak azalarak yaklaşık **0.93** civarında dengeye oturduğu görülmektedir (Şekil 2.1). Buna paralel olarak doğrulama doğruluğu artarak **0.68–0.69** bandında plato yapmaktadır (Şekil 2.2).

XLM-R tarafında da kayıp azalmaktadır; ancak BERTurk'e kıyasla **daha yüksek** bir değerlendirme kaybında (yaklaşık **1.0** civarı) dengelenmektedir (Şekil 2.3). Doğrulama doğruluğu yaklaşık **0.65** seviyelerine kadar yükselmiş ve bazı adımlarda küçük dalgalanmalar göstermiştir (Şekil 2.4).

Eğitim süreci boyunca loss ve doğruluk değişimi görselleştirilmiştir:



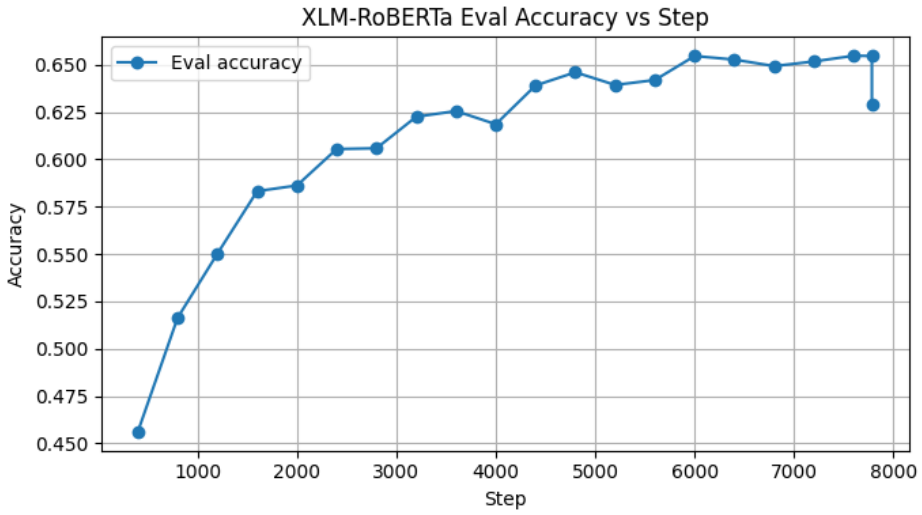
Şekil 2.1 BERTurk modeli için eğitim ve doğrulama kaybının (loss) adımlara (step) göre değişimi.



Şekil 2.2 BERTurk modeli için doğrulama doğruluğunun (eval accuracy) adımlara (step) göre değişimi.

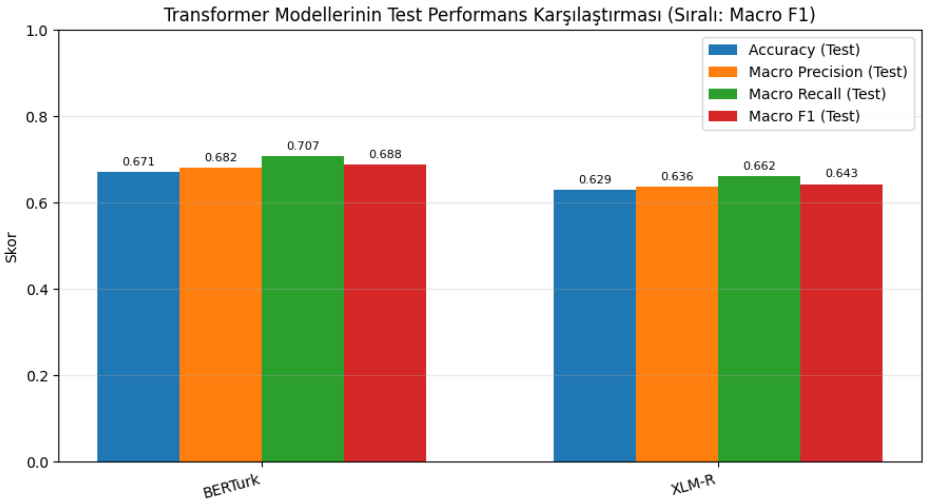
Şekil 2.3 XLM-R modeli için eğitim ve doğrulama kaybının (loss) adımlara (step) göre değişimi.

- **Şekil 2.4de (XLM-R Eval Accuracy–Step):** Doğrulama doğruluğu yaklaşık **0.65** seviyelerine kadar yükselmiş, bazı adımlarda küçük dalgalanmalar göstermiştir.



Şekil 2.4 XLM-R modeli için doğrulama doğruluğunun (eval accuracy) adımlara (step) göre değişimi.

Test metrikleri doğrudan kıyaslandığında BERTurk'ün tüm metriklerde XLM-R'ın üzerinde olduğu net biçimde görülmektedir (Şekil 2.5).

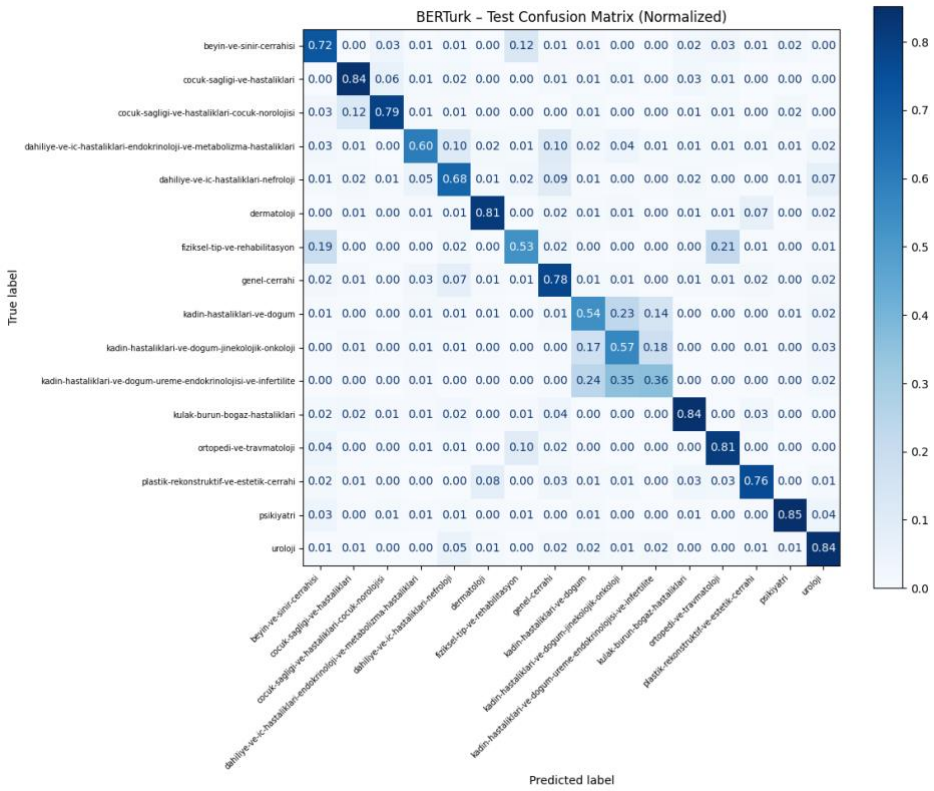


Şekil 2.5 Transformer modellerin test kümesi performans karşılaştırması (Accuracy, Macro Precision, Macro Recall, Macro F1).

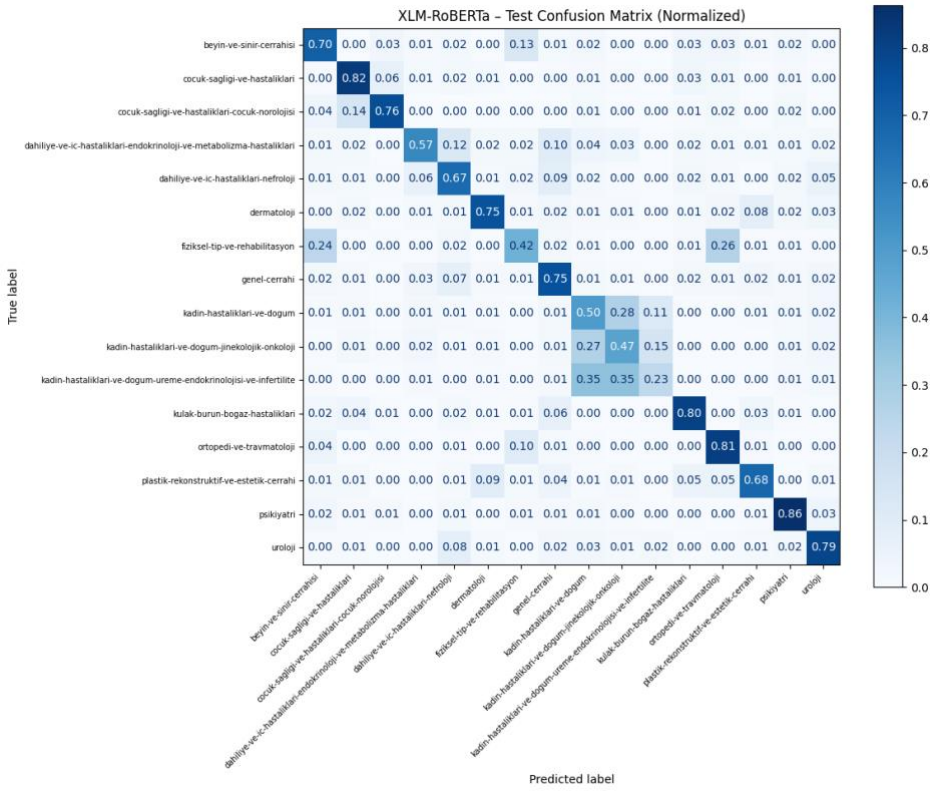
2.2 Confusion matrix analizi

Sınıf bazlı hata örüntülerini incelemek için confusion matrix görselleri değerlendirilmiştir. BERTurk için normalize edilmiş test confusion matrix'te özellikle içerik olarak yakın branşlarda karışmaların arttığı görülmektedir (Şekil 2.6). Öne çıkan örüntüler; kadın-doğum alt uzmanlıkları arasında karşılıklı karışmalar ve kas-iskelet şikâyetlerinde fiziksel tıp/rehabilitasyon ile ortopedi arasında sınırlı ancak belirgin hatalardır.

XLM-R tarafında benzer branş kümelerinde karışmaların BERTurk'e kıyasla daha belirgin olduğu; özellikle kadın-doğum alt uzmanlıkları ile fiziksel tıp/ortopedi ekseninde hataların arttığı izlenmektedir (Şekil 2.7). Bu durum XLM-R'ın test Macro F1 düşüşüyle tutarlıdır.



Şekil 2.6 BERTurk modeli için test kümesinde normalize edilmiş confusion matrix.



Şekil 2.7 XLM-R modeli için test kümesinde normalize edilmiş confusion matrix.

Bu görseller birlikte değerlendirildiğinde, hataların önemli bir bölümünün "etiket sınırlarının pratikte örtüşmesi" (semptom ve terminoloji benzerliği) kaynaklı olduğu; dolayısıyla performans üst sınırının yalnızca model kapasitesine değil veri/etiket ayrıştırılabilirliğine de güçlü biçimde bağlı olduğu anlaşılmaktadır.

3) Üretimsel Bileşen: Llama 3.1–8B (LoRA + 4-bit) Fine-tuning Sonuçları

Sınıflandırma görevinden farklı olarak üretimsel bileşende amaç, kullanıcı sorusuna güvenli sınırlar içinde bilgilendirici yanıt üreten bir asistan elde etmektir. Bu nedenle değerlendirme metrikleri sınıflandırmadaki Accuracy/F1 ile doğrudan kıyaslanmamış; eğitim süreci **eğitim/doğrulama loss** üzerinden izlenmiştir.

Llama 3.1–8B-Instruct temel modeli, **4-bit nicemleme** altında **Unsloth** ile **LoRA (r=16)** kullanılarak eğitilmiştir. Eğitim verisi, temizleme ve filtreleme sonrası **20 branşta dengelenmiş 60.000 örnek** olacak şekilde hazırlanmış; **54.000 eğitim / 6.000 doğrulama** ayrımıyla **1 epoch** SFT uygulanmıştır.

Aşağıdaki ara değerlendirme çıktıları, fine-tuning'in doğrulama kaybını düşürerek öğrenme sağladığını göstermektedir:

Llama 3.1 LoRA Fine-tuning (SFT) ara değerlendirme sonuçları

Step	Training Loss	Validation Loss
250	2.198600	2.273541
500	2.100200	2.164358
750	2.100300	2.126187

Validation loss'un **2.27 → 2.12** seviyesine düşmesi, modelin hedeflenen Türkçe tıbbi diyalog formatına uyumlandığını göstermektedir. Bu modül, raporun genel mimarisinde sınıflandırma bileşeninin "branş yönlendirme" çıktısını tamamlayarak, kullanıcıya "bilgilendirme" sağlayan üretimsel katmanı temsil etmektedir.

TARTIŞMA

Bu çalışmada elde edilen bulgular, Türkçe sağlık soru metinlerinden uzmanlık alanı tahmini görevinde transformer tabanlı yaklaşımların klasik TF-IDF tabanlı yöntemlere kıyasla daha yüksek genel başarı sağladığını, ancak artışın sınırlı kaldığını göstermektedir. Bu durum, problemin doğası gereği beklenebilir bir sonuçtur: Kullanıcıların serbest metin biçiminde yazdığı sağlık soruları yüksek düzeyde gürültü (yazım hataları, eksik noktalama, konuşma dili, kısa ve bağlamı zayıf ifadeler) içerir ve bazı uzmanlık alanları semptom/terminoloji bakımından doğal olarak birbirine yakındır. Dolayısıyla model başarısı yalnızca mimari kapasiteyle değil, veri kalitesi ve etiket ayrıştırılabilirliği ile de güçlü biçimde belirlenmektedir.

Elde edilen sonuçlar genel beklenti ile uyumludur. Bağlamsal (contextual) temsiller üreten BERT türevi modellerin, n-gram frekanslarına dayalı TF-IDF temsilinin ötesine geçerek cümle içi bağlamı ve ifade çeşitliliğini daha iyi yakalaması beklenir. Çalışmada da transformer modellerin özellikle macro metriklerde daha dengeli performans sergilemesi bu beklentiyi desteklemektedir. Bununla birlikte, klasik yöntemlerin de rekabetçi kalması; branşları ayırt etmede belirleyici olabilen anahtar kelime ve n-gram örüntülerinin TF-IDF ile etkili biçimde yakalanabildiğini göstermektedir. Özellikle Linear SVM gibi maksimum marjin tabanlı doğrusal sınıflandırıcılar, yüksek boyutlu seyrek uzaylarda güçlü bir baseline oluşturabilmektedir.

Transformer modeller arasında BERTurk'ün XLM-R'a göre daha başarılı olması, temel olarak BERTurk'ün Türkçe üzerinde monolingual olarak ön-eğitilmiş olmasına bağlanabilir. Türkçe'nin eklemeli (aglutinatif) yapısı nedeniyle aynı kökten çok sayıda yüzey biçiminin türetilmesi, alt-kelime (subword) temsillerinin niteliğini kritik hâle getirir. BERTurk'ün Türkçe dil istatistiklerini daha odaklı biçimde öğrenmesi, sağlık sorularında sık görülen günlük dil + tıbbi terim karışımını daha iyi modellemesine katkı sağlayabilir. Buna karşılık XLM-R, çok dilli ön-eğitim nedeniyle kapasitesini birçok dile paylaştırır; bu durum Türkçe'ye özgü morfolojik/ifade nüanslarının monolingual bir modele kıyasla daha sınırlı temsil edilmesine yol açabilmektedir. Ayrıca XLM-R'ın doğrulamada kabul edilebilir performans gösterip testte düşmesi, doğrulama setine görece daha iyi uyumlanma ve testte genellemenin sınırlı kalması (dağılım farkı ve örnekleme etkileri) olasılığını güçlendirmektedir.

Klasik ML modellerinin benzer hata örüntüleri göstermesi, bu yaklaşımların bağlamdan ziyade yüzey biçimlerine dayanmasıyla açıklanabilir. TF-IDF + n-gram temsili; belirgin branş ipuçları taşıyan kelime kalıplarını yakalamada güçlü olsa da, semptomların ortaklaştığı branş kümelerinde (ör. kadın-doğum alt uzmanlıkları, fiziksel tıp/rehabilitasyon ile ortopedi gibi kesişen şikâyet alanları) bağlamı modelleyemediği için karışmaların tamamen ortadan kalkması beklenmez. Transformer tarafında da benzer karışmaların sürmesi, hataların önemli bölümünün "model yetersizliği"nden ziyade "etiket sınırlarının pratikte örtüşmesi" ile ilişkili olabileceğini düşündürmektedir.

Üretimsel bileşen (Llama 3.1–8B) açısından bakıldığında, LoRA ve 4-bit nicemleme altında yapılan ince ayarın doğrulama kaybını düşürmesi, modelin hedeflenen Türkçe tıbbi diyalog formatına uyumlandığını göstermektedir. Bununla birlikte, üretimsel performansın yalnızca eğitim/doğrulama kaybı ile değerlendirilmesi sınırlıdır; çünkü düşük loss her zaman güvenli, tutarlı ve doğru bilgilendirme kalitesi anlamına gelmeyebilir. Sağlık alanında üretimsel

sistemlerde “halüsinasyon”, aşırı kesin ifade kullanımı veya ilaç/doz gibi riskli önerilere kayma gibi hatalar kritik olduğundan, ilerleyen aşamalarda içerik güvenliği ve doğruluk odaklı nitel değerlendirmelerin (ör. uzman değerlendirmesi, güvenlik kontrol listeleri, vaka bazlı inceleme) eklenmesi daha sağlıklı olacaktır.

Çalışmanın başlıca sınırlılıkları şu şekilde özetlenebilir:

- **Veri gürültüsü ve heterojenliği:** Gerçek kullanıcı metinleri yazım hataları, eksik bağlam, kısaltmalar ve düzensiz dil kullanımı içerebilir. Temizleme adımları gürültüyü azaltmakla birlikte, anlamı etkileyebilecek belirsizlikleri tamamen ortadan kaldıramaz.
- **Etiket ayrıştırılabilirliği:** Bazı uzmanlık alanları semptom ve terminoloji bakımından doğal olarak örtüşür. Bu durum, sınıflandırma için üst sınırı düşürür ve confusion matrix’te tutarlı karışmalara yol açar.
- **Dağılım farkı ve genelleme:** Eğitimde dengeleme yapılmasına karşın test dağılımı farklı olabilir; bu da bazı sınıflarda modelin daha zor örneklerle karşılaşmasına veya domain shift etkisine neden olabilir.
- **Dizi uzunluğu kısıtı:** Transformer modellerde maksimum uzunluk kısıtı (ör. 128 token) uzun sorularda bilgi kaybı yaratabilir; kritik semptom veya bağlam parçaları kırpılabilir.
- **Sınırlı hiperparametre araması:** Fine-tuning ayarları makul bir başlangıç sunsa da, learning rate, max_length, warmup, epoch, weight decay gibi hiperparametreler üzerinde kapsamlı arama yapılmaması erişilebilecek tavan performansı sınırlayabilir.
- **Üretimsel değerlendirme kapsamı:** LLM tarafında nicel kayıp düşüşü gözlenmiş olsa da, güvenli/yararlı yanıt kalitesinin sistematik ölçümü (riskli içerik oranı, güvenlik ihlali, tutarlılık) bu çalışmada sınırlı kalmıştır.

Genel olarak sonuçlar, transformer tabanlı yöntemlerin avantajını doğrularken, performansı belirleyen en kritik faktörlerden birinin veri/etiket yapısı olduğunu göstermektedir. İlerleyen çalışmalarda (i) daha tutarlı normalizasyon ve veri kalite kontrolü, (ii) daha iyi etiket şeması veya hiyerarşik branşlandırma, (iii) sınıf-özel örnek artırma ve hard-negative örnekleme gibi stratejiler, (iv) daha kapsamlı hiperparametre optimizasyonu ve (v) üretimsel modül için güvenlik ve doğruluk odaklı değerlendirme protokollerinin eklenmesiyle sonuçların güçlendirilmesi beklenmektedir.

SONUÇ VE GELECEK ÇALIŞMALAR

Bu çalışmada, Türkçe sağlık platformlarında kullanıcıların serbest metin biçiminde ilettiği hasta sorularından hareketle iki bileşenli bir NLP sistemi ele alınmıştır: (i) sorunun ilgili doktor uzmanlık alanına otomatik yönlendirilmesi (16 sınıflı çok sınıflı metin sınıflandırma) ve (ii) güvenli sınırlar içinde bilgilendirici yanıt üretimi (LLM tabanlı generative QA). Sınıflandırma bileşeni için Doktorsitesi kaynaklı açık veri üzerinde kapsamlı bir veri hazırlama hattı (gürültü/PII temizliği, tekrarların giderilmesi, sınıf filtresi ve dengeleme) uygulanmış; klasik TF-IDF tabanlı yaklaşımlar (Logistic Regression, Linear SVM) ile transformer tabanlı modeller (BERTurk, XLM-R) karşılaştırmalı olarak değerlendirilmiştir. Üretimsel bileşende ise Llama 3.1–8B-Instruct temel modeli, 4-bit nicemleme altında Unsloth ile LoRA kullanılarak Türkçe tıbbi diyalog verisine uyarlanmıştır.

Deneysel bulgular, transformer tabanlı yaklaşımların genel olarak daha yüksek başarı sağladığını ve özellikle Türkçe odaklı ön-eğitimli BERTurk modelinin macro metriklerde en dengeli performansı verdiğini göstermiştir. Bununla birlikte iyi yapılandırılmış bir ön işleme hattı ile desteklenen TF-IDF + Linear SVM yaklaşımı da rekabetçi sonuçlar üreterek, düşük maliyetli ve güçlü bir baseline olarak öne çıkmıştır. Üretimsel tarafta gözlenen doğrulama kaybı düşüşü, LoRA + düşük-bit nicemleme ile kaynak-verimli bir biçimde alan uyarlaması yapılabildiğini ve modelin hedeflenen diyalog formatına uyumlandığını göstermektedir. Genel olarak çalışma, Türkçe sağlık verileri gibi gürültülü ve alan-özgü senaryolarda hem veri hazırlama kalitesinin hem de doğru model ailesi seçiminin performans üzerinde belirleyici olduğunu ortaya koymuştur.

Çalışmanın sınırlılıkları ve elde edilen bulgular ışığında, gelecekte yapılabilecek geliştirmeler aşağıda özetlenmiştir:

- **Daha büyük ve daha kapsamlı veri ile sınıf uzayının genişletilmesi:** Bu çalışmada sınıflandırma tarafında 16 branş ile çalışılmıştır. Gelecekte daha fazla veriyle nadir görülen alt uzmanlıkların da kapsama alınması, modelin gerçek dünyadaki yönlendirme kapsamını artırabilir. Alternatif olarak hiyerarşik etiketleme (üst branş → alt branş) stratejisiyle daha tutarlı bir sınıf şeması oluşturulabilir.
- **Daha güçlü transformer mimarileri ve alan uyarlaması:** BERT ailesinin farklı varyantları (ör. RoBERTa türevleri, DeBERTa benzeri mimariler veya Türkçe odaklı daha güncel encoder modelleri) ile karşılaştırmalar genişletilebilir. Ayrıca tıbbi alan uyarlaması için ek ön-eğitim (domain-adaptive pretraining) veya devam ön-eğitim (continued pretraining) yaklaşımları, özellikle terminoloji yoğun sınıflarda ayrıştırmayı güçlendirebilir.
- **Çok dilli yaklaşımlar ve aktarım öğrenme:** Türkçe verinin sınırlı kaldığı alt branşlarda, çok dilli veri ile desteklenen eğitim (multilingual transfer) veya çapraz-dil veri artırma (cross-lingual augmentation) denenebilir. Bu sayede daha geniş ifade çeşitliliğine dayanıklı modeller elde edilebilir.
- **Daha kapsamlı hiperparametre optimizasyonu ve uzun bağlam yönetimi:** Learning rate, warmup, max_length, epoch sayısı, sınıf ağırlıkları ve düzenlileştirme gibi hiperparametrelerin sistematik aranması performansı artırabilir. Ayrıca uzun soru metinlerinde bilgi kaybını azaltmak için daha yüksek bağlam uzunluğu (max_length) veya akıllı kırpma/özetleme stratejileri uygulanabilir.
- **Hibrit (ensemble) ve hata odaklı iyileştirme:** Klasik modellerin yakaladığı anahtar kelime sinyali ile transformer modellerin bağlamsal sinyalini birleştiren ensemble yaklaşımlar, özellikle sınırda örneklerde başarıyı artırabilir. Confusion matrix üzerinden "sık karışan sınıf çiftleri" belirlenerek hard-negative örnekleme ve sınıf-özel veri artırma stratejileri uygulanabilir.
- **Üretimsel modül için güvenlik ve kalite değerlendirmesinin derinleştirilmesi:** LLM tarafında loss düşüşü tek başına yeterli değildir. Gelecek çalışmalarda, güvenli yanıt üretimi için politika-tabanlı denetimler, red/uyarı şablonları, içerik risk analizi ve uzman değerlendirmesi gibi nitel ölçütlerle sistematik bir değerlendirme protokolü oluşturulmalıdır. Ayrıca, üretimsel modelin "ilaç/doz önerisi" gibi riskli çıktılara kaymasını azaltmak için veri seçimi ve güvenlik odaklı eğitim stratejileri güçlendirilebilir.

Bu öneriler doğrultusunda, sistemin hem branş yönlendirme doğruluğu hem de bilgilendirici yanıtların güvenilirliği artırılarak daha gerçekçi "yönlendirme + bilgilendirme" senaryosuna yakın, uçtan uca bir sağlık destek asistanı mimarisi geliştirilebilir.

KAYNAKÇA

[1] A. Alomari, H. Faris, and P. A. Castillo, "Specialty detection in the context of telemedicine in a highly imbalanced multi-class distribution," arXiv:2402.14039, 2024. [Online]. Available: https://arxiv.org/abs/2402.14039. Accessed: Dec. 2025.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv:1810.04805, 2018. [Online]. Available: https://arxiv.org/abs/1810.04805. Accessed: Dec. 2025.

[3] A. Conneau, K. Khandelwal, N. Goyal, et al., "Unsupervised Cross-lingual Representation Learning at Scale," arXiv:1911.02116, 2019. [Online]. Available: https://arxiv.org/abs/1911.02116. Accessed: Dec. 2025.

[4] Ç. Çöltekin, A. S. Doğruöz, and Ö. Çetinoğlu, "Resources for Turkish Natural Language Processing: A critical survey," arXiv:2204.05042, 2023. [Online]. Available: https://arxiv.org/abs/2204.05042. Accessed: Dec. 2025.

[5] dbmdz, "bert-base-turkish-cased," Hugging Face Model Card. [Online]. Available: https://huggingface.co/dbmdz/bert-base-turkish-cased. Accessed: Dec. 2025.

[6] meta-llama, "Meta-Llama-3.1-8B-Instruct," Hugging Face Model Card. [Online]. Available: https://huggingface.co/meta-llama/Meta-Llama-3.1-8B-Instruct. Accessed: Dec. 2025.

[7] A. Bayram, "Doktor-Llama-3-8b," Hugging Face Model Card. [Online]. Available: https://huggingface.co/alibayram/Doktor-Llama-3-8b. Accessed: Dec. 2025.

[8] kayrab, "ytu_doktor_gpt2-medium," Hugging Face Model Card. [Online]. Available: https://huggingface.co/kayrab/ytu_doktor_gpt2-medium. Accessed: Dec. 2025.

[9] A. Bayram, "DoktorGemma2-9b," Hugging Face Model Card. [Online]. Available: https://huggingface.co/alibayram/DoktorGemma2-9b. Accessed: Dec. 2025.

[10] E. J. Hu, Y. Shen, P. Wallis, et al., "LoRA: Low-Rank Adaptation of Large Language Models," arXiv:2106.09685, 2021. [Online]. Available: https://arxiv.org/abs/2106.09685. Accessed: Dec. 2025.

[11] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, “QLoRA: Efficient Finetuning of Quantized Large Language Models,” arXiv:2305.14314, 2023. [Online]. Available: https://arxiv.org/abs/2305.14314. Accessed: Dec. 2025.

[12] unslothai, “unsloth: Fine-tuning & Reinforcement Learning for LLMs,” GitHub repository. [Online]. Available: https://github.com/unslothai/unsloth. Accessed: Dec. 2025.

[13] Ollama, “Ollama Documentation.” [Online]. Available: https://docs.ollama.com. Accessed: Dec. 2025.

[14] alibayram, “doktorsitesi,” Hugging Face Dataset Card. [Online]. Available: https://huggingface.co/datasets/alibayram/doktorsitesi. Accessed: Dec. 2025.

```
 dataset{bayram_2024_12770916,
 author      = {Bayram, M. Ali},
 title       = {{Türkçe Tıbbi Soru-Cevap Veri Seti: 167 Bin Sağlık
               Sorusu ve Cevabı}},
 month       = jul,
 year        = 2024,
 publisher   = {Zenodo},
 doi         = {10.5281/zenodo.12770916},
 url         = {https://doi.org/10.5281/zenodo.12770916}
}
```

[15] kayrab, “patient-doctor-qa-tr-167732,” Hugging Face Dataset Card. [Online]. Available: https://huggingface.co/datasets/kayrab/patient-doctor-qa-tr-167732. Accessed: Dec. 2025.

```
@dataset{kayrab2024patient-doctor-qa-tr-167732,
 author      = {Muhammed Kayra Bulut},
 title       = {Patient Doctor Q&A TR 167732},
 year        = 2024,
 url         = {https://huggingface.co/datasets/kayrab/patient-doctor-qa-tr-167732},
}
```