T.C.
EGE ÜNİVERSİTESİ
Fen Bilimleri Enstitüsü

TÜRKÇE KLİNİK METİNLERİN DERİN ÖĞRENME YAKLAŞIMLARI İLE SINIFLANDIRILMASI

Hazal TÜRKMEN

DOKTORA TEZİ

Danışman: Prof. Dr. Oğuz DİKENELLİ

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Doktora Programı

İzmir
Temmuz, 2023

T.C.
EGE ÜNİVERSİTESİ
Fen Bilimleri Enstitüsü

TÜRKÇE KLİNİK METİNLERİN DERİN ÖĞRENME YAKLAŞIMLARI İLE SINIFLANDIRILMASI

Hazal TÜRKMEN

DOKTORA TEZİ

Danışman: Prof. Dr. Oğuz DİKENELLİ

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Doktora Programı

İzmir
Temmuz, 2023

EGE ÜNİVERSİTESİ FEN BİLİMLERİ ENSTİTÜSÜ

KABUL VE ONAY SAYFASI

Hazal Türkmen tarafından **DOKTORA** tezi olarak sunulan "TÜRKÇE KLİNİK METİNLERİN DERİN ÖĞRENME YAKLAŞIMLARI İLE SINIFLANDIRILMASI" başlıklı bu çalışma EÜ Lisansüstü Eğitim ve Öğretim Yönetmeliği ile EÜ Fen Bilimleri Enstitüsü Eğitim ve Öğretim Yönergesi'nin ilgili hükümleri uyarınca tarafımızdan değerlendirilerek savunmaya değer bulunmuş ve 12/07/2023 tarihinde yapılan tez savunma sınavında aday oybirliği ile başarılı bulunmuştur.

Jüri üyeleri		**İmza**
Jüri Başkanı	: Prof. Dr. Oğuz Dikenelli	
Raportör Üye	: Dr. Öğr. Üyesi Özgür Gümüş	
Üye	: Prof. Dr. Murat Osman Ünalır	
Üye	: Prof. Dr. Tuğba Özacar Öztürk	
Üye	: Doç. Dr. Selma Tekir	

EGE ÜNİVERSİTESİ FEN BİLİMLERİ ENSTİTÜSÜ

ETİK KURALLARA UYGUNLUK BEYANI

EÜ Lisansüstü Eğitim ve Öğretim Yönetmeliğinin ilgili hükümleri uyarınca Doktora Tezi olarak sunduğum "TÜRKÇE KLİNİK METİNLERİN DERİN ÖĞRENME YAKLAŞIMLARI İLE SINIFLANDIRILMASI" başlıklı bu tezin kendi çalışmam olduğunu, sunduğum tüm sonuç, doküman, bilgi ve belgeleri bizzat ve bu tez çalışması kapsamında elde ettiğimi, bu tez çalışmasıyla elde edilmeyen bütün bilgi ve yorumlara atıf yaptığımı ve bunları kaynaklar listesinde usulüne uygun olarak verdiğimi, tez çalışması ve yazımı sırasında patent ve telif haklarını ihlal edici bir davranışımın olmadığını, bu tezin herhangi bir bölümünü bu üniversite veya diğer bir üniversitede başka bir tez çalışması içinde sunmadığımı, bu tezin planlanmasından yazımına kadar bütün safhalarda bilimsel etik kurallarına uygun olarak davrandığımı ve aksinin ortaya çıkması durumunda her türlü yasal sonucu kabul edeceğimi beyan ederim.

12/07/2023

İmzası

Hazal TÜRKMEN

ÖZET

TÜRKÇE KLİNİK METİNLERİN DERİN ÖĞRENME YAKLAŞIMLARI İLE SINIFLANDIRILMASI

TÜRKMEN, Hazal

Doktora Tezi, Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanı: Prof. Dr. Oğuz Dikenelli

Temmuz 2023, 87 sayfa

Doğal dil işleme (DDİ) alanında son yıllarda kaydedilen önemli ilerlemeler, büyük dil modellerinin (BDM'ler) ortaya çıkışıyla birlikte devrim niteliğinde araştırma ve gelişmeleri mümkün kılmıştır. Özellikle alan içi derlemlerle desteklenmiş ve önceden eğitilmiş olan BDM'lerin sunduğu imkanlar, bu devrim niteliğindeki ilerlemeyi daha da belirgin hale getirmiştir. İngilizce dilinde biyomedikal ve klinik DDİ görevlerinde bu tür modellerin oldukça başarılı olduğu birçok çalışma ile açıkça gösterirken ne yazık ki az kaynaklı dillerde bu tür çalışmaların sayısı oldukça sınırlıdır. Tez çalışması bu noktada, dil kaynaklarının sınırlılıklarını aşma hedefiyle hareket ederek transformer tabanlı iki farklı BDM olan BioBERTurk ve TurkRADBERT dil modelleri ailesini geliştirme yoluna gitmiştir. Literatürde ilk kez geliştirilen Türkçe dilindeki biyomedikal ve klinik alanındaki modellerin performansını iyileştirebilmek amacıyla da farklı dil ön eğitim stratejilerinin etkisini araştırılmıştır. Dil modellerinin performanslarının değerlendirilmesi için tez kapsamında oluşturulan ve Türkçe kafa BT radyoloji raporlarının doküman seviyesinde uzman radyologlar tarafından tasarlanarak etiketlenmiş üç farklı veri kümesi ilk kez oluşturulmuştur. BioBERTurk dil modeli ailesi, biyomedikal derlemle eğitim ve sürekli ön eğitim stratejileri kullanılarak diğer Türkçe modellere kıyasla daha iyi performans göstermiş, ancak radyoloji ve medikal derlemelerini birleştiren ve sıfırdan eğitilen model en düşük başarıyı sergilemiştir. Tez kapsamında geliştirilen bir diğer dil modelleri ailesi TurkRadBERT, klinik alanda Türkçede sınırlı dil kaynaklarıyla, radyoloji raporlarını içeren çok

etiketli bir sınıflandırma görevinde farklı ön eğitim metodolojilerinin Türkçe klinik dil modellerinin performansına olan etkisini araştırmıştır. Genel Türkçe BERT modeli BERTurk ve TurkRadBERT-task v1, genel alan derleminden elde edilen bilgi sayesinde en iyi genel performansı sergilemiştir. Sonuçlar, optimum performans için genel alan bilgisi ve göreve özgü ince ayarın birleşiminin ve alana özgü kelime sözlüğünün öneminin altını çizmektedir. Özetle bu tez çalışması Türkçe medikal alanında dil modelleri geliştirmek için değerli bilgiler sunmaktadır ve klinik alandaki diğer düşük kaynaklı diller için ön eğitim teknikleri konusunda gelecekteki araştırmalara rehberlik niteliğinde bilgiler sağlamaktadır.

Anahtar sözcükler: doğal dil işleme, dil modeli, radyoloji, az-kaynaklı diller

ABSTRACT

CLASSIFICATION OF TURKISH CLINICAL NOTES USING DEEP LEARNING TECHNIQUES

TÜRKMEN, Hazal

Ph.D. Thesis in Department of Computer Engineering

Supervisor: Prof. Dr. Oğuz Dikenelli

July 2023, 87 pages

The significant advances made in the field of natural language processing (NLP) in recent years, together with the emergence of large language models (LMMs), have enabled revolutionary research and development. This revolutionary progress has been made even more evident by the possibilities offered by LLMs, especially when they are pre-trained and supported by in-domain corpora. While many studies have clearly demonstrated that such models are highly successful in biomedicine and clinical NLP tasks in English, unfortunately, the number of such studies in low-resource languages is very limited. At this point, this thesis aims to overcome the limitations of language resources and develop two different families of transfomer-based LLMs, BioBERTurk and TurkRADBERT language models. In order to improve the performance of the Turkish language biomedicine models developed for the first time in the literature, the effect of different language pre-training strategies was investigated. In order to evaluate the performance of the language models, three different datasets, designed and labeled by expert radiologists at the document level of Turkish head CT radiology reports, were created for the first time.The BioBERTurk family of language models performed better than the other Turkish models using the biomedical corpus training and continuous pre-training strategies, but the model combining radiology and biomedicine corpora and trained from scratch showed the lowest performance. TurkRADBERT, another family of language models developed in this thesis, investigated the effect of different pre-training methodologies on the performance of Turkish

clinical language models on a multi-label classification task involving radiology reports in the clinical domain with limited language resources in Turkish. The generic Turkish BERT model BERTurk and TurkRadBERT-task v1 showed the best overall performance thanks to the knowledge gained from the general domain corpus. The results underline the importance of the combination of general domain knowledge, task-specific fine-tuning and domain-specific vocabulary for optimal performance. In summary, this thesis provides valuable insights for developing language models in Turkish biomedicine and provides guidance for future research on pre-training techniques for other low-resource languages in the clinical domain.

Keywords: natural language processing, language model, radiology, low-resource language

ÖNSÖZ

Bu tezin ortaya çıkmasına katkıda bulunan pek çok kişiye büyük minnettarlık duyuyorum.

Her şeyden önce doktora yolculuğum boyunca danışmanım Prof.Dr. Oğuz Dikenelli'ye son derece minettarım. Kendisi geçtiğimiz altı yıl boyunca inanlımaz bir akıl hocası ve mantığın sesi oldu. Bu tez, yıllar boyunca danışman hocamın bana verdiği değerli geri bildirimleri, tavsiyeleri, cesaretlendirmesi ve motivasyonu olmadan mümkün olamazdı. Kendisine çok teşekkür ediyorum. Ayrıca araştırma ve tez çalışmasının şekillenmesinde değerli görüş ve önerileri ile katkıda bulunan jüri üyesi hocalarım Dr. Öğr. Üyesi Özgür Gümüş ve Dr. Öğr. Üyesi Selma Tekir'e de sonsuz teşekkürlerimi sunarım.

Ege Üniversitesi Tıp Fakültesi Radyoloji Bölümün'den Prof. Dr. Süha Süreyya Özbek, Prof. Dr. Cem Çallı ve Doç. Dr Cenk Erarslan hocalarıma verinin oluşturulmasındaki destekleri ve katkılarından dolayı ayrıca teşekkürlerimi bir borç bilirim.

Tez çalışmasında sağladıkları teknik destek ile uygulamanın gelişmesine katkı sunan Google TRC ekibine teşekkür ederim.

Hayatımın zor ve aynı zamanda harika bir dönemiydi. Ancak bunların hiçbirini ailemin desteği olmadan yapamazdım. Bana inandıkları için ve her zaman yanımda oldukları için aileme sonsuz teşekkürlerimi sunarım.

İZMİR

12/07/2023

Hazal TÜRKMEN

İÇİNDEKİLER

Sayfa

ŞEKİLLER DİZİNİ

Şekil **Sayfa**

TABLOLAR DİZİNİ

Tablo **Sayfa**

SİMGELER VE KISALTMALAR DİZİNİ

Simgeler	Açıklama
$BERT$	Çift Yönlü Kodlayıcı Temsilleri
$CBOW$	Sürekli Kelime Torbası(Continuous Bag-of-Words)
$LSTM$	Uzun Kısa Süreli Bellek Sinir Ağı(Long short-term memory)
RNN	Tekrarlayan Sinir Ağları(Recursive Neural Networks)

1 GİRİŞ

Günümüzde gelişmiş Yapay zeka (YZ) teknolojilerinin popülerliği gittikçe artmaktadır. Bu artışın en büyük sebebi insan beyninden esinlenerek tasarlanan YZ tekniklerinden derin öğrenme gibi algoritmaların kullanımı yaygınlaşarak tekrardan doğması gösterilmektedir. Bulunduğumuz teknoloji çağına tsunami etkisi yaratan derin öğrenme görüntü işleme, makine çevirisi ve otonom araçlar gibi birçok alana yenilikler getirmiştir. Dönüşümü gerçekleştireceği bir sonraki alanın ise sağlık hizmetleri olduğu düşünülmektedir(He et al., 2019). Medikal bilgi depolayan elektronik sağlık kayıtlarının (ESK) sağlık kurumlarındaki hızlı adaptasyonu bu dönüşümü tetikleyen en büyük etkendir. Öncelikle hasta bilgilerini arşivlemek gibi birçok sağlık görevlerinin yerine getirilmesi için tasarlanan ESK'lar, medikal bilginin artışıyla araştırmacılar tarafından derin öğrenme uygulamalarının ana kaynağı olarak kullanılmaya başlamıştır (Shah and Khan, 2020). Son araştırmalar, ESK verilerinin ikincil kullanımının, derin öğrenmenin gücünden yararlanarak sağlık hizmeti sağlayıcılarına daha doğru ve kişiselleştirilmiş klinik karar desteği sağlayabileceğini göstermiştir.

Elektronik sağlık kayıtları (ESK), bünyesinde barındırdığı çeşitli veri türleri ile sağlık hizmetlerinin etkin bir şekilde sunulmasına imkan sağlar. ESK'ların bir parçası olan radyoloji raporları, hastaların radyolojik görüntülemeleri (örneğin X-ışını, MRI veya CT taramaları) hakkında karmaşık tıbbi terminoloji ve detaylı bilgileri içeren serbest metinler şeklindedir. Bu metinlerin doğru ve hızlı bir biçimde analiz edilmesi, hastaların durumlarını doğru bir şekilde değerlendirebilmek ve etkili tedavi planları oluşturabilmek için hayati önem taşır. Fakat, radyoloji raporlarındaki serbest metinlerin düzensiz yapısı, metinlerin insanlar tarafından yapısal veriye dönüştürülmesini zorlaştırır ve bu süreç zaman alıcı ve maliyetli olabilir. İşte tam bu noktada, Doğal dil işleme (DDİ) ve dil modelleme gibi derin öğrenme teknikleri devreye girer ve radyolojik metinleri otomatik olarak analiz etme, hızlı bir şekilde yapısal veriye dönüştürme potansiyeli sunar. Ayrıca dil modellemesindeki son teknolojiler,

radyoloji raporlarının değerlendirilmesi sürecini insandan daha verimli ve etkin bir hale getirme kapasitesine sahip duruma gelmiştir.

Ne yazık ki, bu alandaki mevcut çalışmaların çoğu İngilizce metinlere yöneliktir ve Türkçe metinler üzerinde bu tür tekniklerin kullanılabilirliği hakkında yeterli araştırma bulunmamaktadır. Buradan yola çıkarak bu tez çalışması, Türkçe radyoloji raporlarının dil modelleme yöntemleri ile nasıl daha etkili bir şekilde analiz edilebileceğine odaklanmaktadır. Tez çalışmanın amacı, bu alanda bir boşluğu doldurmak ve Türkçe radyoloji raporlarından anlamlı bilgiler çıkarmada dil modelleme tekniklerinin etkinliğini incelemektir. Tezin bu bölümünde, çalışmanın motivasyonu ve arka planı ele alınmıştır. İkinci bölümde, araştırma probleminin detayları ve tez çalışmanın literatüre katkıları anlatılmıştır. Son olarak da, tezin genel yapısı ve ilerleyen bölümler hakkında bir bakış sunulmuştur.

1.1 Motivasyon

ESK'ların ikincil kullanımının temelinde, ESK'ların "gerçek dünya" hastalık ve bakım süreçlerini yakalayarak, geleneksel randomize kontrollü çalışmalarından (RKÇ) daha zengin ve daha genelleştirilebilir veriler sunması ve bu nedenle karşılaştırmalı etkililik araştırmaları için daha uygun bir kaynak olabilmesi yatmaktadır (Liu et al., 2012). Tıp biliminde RKÇ'lerden elde edilen bilgi, kontrol edilmemiş karıştırıcı faktörler için iyi tasarlanmış ve yürütülmüş olduklarında, güvenilirlik ve etkililik açısından altın standart kabul edilmektedir. Ancak bu koşullar sağlansa bile RKÇ bulgularının günlük uygulamada görülen hastaların türüne uyarlanması genellikle zordur. Bu durumun temel nedeni, randomize kontrollü bir çalışmanın yürütülmesi son derece maliyetli ve zaman alıcı olmasından dolayı seyrek çalışmalar olmasıdır. Ayrıca maliyetleri nedeniyle, RKÇ uygulamalarında sık sık ciddi dışlama kriterleri uygulanır ve yaş, ciddi engellilikler, gebelikler ve çoklu ilaç kullanımı nedeniyle riskli hastalar çalışmadan çıkarılır. Bundan dolayı RKÇ uygulamaları tüm klinik sorular için uygun olmayabilir ve klinik kararlar vermek için araştırma kanıt tabanın daralmasına neden olur (Sackett et al., 1996). Kanıt tabanının

daralması durumda da klinisyenler sıklıkla bilimsel tıp sonuçlarından daha az rehberlik alarak karar vermek zorunda kaldıkları çıkarımsal boşluk olarak adlandırılan durumlarla karşı karşıya kalırlar (Stewart et al., 2007). Büyük boyuttaki hasta kayıtlarının sistematik, güvenli bir şekilde saklayan Elektronik Sağlık Kayıtlarının yaygınlaşması kanıt tabanını genişleterek RKÇ bulgularına göre daha efektif ve bilimsel çalışmalar için daha genellenebilir verilerin sunulmasına olanak sağlamıştır. Bu nedenle, ESK'ların yaygın kullanımı, yönetimsel (örneğin, sigorta talepleri) veri kaynakları gibi diğer veri kaynaklarından daha kapsamlı bir teşhis, tedavi ve sonuçlar görüntülemesi sağlayabilir ve kliniseyenin karar sürecini hızlandırabilir. ESK kaynaklarından bu şekilde elde edilen bilginin keşfi, Uygulama Temelli Tıp olarak adlandırılmıştır ve verinin üstten aşağı bir süreçle sistematik bir şekilde oluşturulduğu Kanıta Dayalı Tıp (KDT) paradigmasını rutin bakımda bilgi keşfini ilk aşamaya yerleştirerek tersine çevirmiştir (Weng et al., 2012; Shah and Tenenbaum, 2012; Kanwal et al., 2022).

Klinisyenlerin karar verme sürecinde ESK'daki radyoloji raporları kritik bir rol oynamaktadır. Radyoloji, doğası gereği, radyoloji raporları, klinik notlar, tıbbi görüntülemeye bağlı notlar ve daha fazlasını içeren geniş bir metin verisi oluşturur. Bu tür metinlerin örnekleri arasında radyografik bulgular, Bilgisayarlı Tomografi (BT) taramaları için notlar ve Manyetik Rezonans Görüntüleme (MRI) raporları yer alır; tüm bunlar sofistike bir anlayış ve yorumlama gerektirmektedir. Örneğin, bir hekim bir hastanın durumunu değerlendirdiğinde, genellikle bir radyoloji raporunu referans alır. Bu rapor hastanın mevcut durumunu, olası teşhisleri ve tedavi seçeneklerini belirlemeye yardımcı olabilecek detaylı bilgiler sağlar. Ancak radyoloji raporlarının çözümlemesi ve anlaşılması genellikle karmaşıktır ve zaman alıcıdır. Çünkü bu raporlar genellikle serbest metin formatındadır ve tıbbi terminoloji ve karmaşık bilgilerle doludur. İşte burada, DDİ ve büyük dil modellerinin önemi ortaya çıkmaktadır.

Doğal Dil İşleme (DDİ), dilin anlaşılmasını ve işlenmesini otomatikleştiren bir bilgisayar bilimleri dalıdır. Bu alandaki teknikler serbest metin for-

matındaki radyoloji raporlarını anlamlandırmak ve anlamlı bilgilere dönüştürmek için kullanılabilir. Son yıllarda ise Doğal dil işleme alanında, dil modelleri önemli bir dönüşüm geçirerek serbest metinlerin anlamlı bilgilere dönüşüm sürecinde yepyeni bir dönem başlatmıştır. Özellikle radyoloji raporlarına odaklanıldığında bu raporların içerdiği detaylı ve genellikle karmaşık bilgiler gelişmiş dil modelleri ile yapılandırılabilir ve analiz edilebilir hale gelmiştir. Bununla birlikte dil modelleri ve DDİ tekniklerinin kullanılması bu raporların daha hızlı ve daha doğru bir şekilde analiz edilmesine olanak sağlayarak daha etkili ve bilinçli klinik kararların verilmesinde önemli bir rol oynamaktadır. Ayrıca gelişmiş dil modelleri klinesyenlere daha geneleştirebilir kanıt tabanı sağlanarak retrospektif çalışmaların yanı sıra, prospektif klinik uygulamalar ve araştırmalar için de değerli bir araç olabilmektedir. Ancak bu alandaki önemli dar boğazlardan birisi mevcut araştırmaların çoğu İngilizce tıbbi metinlere odaklanmıştır ve diğer dillerdeki klinik metinler için bu tür tekniklerin kullanımına yönelik araştırmalar ingilizceye göre çok azdır. Türkçe, zengin tıbbi bir kelime dağarcığına sahip geniş bir konuşma dili olmasına rağmen, Türkçe klinik metinler için DDİ ve derin öğrenme tekniklerinin kullanımı konusunda sınırlı araştırma yapılmıştır. Bu durum Türkçe anlayan medikal yapay zeka uygulamalarının geliştirilmesinin önündeki en önemli problemlerden biridir. Bu tezin motivasyonu, değerli ve genellikle yapılandırılmamış bilgiler içeren Türkçe rayoloji raporlarının otamatik analiz edilebilme ihtiyacına cevap vermektir. Derin dil modelleri ve DDİ tekniklerini kullanarak, tez Türkçe radyoloji raporları için bir sınıflandırmayı gerçekleyen bir derin dil modeli geliştirmeyi amaçlamaktadır. Bu model, sağlık profesyonellerine retrospektif çalışmalarda yardımcı olacak ve klinik karar verme konusunda kavramsal boşluğu daraltacaktır.

1.2 Araştırma Bağlamı

Tez çalışması Ege Üniversitesi Etik Kurulu tarafından “UH150040389” çalışma numarası ile onaylanmıştır. Çalışma retrospektif olarak, tümü ile kimliksiz (anonim) hale getirilmiş BT tetkiki rapor metinlerinin, EÜTF Hastanesi

içinde ve bilgisayar ortamında gerekli önlemler alınarak işlenmesi söz konusu olacağı için hasta onamı alınmamıştır. Çalışma ayrıca TPU Research Cloud programı (TRC) ve Google'ın CURe programı tarafından desteklenmiştir.

1.3 Problem tanımı ve zorluklar

Bu tezin ele aldığı problem, Türkçe radyoloji raporlarının otomatik olarak analiz edilmesi ihtiyacıdır. Sağlık kurumları tarafından üretilen hasta raporları yapılandırılmamış doğası ve büyük hacmi, bunları manuel olarak analiz etmeye engel olmaktadır. Bu nedenle, hasta raporlarını otomatik olarak analiz edebilen ve ilgili bilgileri çıkaran bir sınıflayıcı dil modeline ihtiyaç vardır. Aşağıda çalışma alanının zorluklarından bahsedilmiştir.

1.3.1 Radyoloji Raporları

Elektronik Sağlık Kayıtları (ESK), sağlık hizmetlerindeki tüm hasta bilgilerini içerir. Bu veri çeşitliliği, demografik bilgilerden laboratuvar sonuçlarına, klinik notlardan radyoloji raporlarına kadar geniş bir yelpazeyi kapsar. Radyoloji raporları, ESK sistemi içerisinde önemli bir bileşeni oluşturur ve genellikle hasta hakkında detaylı ve spesifik bilgiler sağlar. Radyoloji raporları, radyolojik görüntüleme sonuçlarını ve bu görüntülerin profesyonel yorumlarını içerir. Raporlar genellikle hastanın vücudunun belirli bir bölgesini görselleştirir ve varsa anormallikleri belirtir (Dunnick and Langlotz, 2008). Buna ek olarak, hekimlerin hastanın sağlık durumu hakkında daha bilinçli kararlar almasını sağlayan önemli bir araçtır. Örneğin, bir radyolog, bir MR görüntüsünde görünen bir tümörün boyutunu ve yerleşimini belirtir, bu bilgiler daha sonra onkolog tarafından tedavi planını oluşturmak için kullanılır(Shamshad et al., 2023). Bu tez çalışması kapsamında Ege Üniversitesi Hastanesi Radyoloji Bölümünden alınan hasta raporlarının işlenmesi ve Türkçe DDİ teknolojileri ile analiz edilmesi ele alınmıştır. Yapılan çalışma retrospektif olarak tümü ile kimliksizleştirilmiş hasta metinlerinden oluşturuğu için hasta gizlilik ve güvenlik endişesi bulunmamaktadır. Ancak hasta mahremiyetinin ve etik standartlar tarafından verilerin kimliği korunduğu için açık olarak paylaşılmamıştır. Verinin

detaylı açıklaması tezin ikinci bölümünde yapılmaktadır. Hasta raporlarının DDİ uygulamalarında kullanılması sırasında karşılaşılan zorluklar aşağıda sıralanmıştır.

- Veri kalitesi ve tutarlılık: Tıbbi metinler eksik, tutarsız olabilir ve derin öğrenme modellerinin performansını etkileyebilecek hatalar içerebilir. Bu hatalar klinik notların doğru bir şekilde sınıflandırılmasını zorlaştıran yazım hataları, standart olmayan kısaltmalar ve tekrarlayan veriler şeklinde sıralanabilir.
- Alana özgü dil: Klinik notlar genellikle, derin öğrenme modellerinin anlaması zor olabilecek alana özgü dil ve tıbbi jargon içerir. Ek olarak, Türkçe klinik dili, klinik terimler için standart bir kelime dağarcığı geliştirmeyi zorlaştırabilen karmaşık bir sözdizimine sahiptir.
- Dengesiz veri dağılımı: Raporlarda belirli klinik not kategorileri diğerlerinden daha yaygın olabilir ve bu da dengesiz veri dağılımına neden olabilir. Bu özellikle nadir kategoriler için derin dil modellerinin performansının ve doğruluğunun düşmesine yol açabilmektedir.

Tez çalışmasında Türkçe klinik metinlerin işlenmesindeki yukarıda verilen sorunlara yönelik alınan önlemler ve çözüm önerilerine değinilmiştir. Bu kapsamda kullanılacak olan serbest metinlerin Türkçe dilinde olması teze ulusal düzeyde katkı sağlanmış ve Türkiye'de klinik doğal dil işleme çalışmak isteyen araştırmacılar için yol gösterici bir nitelik katmıştır.

1.3.2 Türkçe Klinik Metinlerin Dil modelleri ile Sınıflandırılması

Klinik DDİ teknikleri, yapılandırılmamış klinik metin verilerinden anlamlı bilgiler çıkarmak için kullanılır. Metin sınıflandırması da, hasta raporları gibi klinik metin belgelerinin içeriklerine göre otomatik olarak sınıflandırılmasını sağladığından, tıp alanında DDİ'nin yaygın bir uygulamasıdır. Son zamanlarda dil modelleri, derin öğrenme teknikleri kullanılarak büyük miktarda dil verisinden anlam öğrenme konusunda büyük başarılar elde etmiştir.
Birçok mevcut derin öğrenme modeli denetimlidir ve dolayısıyla eğitim için

etiketli veri gerektirir. Hasta metinlerinin uzman tarafından etiketlenmesi, görevin bilişsel karmaşıklığı ve veri kalitesindeki değişkenlik nedeniyle zorlu bir iştir. Temelinde sinir ağları olan derin öğrenme modellerini eğitmek için büyük miktarda metin gerekmektedir ve ne yazık ki etiketli hasta verileri genellikle yetersizdir. Ayrıca, bu alandaki araştırmacılara açık etiketli birçok veri İngilizcedir (Névéol et al., 2018; Boudjellal et al., 2021; Schneider et al., 2020; Barash et al., 2020) ve Türkçe gibi diğer diller göz ardı edilmektedir. Derin öğrenme modeli olan BERT tabanlı transformer dil modellerinin (Devlin et al., 2018) büyük bir metin verisinde önceden eğitmek, modelin daha az miktarda etiketli veri kullanarak belirli klinik uygulamaları için ince ayar yapılabilen genel dil kalıplarını ve temsillerini öğrenmesine izin verdiğinden büyük etiketli veri kümesine olan ihtiyacı hafifletmeye yardımcı olabilmektedir. Tez çalışmasında Türkçe dilinde yazılmış olan klinik metinlerin sınıflandırmayı sağlayan transformer dil modelinin oluşturulması amaçlanmıştır. Literatürde Türkçe diline ait klinik alanda etiketli veri kümesi olmamasından dolayı Ege Üniversitesi Hastanesi'den hasta raporları üzerinden metin sınıflama görevi oluşturulmuştur. Bir diğer zorluk da ince ayar yöntemi için Türkçe dilinde açık büyük boyutta biyomedikal ve klinik metin derlemleri olmamasından kaynaklanan dil modellerinin eksiğidir. Buradan yola çıkarak tez kapsamında ilk Türkçe biyomedikal ve klinik metin derlemleri oluşturularak alana özgü transformer dil modelleri oluşturulmuştur. Bu açıdan tez yenilikçi yöntemleri kapsayarak alanındaki öncü çalışmadır.

1.4 Tezin katkıları

Tezin literatüre katkıları aşağıdaki gibi özetlenebilir:

- Literatürde ilk kez, Türkçe medikal alana özgü makale metinlerinden oluşan bir derlem oluşturulmuş ve bu metinler üzerinde eğitim vererek BERT dil modelleri geliştirilmiştir. Modeller ve metin derlemi, bilimsel topluluğun genel faydasına olacak şekilde Github üzerinde araştırmacılara olarak sunulmuştur. Tez çalışması, bu açıdan Türkçe

medikal dil işleme alanında önemli bir ilerlemeyi temsil etmektedir ve diğer araştırmacıların kendi projelerinde kullanabilecekleri değerli bir kaynak sağlamaktadır.

- Türkçe klinik dilinde bir DDİ görevinin eksikliğinden dolayı performans değerlendirmesi ve dil modellerinin karşılaştırılması için Türkçe kafa BT raporlarından oluşan çok sınıflı bir sınıflama görevi geliştirilmiştir. Bu görev, radyoloji uzmanları ile birlikte tasarlanıp uygulanmıştır. Raporlar, bulgular ve izlenimler olarak iki farklı bölüme ayrılmış ve her biri "normal", "anormal" ve "seri dışı" sınıflarına göre etiketlenmiştir. Ayrıca yapılan çalışma, Türkçe dilinde klinik DDİ görevi oluşturan ilk çalışma olması bakımından önemlidir.
- Biyomedikal metin derleminin yanında ilk defa ufak boyutta Ege Üniversitesi Hastanesi'nin Türkçe kafa BT radyoloji raporlarından oluşan ve dil modelleri için kaynak sağlayan görevle ilgili küçük derlem oluşturulmuştur.
- Tez çalışmasında Türkçe klinik alanında çok etiketli metin sınıflandırma çalışmalarındaki mevcut bir boşluğu ele alınarak, klinik dil modellerini değerlendirmek için çok etiketli bir klinik sınıflandırma görevinin radyoloji uzmanlarıyla birlikte tasarımı ve geliştirilmesi sağlanmıştır.
- Sınırlı klinik görev radyolojisi verilerinden yararlanarak eşzamanlı ön eğitimin yeni bir araştırması yapılmış ve bu yaklaşımın sınırlı Türkçe klinik radyoloji verileri bağlamında kullanımına yeni bir bakış açısı sunulmuştur.
- Eşzamanlı ön eğitim yöntemi, Google'ın özel yüksek hızlı ağ arayüzleriyle bağlanan dağıtık mimarideki TPU cihazların (Google Cloud TPU-pod) üzerinde başarılı bir şekilde uygulanarak, bu yöntemin açık kaynaklı ilk uygulaması gerçekleştirilmiştir. Bu açıdan tezdeki uygulama teknik bir katkı sunmuştur.
- Oluşturulan görev odaklı Türkçe klinik derlem ile TurkRadBERT-task v1, TurkRadBERT-task v2, TurkRadBERT-sim v1 ve TurkRadBERT-sim v2 olmak üzere dört farklı dil modeli geliştirilmiştir. Farklı ön eğitim

tekniklerinin Türkçe klinik dil modellerinin çok etiketli bir sınıflandırma görevindeki performansı üzerindeki etkisi incelenmiştir ve sınırlı dil modeli kaynağı durumunda modellerin nasıl eğitileceği konusunda geniş çapta karşılaştırmalar yapılmıştır. Yapılan çalışmalar böylelikle düşük kaynaklı veri durumunda veya inglizce dışında başka dillerde çalışmak isteyen araştırmacılar için bir yol planı niteliğini taşımaktadır.

1.5 Tez çalışmasında yapılan yayınlar

- Türkmen, H., Dikenelli, O., Eraslan, C., Çalli, M. C., & Ozbek, S. S. (2022, June). Developing Pretrained Language Models for Turkish Biomedical Domain. In 2022 IEEE 10th International Conference on Healthcare Informatics (ICHI) (pp. 597-598). IEEE.
- Türkmen, H., Dikenelli, O., Eraslan, C., Çallı, M. C., & Özbek, S. S. (2023). Harnessing the Power of BERT in the Turkish Clinical Domain: Pretraining Approaches for Limited Data Scenarios. arXiv preprint arXiv:2305.03788. (Clinical NLP workshop kabul edildi)
- Turkmen, H., Dikenelli, O., Eraslan, C., & Callı, M. C. (2022). BioBERTurk: Exploring Turkish Biomedical Language Model Development Strategies in Low Resource Setting. Journal of Healthcare Informatics Research (Kabul edildi)

1.6 Tezin organizasyonu

Bu araştırmanın motivasyon ve amacı önerilen yöntemler ve yaklaşımlar ile tezin katkısı ilk bölümde tanıtılmıştır. Tezin geri kalan bölümleri aşağıdaki gibi organize edilmiştir.

Bölüm 2, ikili ve çoklu etiketler üzerine genel bir bakış ve detaylı istatistikleri içermektedir. Bu bölüm, tez kapsamında ele alınan sınıflandırma problemlerinin niteliklerini derinlemesine sunmaktadır. Özellikle sınıflama için oluşturulan veri kümelerinin gelişim sürecine dair ayrıntılı bilgi sağlamaktadır. Ek olarak, veri kümelerinin etiketlerinin özet istatistiklerini ve metin ön işleme

süreçlerini kapsamlı bir şekilde ele almakatdır. Bu bölüm bu tezin daha sonraki bölümlerinde kullanılan yöntem ve tekniklere dair temel bilgileri okuyucuya sağlamak amacıyla hazırlanmıştır.

Bölüm 3, dil modelleri ve sinir ağlarının Türkçe BT raporlarının sınıflandır ma problemlerini ele alma bağlamındaki temel teorik altyapısını sunmaktadır. Bu bölümde, seçilen deneysel metodolojinin ve değerlendirme ölçütlerinin detaylarına da değinilmektedir. Radyoloji metinlerinin sınıflandırılmasındaki alana özgü dil modellerinin literatür taraması bu bölümde gerçekleştirilirken, radyoloji alanına özgü biyomedikal doğal dil işleme uygulamaları da dahil olmak üzere sınıflandırma tekniklerini kapsayan geniş bir literatür incelemesi de sunulmaktadır. Bu bölüm, çalışmanın metodolojisinin ve araştırma çerçevesinin anlaşılmasına yönelik genel bir zemin oluşturmayı hedeflemektedir.

Bölüm 4, Türkçe'nin sınırlı kaynağa sahip olduğu bir dil bağlamında, bir dil modelinin nasıl oluşturulacağına dair genel bir çerçeve sunar ve BERT dil modeli geliştirme sürecinin aşamaları ve bu aşamaların nasıl bir deneysel yapı içinde uygulanabileceğine odaklanmaktadır. Bu genel çerçeve sunulduktan sonra, tez kapsamında geliştirilen dil modellerinin deneysel süreçlerin pratikte nasıl uygulandığına dair genel bir bakış verilmiştir. Bu bölüm bir sonraki iki bölüm ile ilgili genel bir çerçeve sunmaktadır.

Bölüm 5, tez kapsamında geliştirilen Türkçe biyomedikal alanda BERT tabanlı dil modellerinin üzerine yoğunlaşmaktadır. Bu bölüm, ön eğitim yaklaşımları ve ön eğitim için kullanılacak derlem seçimi konularında kritik kararların alınmasının derinlemesine incelenmesiyle bu alandaki bilgiyi genişletmektedir. Ayrıca, bu dil bağlamında ilk kez gerçekleştirilen ve Türkçe biyomedikal alana özgü dört adet dil modelinin oluşturulması süreci detaylarıyla sunulmaktadır. Model performansının değerlendirilmesi için oluşturulan etiketli veri kümeleri, kafa BT radyoloji raporlarını sınıflandırmak amacıyla kullanılmıştır. Son olarak bölüm farklı ön eğitim yaklaşımlarının ve ufak boyuttaki görev özgü derlemlerin model performansı üzerindeki etkilerini değerlendiren kapsamlı bir inceleme sunmaktadır.

Bölüm 6, Türkçe için özel olarak uyarlanmış klinik dil modellerinin geliştirilmesi ve değerlendirilmesini araştırmaktadır. Radyoloji raporlarını içeren çok etiketli bir sınıflandırma görevine odaklanarak, çeşitli ön eğitim metodolojilerinin bu modeller üzerindeki etkisi üzerine kapsamlı bir çalışma sunmaktadır. Bölüm, kısıtlı dil kaynaklarının getirdiği sınırlamaların ve zorluğun üstesinden gelmek için yeni çözümler sunmaktadır. Yeni geliştirilen dört model tanıtılmakta ve performansları karşılaştırılmakta, alana özgü kelime sözlüğünün ve genel alan bilgisi ile göreve özgü ince ayar arasındaki dengenin önemi vurgulanmaktadır. Bu bölüm, özellikle klinik alandaki düşük kaynaklı diller için DDİ alanında gelecekteki araştırmalar için değerli bilgiler ve rehberlik sağlamaktadır.

Tezin son bölümünde ise tez özetinin yanında gelecek çalışmalar hakkında planlar ve öneriler yapılmıştır. Tezin alana verdiği katkılar ve karşılaştığı kısıtlamalar bu bölümde tartışılmıştır.

2 VERİ ve SINIF

Bu bölüm, tez boyunca geliştirilen Türkçe alana özgü dil modeli performans analizleri için oluşturulan etiketli veri kümelerinin detaylarını ayrıntılı bir şekilde ele almaktadır. Çalışmada çok sınıflı ve çok etiketli sınıflandırma hakkında kapsamlı bilgi verilirken, veri ve etiketlerin genel özet istatistikleri de sunulmaktadır. Etiketli veri için kullanılan Türkçe radyoloji raporlarının ön işlemesinde uygulanan adımlar da bu bölümde özetlenmiştir. Bu sayede, tez çalışmasının veri ve analiz yöntemleri hakkında kapsamlı bir anlayış sunulmuştur.

2.1 Türkçe Radyoloji Raporları

ESK verileri, hem yapılandırılmış (ör. laboratuvar değerleri yaşamsal belirtiler) hem de yapılandırılmamış (ör. klinik notlar, radyoloji raporları) metin öğelerinden oluşur. Bu yapılandırılmamış veriler, otomatik metin analiz sistemleri geliştirilirse birçok amaç için kullanılabilecek zengin bilgiler içermektedir (Shin et al., 2017).

Tez kapsamında otomatik olarak metinlerin analizini gerçekleyen modeller için Ege Üniversitesi Hastanesi ESK sisteminden nöroloji ve acil servisindeki Ocak 2018 ile Haziran 2018 tarihleri arasına ait bilgisayarlı tomografi (BT) tetkikleri (8 >) için 40.306 doğrulanmış Türkçe radyoloji raporları kullanımıştır. Toplanan radyoloji raporları bulgular ve izlenim olmak üzere iki kısımdan oluşmaktadır. Bir radyoloji raporunun en uzun bölümü olan bulgular kısmı, radyoloğun çalışmada incelenen her bir anatomi parçasını taramada gördüklerine ilişkin notlarla birlikte listelediği yerdir. Taranan bölgenin normal mi, anormal mi yoksa potansiyel olarak anormal mi olduğu bilgileri yer almaktadır. İzlenim bölümünde ise radyolog bulguları özetler ve gördüğü en önemli bulguları ve bu bulguların olası nedenlerini bildirir. Bu bölüm karar vermek için en önemli bilgileri sunmaktadır (Hartung et al., 2020). Tez çalışmasında radyoloji raporlarının sınıflandırma problemindeki ilk durum çalışmasında bulgular ve izlenim için iki farklı veri kümesi oluşturulup, uzun

ama belirsizliği daha fazla ifadeleri içeren bulguların mı yoksa, kısa ama bilgilerin süzülüp özetlendiği izlenimin kullanımının metin analizinde daha başarılı olacağının sorusu araştırılmıştır. 12 farklı bulgunun etiketlendiği ikinci durum çalışmasında ise bulugular ve izlenim olarak bölünmeden bir radyoloji raporunun tümü kullanılmıştır. Durum çalışmaları ve veri kümelerinin detaylı analizi aşağıda verilmiştir. Veri kümeleri oluşturulmadan önce tüm radyoloji raporları ortak bir metin ön işleme sürecinden geçirilmiştir. Ön işlemede gerçeklenen adımlar aşağıda maddeler halinde özetlenmiştir.

- 100 karakterden az cümleler az bilgi içerdiğinden filterelenmiştir.
- Fazlalık boşluk karakterleri, radyoloji alanındaki özel kodlamalar filterlenmiştir.
- Metin de varsa özel isimler anonimleştirilmiştir.
- Tekrarlayan raporlar tekilleştirilmiştir.

Daha sonra ön işlemeden geçirilen hasta raporları iki farklı durum çalışması için radyologlar tarafından etiketlenmiştir.

2.2 Genel Bakış: Durum çalışmaları

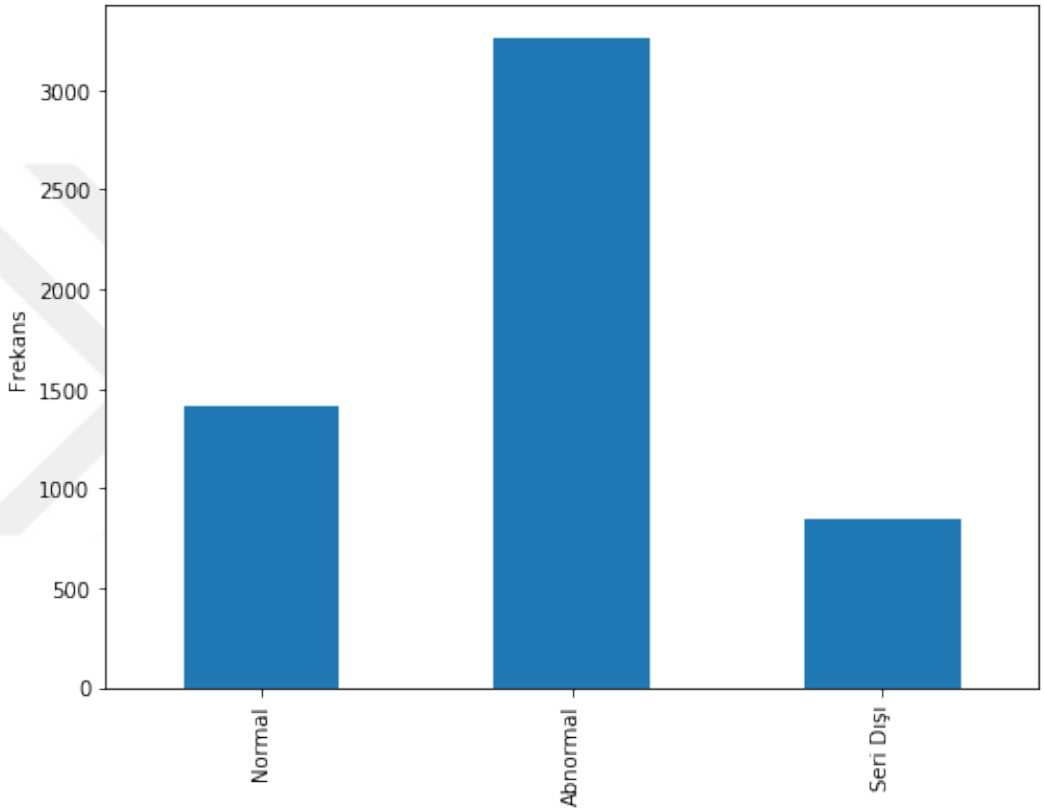
Çalışma Ege Üniversitesi hastanesinin nöroloji ve acil servsi bölümlerinde yaygın olarak geçen tıbbi bulguların çok sınıf ile ve çok etiketli sınıflandırmasını ele almaktadır. Bu bölüm, bu iki bulgu çalışmasına genel bir bakış sunar.

2.2.1 Türkçe kafa BT raporlarının sınıflandırılması

Acil servis (AS), sağlık hizmeti ortamlarındaki en aşırı kalabalık ve dinamik ortamlardan biridir (Carter et al., 2014). Bu kalabalıklılığı önemli bir faktörü de hızlı hasta teşhisinin temel bir bileşeni olan kontrastsız kafa bilgisayarlı tomografisinin (BT), acil serviste en sık yapılan BT taraması olmasıdır (Johnson and Winkelman, 2011; Ginde et al., 2008). Son yirmi yılda, acil servis ortamında BT kullanım oranları istikrarlı bir şekilde artmıştır (Raja et al., 2014). Örneğin yapılan bir çalışmaya göre BT cihazı başına çekim sayısı OECD

ortalaması 6.806 iken, bu rakam Türkiye'de 13.993 olarak Japonya'dan sonra en yüksek ikinci sıradaki ülkedir (DURMUŞ and GÜNEYSU, 2020). Kafa BT'si kanama, tümörler ve iskemi (başlangıç zamanına bağlı olarak) gibi acil intrakraniyal patolojileri hızlı bir şekilde teşhis edebilir ve böylece acil servis hasta bakımının desteklenmesine yardımcı olabilmektedir. Ancak yüksek maliyetli ve radyasyonlu bir tetkik olması sebebiyle de BT tetkikinin yapılamsı konusunda mümkün olduğunca titiz davranılması ayrı bir önem arz etmektedir (DURMUŞ and GÜNEYSU, 2020). Acil servislerdeki stresli ortamda yüksek kaliteli ve verimli bakım sağlamak için hastaların BT çekimindeki iş hacmini optimize eden, triyajda gerekli durumları önceliklendiren çeşitli DDİ uygulamaları mevcuttur. Bu tez çalışmasında Ege Üniversitesi Hastanesi nöroloji ve acil servis bölümlerinden Türkçe BT tetkit raporlarının retrospektif olarak sınıflayan DDİ modelinin geliştirilmesi amaçlanmıştır. Modelin geliştirilmesinde İki kullanım durumunu incelenmiştir: (1) raporların normal, abnormal (patolojik) ve seri dışı olarak sınıflandırıldığı genel etiketleme kullanım durumu; (2) raporların 12 farklı bulgu ile etiketlendiği belirli etiketleme kullanım durumudur. **Genel sınıflama kullanım durumu:** Genel sınıflama durumunda ön işlemeden geçirilen radyoloji raporları tekrardan 300 karakterden büyük olan raporlar filtrelenerek 5533 rapor etiketleme sürecinde kullanılmıştır. Raporların normal, abnormal (patolojik) ve seri dışı olarak etiketleme süreci, radyoloji raporlamasında uzun yıllara dayanan deneyime sahip üç radyolog (C.E, M.C.C¸ ve S.S.O) tarafından üç aşamada gerçekleştirilmiştir. Her aşamada, önce iki etiketleyici (C.E, M.C.C) raporların bir alt kümesini bağımsız olarak etiketledikten sonra üçüncü etiketleyici, çelişkili olanları belirlemek için bu etiketleri kontrol etmiştir. Her aşamanın sonunda, üç etiketleyici de tamamen üzerinde anlaşmaya varılan etiketler oluşturarak bir fikir birliği oluşturmuştur. Etiketleme görevi etiketleyicilerin işini kolaylaştırmak için bir elektronik tablo dosyasında uygulanmıştır. Ardından, etiketli veri kümesinin son hali, ayrı ayrı değerlendirmek için izlenimler ve bulgular olmak üzere iki veri kümesine ayrılmıştır. İzlenim veri kümesinde 28704 cümle ve 13892 sözcük vardır; Öte yandan, bulgu veri kümesi 78939 cümle ve 17348 sözcük içermektedir.

Bir raporun bulgular bölümü genellikle, daha uzun metin içeren modellerin performansını gözlemlenmesini sağlayan daha uzun metinler içermektedir. Bütün veri kümeleri daha sonra ince ayar için test (%10), doğrulama (%10) ve eğitim (%80) olarak rastgele ayrılmıştır. İki veri kümesinin sınıf dağılımları aynıdır ve bu dağılım Şekil 2.1'de gösterilmektedir. Şekil 2.1'de görüldüğü gibi, radyoloji metin işlemede veri kümelerinde yaygın görülen dengesiz bir dağılımı vardır.



Şekil 2.1: Genel sınıflama durumunun etiket dağılımı

Özel sınıflama kullanım durumu: Özel sınıflama kullanım durumu için rastgele seçilen 2000 Türkçe kafa BT raporunu kullanarak çok etiketli sınıflandırma görevi geliştirilmiştir. Veri kümesinde 20618 cümle ve 249072 sözcük bulunmaktadır. Bu sınıflandırma görevinin amacı, serbest metin formatında sunulan bir radyoloji raporunda klinik açıdan önemli 12 gözlemlerin varlığını tespit etmektir. Bunlar 'İntraventriküler', 'Gliosis', 'Epi-

dural', 'Hidrosefali', 'Ensefalomalazi', 'Kronik iskemik değişiklikler', 'Lakuna', 'Lökoaraiosis', 'Mega sisterna magna', 'Meningioma', 'Subaraknoid Kanama', 'Subdural', 'Bulgu Yok' eitketleridir. Sınıflandırma süreci, rapordaki cümlelerin gözden geçirilmesini ve olumlu ya da olumsuz olmak üzere iki sınıftan birine dahil edilmesini içerir. 13. gözlem olan "Bulgu Yok", herhangi bir bulgunun olmadığını göstermektedir. Radyoloji uzmanları veri kümesini bu açıklama şemasına göre etiketlemiştir ve etiketleme süreci genel kullanım durumunda izlenilen süreç ile aynıdır. Etiketleme süreci üç deneyimli radyologun (C.E, M.C.C ve S.S.O) katılımıyla üç aşamada gerçekleştirilmiştir. Her aşamada, iki etiketleyici (C.E, M.C.C) bağımsız olarak raporların bir kısmını etiketlemiştir. Daha sonra, üçüncü etiketleyici herhangi bir tutarsızlığı tespit etmek için bu etiketleri incelemiştir. Her aşamanın sonunda, üç etiketleyici de üzerinde anlaşmaya varılan etiketleri oluşturarak bir fikir birliğine varmıştır. Etiketlenen veri kümeleri daha sonra ince ayar için rastgele test (%10), doğrulama (%10) ve eğitim (%80) kümelerine bölünmüştür. Tablo 2.1'da gösterildiği gibi sınıf dağılımları, veri kümelerindeki farklı kategorilerin değişen yaygınlığını göstermektedir. Veri kümeleri, radyoloji alanındaki metin işlemenin tipik bir özelliği olan dengesiz bir dağılım sergilemektedir (Qu et al., 2020).

Tablo 2.1: Veri kümesindeki her etiket için pozitif ve negatif frekansların dağılımı

Kategori	**Pozitif**	**Negatif**
Intraventriküler	22 (%1.1)	1978 (%98.9)
Gliozis	54 (%2.7)	1946 (%97.3)
Epidural	51 (%2.55)	1949 (%97.45)
Hidrosefali	70 (%3.5)	1930 (%96.5)
Ensefalomalasi	177 (%8.85)	1823 (%91.15)
Kronik İskemik Değişiklikler	951 (%47.55)	1049 (%52.45)
Lakuna	138 (%6.9)	1862 (%93.1)
Lökoarayoz	49 (%2.45)	1951 (%97.55)
Mega Sisterna Magna	15 (%0.75)	1985 (%99.25)
Meningiom	39 (%1.95)	1961 (%98.05)
SAK	209 (%10.45)	1791 (%89.55)
Subdural	227 (%11.35)	1773 (%88.65)
Bulgu Yok	299 (%14.95)	1701 (%85.05)

2.3 İstatistik: Metin Sınıflandırma problemi

Bu bölümde bahsedildiği gibi tezde iki farklı sınıflama problemi ele alınmıştır. Bunlar çok sınıflı sınıflama ve çok etiketli sınıflamadır.

Sınıflama problemi için, Tsoumakas ve ark. (Tsoumakas et al., 2010) tarafından belirtilen gösterimler kullanılırsa:

- L = λ j : j = 1...q, sonlu etiket kümesine atıfta bulunur
- D = $(x_i, Y_i), i = 1...m$, çok-etiketli eğitim örneklerinin kümesine atıfta bulunur; burada x_i özellik vektörüdür ve $Y_i \subseteq L$, $i - nci$ örneğinin etiket kümesidir.

Not: $\subseteq$ sembolü, bir kümenin diğer bir kümenin alt kümesi olduğunu gösterir.

Etiket kardinalitesi (LCard), bir veri kümesindeki örneklerin etiket sayısının ortalamasıdır:

$$LCard = \frac{1}{m}\sum_{i=1}^{m}|Y_i|$$

burada m, veri kümesindeki toplam örnek sayısıdır ve $|Y_i|$, veri kümesindeki i-inci örneğe atanan etiket sayısını gösterir. Bu formül, veri kümesindeki örnek başına ortalama etiket sayısını hesaplar ve etiket kardinalitesi olarak bilinir.

Etiket Yoğunluğu (LDens), şu şekildedir:

$$LDens = \frac{LCard}{q}$$

Burada LCard, veri kümesindeki örneklerin etiket sayılarının ortalamasıdır. Denklem, veri kümesindeki etiket yoğunluğunu hesaplamak için kullanılır. Tek sınıflı sınıflandırma için LCard her zaman 1 olacak ve çok etiketli sınıflama için LCard > 1 olacaktır. Bu tez çalışmasında genel kullanım durumundaki veri kümesinin tek sınıflı bir problem olduğundan dolayı LCard 1 iken, özel kullanım durumundaki veri kümesi çok etiketli bir problem ve LCard'ı 1.15 olarak hesaplanmıştır. LCard 1.15, her örneğin kendisiyle ilişkilendirilmiş ortalama 1'den biraz fazla etikete sahip olduğunu göstermektedir.

3 ARKAPLAN VE İLGİLİ ÇALIŞMALAR

Bu bölümde, tezdeki tüm modellerin yapı taşını oluşturan temel kavramlar ve araştırma eğilimleri tanıtılmaktadır.

İlk olarak metin sınıflandırma alanındaki sınıflandırma teknikleri ili iligli kısa bir genel bakış verilecektir. Sonraki bölümü oluşturan dil modelleri ise tez boyunca büyük bir rol oynadığında dolayı ayrı bir şekilde gruplandırılıp sunulmuştur. Daha sonra dil modelleri ile yürütülen literatürdeki deneysel metodolojiden ve en yaygın kullanılan değerlendirme metriklerinden basedilmiştir. Son olarak radyoloji alanındaki doğal dil işleme çalışmaları sunularak bölüm sonlandırılmıştır.

3.1 Metin Sınıflandırma problemi

Metin sınıflandırması, metin örneklerinin önceden tanımlanmış sınıflara kategorize edilmesiyle ilgili bilgi erişimi çalışma alanıdır. Her bir örneğe bu sınıfların veya etiketlerin nasıl atandığına bağlı olarak, üç ana metin sınıflandırma türü vardır: tek sınıflı, çok sınıflı ve çok etiketli.

Tek sınıflı sınıflandırma, bilinen sınıf etiketleri L'den tek bir sınıf etiketi λ ile ilişkilendirilmiş örnek kümesinden öğrenme işlemidir, burada $|L| > 1$'dir. İkili sınıflandırma durumunda, sınıf etiketleri $|L| = 2$'dir. Çok sınıflı sınıflandırmada tıpkı ikili sınıflandırmada olduğu gibi her veri kümesinin yalnızca bir çıkış özelliği vardır. Bununla birlikte çok sınıflı sınıflandırmada ikili sınıflandırmada görüldüğü gibi yalnızca 1 ve 0'ın aksine, her örnek sonlu ve ayrık bir etiket kümesinden herhangi bir değer alabilir bu durumda $|L| > 2$'dir. Çok-etiketli sınıflandırma probleminde ise, örnekler $Y \subseteq L$ etiket kümesiyle ilişkilendirilir. Tek-sınıflı ve çok-sınıflı problemler arasındaki fark aşağıda gösterilmiştir. Not: Burada $\subseteq$ sembolü, Y kümesinin L kümesinin bir alt kümesi olduğunu ifade eder. Tablo 3.1 ve Tablo'de 3.2 sırasıyla çok sınıflı sınıflama ve çok etiketli sınıflamaya örnektir. Çok sınıflı sınıflama için (soldaki) $Y_i \in \{0, 1\}$, çok etiketli sınıflama için (sağdaki) $Y_i \subseteq L, Y_i \in \{0, 1\}$ olarak kabul edilmektedir.

Tablo 3.1: Çok sınıflı sınıflama örneği

X_1	X_2	X_3	X_4	Y
1	0.3	1	0	Y_1
0	0.5	0	0.6	Y_3
1	1	0.6	1	Y_3
0.5	0.4	0.2	1	Y_2
0	1	1	0	Y_1

Tablo 3.2: Çok etiketli sınıflama örneği

X_1	X_2	X_3	X_4	Y_1	Y_2	Y_3
1	0.3	1	0	1	0	1
0	0.5	0	0.6	1	1	0
1	1	0.6	1	0	1	1
0.5	0.4	0.2	1	1	0	1
0	1	1	0	1	1	1

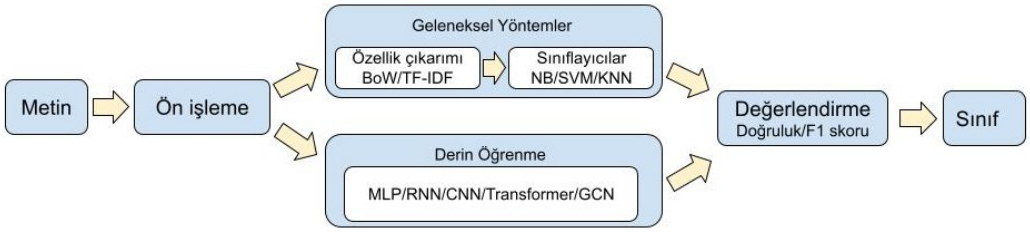
3.2 Metin sınıflandırmada Derin Öğrenme

Doğal Dil İşleme (DDİ), insan-bilgisayar iletişimini sağlayan bir Yapay Zeka (YZ) alanıdır. Daha net bir tanımla DDİ, insan dilini anlamak ve oluşturmak için araçlar üretmeyi amaçlar. Uzun süredir çalışılan ancak son zamanlarda hızla popülerliği artan DDİ alanında insan hayatını kolaylaştırmak için çeşitli uygulamalar geliştirilmiştir. Günümüzde makine çevirisinden (ör. Google Çeviri, DeepL, vb.), Diyalog Sistemleri (ör. Siri, Alexa vb.), metin özetleme (Allahyari et al., 2017; Widyassari et al., 2022) ve elektronik sağlık kayıtlarından bilgi çıkarımına (Hsu et al., 2022) kadar birçok teknolojide DDİ yaklaşımları kullanılmaktadır. Metin sınıflandırma ise kısacası - metne önceden belirlenmiş etiketleri atama işlemi - birçok DDİ uygulamasında önemli ve temel bir görevdir.

Yaşadığımız bilgi patlamasıyla birlikte büyük miktardaki metin verilerini manuel olarak işlemek ve sınıflandırmak zaman alıcı ve zorlu bir iş haline gelmiştir. Bunun yanı sıra, manuel metin sınıflandırmasının doğruluğu maliyet ve uzmanlık gibi insan faktörleri tarafından kolayca olumusuz etkilenebilmektedir. Daha güvenilir ve daha az öznel sonuçlar elde etmek için makine öğrenimi yöntemlerini kullanarak metin sınıflandırma işlemini otomatikleştirilmesi gerekmektedir.

Figür 3.1, geleneksel ve derin analiz ışığında metin sınıflandırmasında yer alan prosedürlerin akış şemasını göstermektedir. Metin verileri sayısal, görüntü veya sinyal verilerinden farklıdır. Dikkatli bir şekilde işlenmesi için DDİ teknik-

lerine ihtiyaç duymaktadır. Model için metin verilerinin ön işlenmesi ilk önemli adımdır. Geleneksel modeller genellikle iyi örnek özelliklerini yapay yöntemlerle elde etmek ve ardından klasik makine öğrenimi algoritmalarıyla sınıflandırmak için gereklidir. Bu nedenle, yöntemin etkililiği büyük ölçüde özellik çıkarımı tarafından kısıtlanmaktadır. Ancak, geleneksel modellere farklı olarak, derin öğrenme özellik mühendisliğini doğrudan çıktılara eşleştiren bir dizi doğrusal olmayan dönüşümleri öğrenerek model uyum sürecine entegre eder.



Şekil 3.1: Metin sınıflandırmasının akış şeması

1960'lardan 2010'lara kadar, metin sınıflandırma yöntemlerinde Naïve Bayes (NB) (Maron, 1961), K-Nearest Neighbor (KNN) (Cover and Hart, 1967) ve Support Vector Machine (SVM) (Joachims, 2005) gibi istatistik tabanlı geleneksel modeller egemen olmuştur. Önceki kural tabanlı yöntemlerle karşılaştırıldığında, bu yöntemler doğruluk ve kararlılık açısından açık avantajlara sahiptir. Ancak, bu yaklaşımlar yoğun özellik mühendisliği yapılmasını zorunda kılmaktadır ve bu da zaman alıcı ve maliyetli işlemlerde oluşmaktadır. Bunun yanı sıra genellikle metin verilerinde doğal sıralı yapıyı veya bağlamsal bilgiyi dikkate almamaktadır ve buda kelime anlamsal bilgisini öğrenmeyi zorlaştırmaktadır. 2010'lardan bu yana, metin sınıflandırması geleneksel modellerden derin öğrenme modellerine doğru giderek değişmiştir. Geleneksel yöntemlere kıyasla, derin öğrenme yöntemleri, insanlar tarafından tasarlanan kurallar ve özelliklerden kaçınarak, metin madenciliği için anlamsal olarak anlamlı temsilleri otomatik olarak sağlamaktadır. Bu nedenle, metin sınıflandırma araştırmalarının çoğu Derin Sinir Ağları (DSA'ler) temelli olup

yüksek hesaplama karmaşıklığına sahip veri odaklı yaklaşımlardır.

Klinik alandaki metin sınıflandırması ise diğer genel alanlardakilere (örneğin, e-posta sınıflandırması) kıyasla üç nedenden dolayı daha zordur. İlk olarak, klinik rapor metinleri genellikle yüksek düzeyde gürültü, seyreklik, karmaşık tıbbi terimler, tıbbi ölçümler ve skorlar (örneğin, 140/65 kan basıncı), kısaltmalar, yanlış yazılmış kelimeler ve kötü dilbilgisi cümleleri içerir (Nguyen and Patrick, 2016). Bu sorunları ele almak amacıyla, klinik metin sınıflandırma görevindeki metinlerin dilbilimsel özelliklerini ortaya çıkarmak için metinlerin ayrıntılı bir ön analizi yapılmalıdır. Bu tür sözcüksel karmaşıklıkları ele almak için, metinlerdeki gürültüyü gidermek için özel ön işleme teknikleri de geliştirilmiş ve kullanılmıştır (Nguyen and Patrick, 2016). İkinci olarak, klinik rapor veri kümeleri, genellikle dengesiz sınıf dağılımına sahiptir; burada bir sınıf, önemli ölçüde ilgi gören (örneğin, kanser pozitif) diğer sınıflara (örneğin, kanser negatif) kıyasla yetersiz temsil edilir. Bu nedenle, bu tür veri kümesindeki görece az sayıdaki pozitif örnekler, nadir görülen olaylar olarak sınıflandırılabilir, göz ardı edilebilir veya aykırı değerler olarak kabul edilebilir ve çoğunluk sınıfına kıyasla daha fazla yanlış sınıflandırmaya neden olabilir. Bu nedenle, normal örnekler arasında bu tür nadir bir sınıfın belirlenmesi gerekmektedir. Aksi takdirde, herhangi bir tanı hatası hastalara stres ve daha fazla komplikasyon getirebilir. Bu nedenle, bir sınıflandırma modeli, veri kümelerinde pozitif sınıfta yüksek tanımlama oranlarına ulaşabilmelidir.

Klinik metin sınıflandırma teknikleri, patoloji raporları, radyoloji raporları, otopsi raporları, ölüm belgeleri ve biyomedikal belgeler gibi birçok türde serbest metinli klinik raporda kullanılmıştır. Yapılan bir araştırmaya göre yapılan çalışmaların çoğunluğu, patoloji raporlarından sonra biyomedikal belgeler, radyoloji raporları ve otopsi raporlarını kullanıldığını göstermektedir (Mujtaba et al., 2019). Yine aynı çalışmaya göre patoloji raporlarınında birçoğu, meme kanseri veya diğer ilgili kanserleri tespit etmek için metin sınıflandırma teknikleri kullanılarak kullanılmıştır. Örneğin, Rani ve ark (Rani et al., 2015), kanser evrelerini metin sınıflandırma teknikleriyle tespit etmek için patoloji raporlarını kullanmışlardır. Radyoloji raporları da klinik metin

sınıflandırması alanında yaygın olarak kullanılmıştır. Zuccon ve ark. (Zuccon et al., 2013), radyoloji raporlarından eklem kırıklarını metin sınıflandırma teknikleriyle tespit eden çalışma yapmışlardır. Shin ve ark (Shin et al., 2017), beyin bilgisayarlı tomografi (beyin veya baş CT raporları) ile ilgili radyoloji raporlarını kullanarak çocukluk çağı travmatik beyin hasarını tespit etmek için kullanmışlardır. Ayrıca insan immun yetmezlik virüsünü (HIV) ve influenza benzeri hastalıkları tespit etmek için radyoloji raporlarını otomatik metin sınıflandırma teknikleriyle kullanan çalışmalar da literatürde mevcuttur (Bates et al., 2016; Pineda et al., 2015).

Bir sonraki bölümde bu tez çalışmasında klinik raporların sınıflandırılmasında kullanılan ve DSA mimarilerinden olan transformer tabanlı dil modeli (BERT) ve dikkat tabanlı yinelemeli sinir ağı dil modelinden teorik olarak bahsedilmiştir.

3.3 Dil Modelleme

Dil modellemesi, metnin olasılık dağılımını tahmin etme problemidir. Verilen n adet kelimenin bir dizisi olduğunda, bir dil modelinin amacı birleşik olayın olasılığını belirlemektir, yani:

$$p(w1, w2, ..., wn), \tag{3.1}$$

burada, her bir w_i bir rasgele değişkendir. Zincir kuralını uyguladığımızda, Denklem (3.1), koşullu olasılıkların bir çarpanları olarak yeniden yazılır:

$$p(w1, w2, ..., wn) = \prod_{i=1}^{n} p(w_i|w_{i-1}, ..., w_1) \tag{3.2}$$

$p(w_i|w_{i-1}, ..., w_1)$ burada, w_i kelimesinin tüm önceki kelimeler $w_{i-1}, ..., w_1$ sonrasında görünme olasılığıdır ve ayrıca (sol) bağlam olarak da adlandırılır. Denklem (3.2) daha işlenebilir olduğundan, genellikle istatistiksel modeller, koşullu olasılıkları $p(w_i|w_{i-1}, ..., w_1)$ veya onların yaklaşımlarını tahmin etmeyi hedefler.

3.3.1 N-grams

Denklem (3.2)'deki tüm koşullu olasılıkların tahmini, belirli ve isteğe bağlı olarak büyük bir derlemedeki gözlemlerden öğrenilebilir. Gerçekten de, w_i belirtecinin $w_{i-1}, ..., w_1$ bağlamı verildiğindeki olasılığı, w_i belirtecinin bu bağlamdan sonra kaç kez ortaya çıktığını sayarak elde edilebilir, bu sayma işlemi, $w_{i-1}, ..., w_1$ bağlamının toplamda kaç kez ortaya çıktığı ile normalize edilir:

$$p(w_i|w_{i-1}, ..., w_1) = \frac{\#(w_1, ..., w_{i-1}, w_i)}{\sum_{w_j \in V} \#(w_1, ..., w_{i-1}, w_j)} \tag{3.3}$$

burada $\#(\cdot)$, bir giriş gözleminin sayma işlemini temsil eder. Denklem (3.3)'deki tüm olası alt dizileri depolamanın ve kullanmanın, bağlam uzunluğu büyümeye başladıkça hafıza ve hesaplama açısından uygulanabilir olmadığı maalesef bir gerçektir. N-gram modelleri, bu sorunu N-1 tokenlik sabit boyutlu bir bağlam üzerinde koşullu bir olasılıkla yaklaşarak aşar. Dolayısıyla, Denklem (3.3) şu hale gelir:

$$p(w_1, w_2, \ldots, w_n) \approx \prod_{i=1}^{n} p(w_i|w_{i-1}, \ldots, w_{i-N+1}) \tag{3.4}$$

3.3.2 Nöral Dil Modelleri

N-gramların genelleştirme sorunlarını aşmak için, ileri beslemeli yapay sinir ağları, Denklem (3.4)'teki koşullu olasılıkları öğrenmek üzere adapte edilebilir:

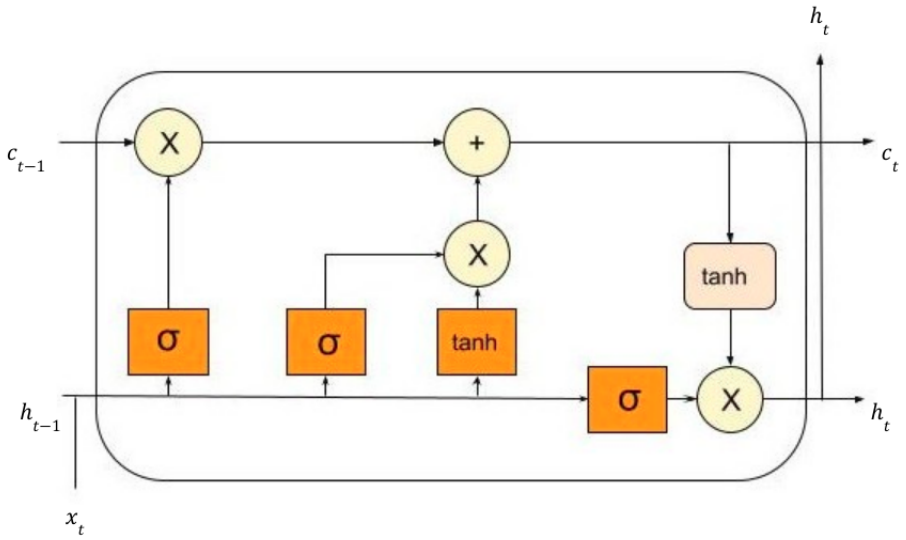
$$p(w_1, w_2, ..., w_n) \approx \prod_{i=1}^{n} p_\theta(w_i|w_{i-1}, ..., w_{i-N+1}) \tag{3.5}$$

Burada, p_θ, ağırlıkları θ tarafından parametrize edilmiş bir sinir ağı tarafından tahmin edilen dağılımı ifade eder. Model, son N-1 kelimeyi giriş olarak alan ve tüm dizi $(w_1, ..., w_n)$ içindeki sonraki kelimenin olasılığını maksimize etmek için optimize edilmiş bir çok-katmanlı perceptron (MLP)'dır. Bu denklem şu şekilde yazılabilir:

$$\max_\theta \prod_{i=1}^{n} p_\theta(w_i|w_{i-1}, ..., w_{i-N+1}). \tag{3.6}$$

Kelimeler birden fazla şekilde temsil edilebilir. Sinir ağları, N-gram modellerine kıyasla birçok avantaj sunar. Görünmeyen dizilere genelleme yapabilir ve eksik veya bilinmeyen verileri işleyebilir. Ayrıca, model boyutu corpus boyutundan bağımsızdır ve ilk sinir dil modeli (Bengio et al., 2000) tarafından önerilmiştir.

Tekrarlayan Sinir Ağı. Tekrarlayan Sinir Ağları (RNN) (Rumelhart et al., 1986), metin gibi sıralı verilerin işlenmesi için tasarlanmış bir sinir ağı grubudur. Veriler, karmaşık zamansal bağımlılıklar ve gizli bilgiler içermektedir. Bilgi sadece bir yönde aktığı beslemeli ileri ağların aksine, RNN'ler döngülere sahiptir ve geçmişten gelen bilginin gelecekte kullanılmasına olanak tanımaktadır. Bu özellik örneğin, "hasta kabul edildi" ve "suçları kabul edildi" arasındaki farkı ayırt etme olasılığını sağlamaktadır. RNN'lerde önemli bir sorun, kaybolan gradyan sorunudur. Geriye doğru hesaplama işlemi yapılırken ağırlıklar birçok kez çarpılır sonuç olarak küçük (kaybolan) veya büyük (patlayan) bir değer ortaya çıkar. Bu RNN'lerin karşılaştığı sorunun çözümü, uzun-kısa süreli bellek(LSTM) ağları olarak adlandırılan bir RNN varyantıdır (Hochreiter and Schmidhuber, 1997). LSTM, giriş kapısı, unutma kapısı ve çıkış kapısı adı verilen yenilikçi bir kapı mekanizmasıyla birlikte sabit bir hata akışı ve uzun süreli bağımlılık sorunlarını önleyen bir mekanizmaya sahiptir. LSTM'deki bellek, bir dahili durumda depolanır ve üç kapı hafızadan hangi bilgilerin dahil edileceği, ekleneceği veya kaldırılacağı konusunda önemli bir rol oynamaktadır. Zamanla bellek hücreleri ağırlıklara dayanarak hangi bilgilerin önemli olduğunu öğrenir. LSTM 'in basit bir mimarisi Şekil 3.2'te sunulmuştur (Graves, 2013).



Şekil 3.2: LSTM

Her bir t zaman adımında , LSTM birimlerini $\mathbb{R}^d$ uzayında bir vektör koleksiyonu olarak tanımlarız: bir giriş kapısı i_t, bir unutma kapısı f_t, bir çıkış kapısı o_t, bir bellek hücresi c_t ve bir gizli durum h_t. d, LSTM birimlerinin sayısıdır. Kapı vektörlerinin i_t, f_t ve o_t girişleri $[0, 1]$ aralığındadır. LSTM geçiş denklemleri aşağıdaki gibidir:

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + V_i c_{t-1}) \tag{3.7}$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + V_f c_{t-1}) \tag{3.8}$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + V_o c_t) \tag{3.9}$$

$$\tilde{c}t = \tanh(W_c x_t + U_c h_{t-1}) \tag{3.10}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{3.11}$$

$$h_t = o_t \odot \tanh(c_t) \tag{3.12}$$

Burada x_t mevcut zaman adımında girdiyi, σ sigmoid fonksiyonunu ifade eder ve $\odot$ eleman bazında çarpma işaretidir. Öznel olarak, unutma kapısı bellek hücresinin her bir biriminin silinme miktarını kontrol ederken, giriş kapısı her bir birimin ne kadar güncelleneceğini kontrol etmektedir. Son olarak çıkış kapısı ise iç bellek durumunun ortaya çıkmasını kontrol etmektedir.

Dikkat tabanlı Uzun Süreli Kısa Bellek Sinir Ağı. RNN metin sınıflandırmasıyla ilgili DDİ görevlerindede mükemmel sonuçlar sağlamaktadır. Ancak, özellikle sınıflandırma hatalarında yetersiz yorumlanabilirlikleri ve gizli verilerin okunamaz olması nedeniyle bu modeller yeterince sezgisel değildir. Bu problemin çözümü için literatürde dikkat odaklı yöntemler metin sınıflandırmasında başarılı bir şekilde kullanılmaktadır. Dikkat mekanizması ilk olarak makine çevirisinde Bahdanau ve ark (Bahdanau et al., 2014) tarafından önerilmiştir. Bir veri dizisinin LSTM kodlaması sırasında, ara hesaplamalar hesaplamalar (durumlar) gerçekleştirilir. Dikkat algoritmasının LSTM ile birlikte kullanımı dizideki kelimelere bağlam eklemek için bu durumları kullanır. Sözcüklere bağlam kazandırmak temsil gücünü artırmaktadır. Ayrıca Yiftach ve ark. (Barash et al., 2020) LSTM modelini dikkat mekanizması ile kullanarak ibranice dilinde yazılmış radyoloji raporlarının sınıflandırılmasında yüksek başarım elde etmiştir.

Formal olarak H LSTM ağının çıktı vektörlerinden oluşan $H = [h_1, h_2, ..., h_n]$ bir matris olsun.

$$M = \tanh(H) \tag{3.13}$$

$$\alpha = \text{softmax}(w^{\mathrm{T}} M) \tag{3.14}$$

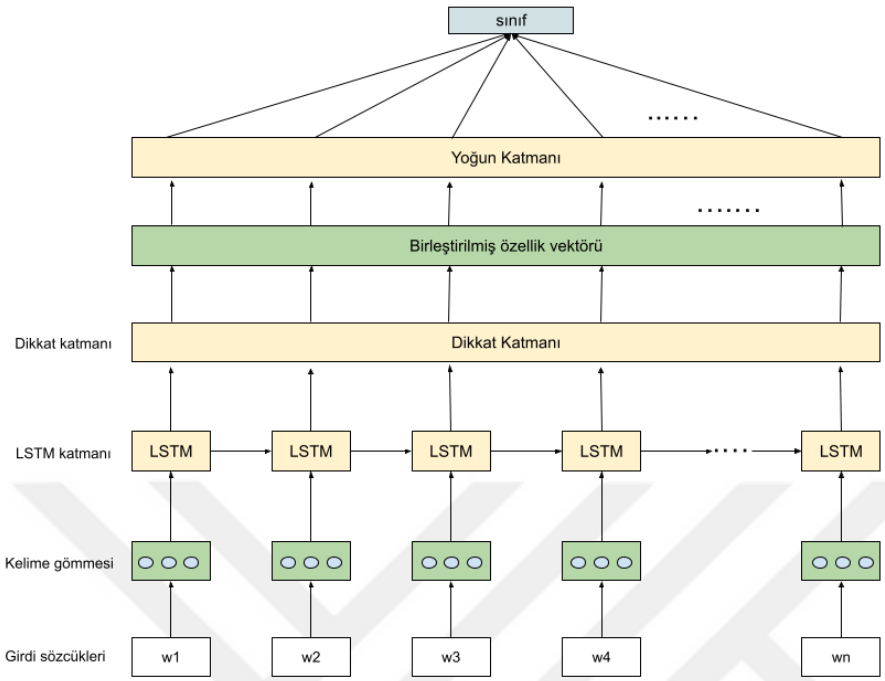
$$r = H\alpha^{\mathrm{T}} \tag{3.15}$$

Burada M, LSTM ağının çıktı vektörlerinin tanh fonksiyonuna sokulmuş halidir. Ağırlıklar α, w^{T} matrisiyle matris çarpımından sonra softmax fonksiyonuna sokulur. Çıktı vektörleri, α vektörü kullanılarak ağırlıklı toplamları alınarak, $r = H \cdot \alpha^T$ şeklinde hesaplanır. Son olarak, dizinin son temsilini, r bağlam vektörüne eleman bazlı olarak tanh aktivasyon fonksiyonu uygulanarak hesaplanır:

$$h^* = \tanh(r) \tag{3.16}$$

Bu mimaride, LSTM girdi dizisinin zamansal bağımlılıklarını yakalarken, dikkat mekanizması, öğrenilmiş ağırlık vektörü w'ye dayanarak her çıktı belirteci için girdi dizisinin farklı bölümlerine seçici olarak odaklanmasına izin verir. Son temsilci h^* sınıflandırma katmanına girdi olarak kullanılarak

tahminler yapılır. Tez çalışmasında temel model olarak kullanılan LSTM-attn modeli Şekil 3.3'de verilmiştir.



Şekil 3.3: LSTM-attn modeli

3.3.3 Transformer mimarileri

Transformatör mimarileri (Vaswani et al., 2017), DDİ'deki önemli gelişme lerden biridir ve birçok dil görevinde en iyi sonuçları (SOTA) elde etmiştir (Devlin et al., 2018; Dai et al., 2019b; Gu et al., 2021). Temalinde Transformer modeller, tekrarlama olmaksızın öz-dikkat mekanizması temelli ileri beslemeli modellerdir. Öz-dikkat, bir kelimeyi işlerken bağlamını göz önünde bulunduran bir yöntemdir. Transformer modelleri, bir dizi içindeki tüm belirteçleri aynı anda paralel olarak alarak uzun mesafeli bağımlılıkların yakalanmasını sağlar. Denklem (3.17), $x_1, x_2, ..., x_n$ dizisi girdi gömülerinin $z_1, z_2, ..., z_n$ 'ye uygun hale getirilmiş bir diziye eşlenmesini sunan öz-dikkat tanımını sunar. Her giriş x_i için üç vektör, sorgu q_i, anahtar k_i ve değer v_i hesaplanır. Dikkat fonksiyonu, aynı anda bir dizi sorgu için hesaplanır ve vektörizasyon işlemi ile bir Q matrisi

oluşturulur. Benzer şekilde, anahtarlar ve değerler, vektörizayon işlemleri ile K ve V matrislerine dönüştürülür. Matrislerin ağırlıkları eğitim sırasında öğrenilir ve bu üç vektörün hesaplanması için kullanılır. Öz-dikkat aynı zamanda ölçekli nokta-çarpım dikkati olarak da bilinir. Q, K, V doğrusal projeksiyonları için öz-dikkat şöyledir:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3.17}$$

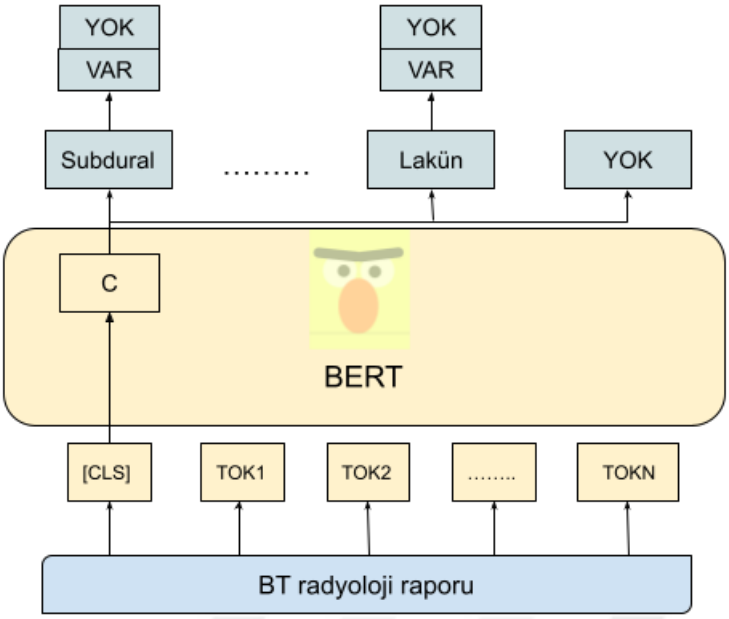
Burada giriş sorguları ve anahtarlarının boyutu d_k, değerlerin boyutu ise d_v'dir. Çoklu ağırlık matrisleri, aynı kelime için farklı sorgu, anahtar ve değer vektörleri üretir ve bu, modelin çoklu temsillere sahip olmasına izin verir; yani, transformatör model birden fazla dikkat başlığı kullanır (Garcia, 2021). Vaswani ve ark. (Vaswani et al., 2017), transformatör mimarisi hakkında daha fazla ayrıntı sunmaktadır.

Transformer mimarilerinin ön eğitiminde maskelenmiş dil modellemesi(MDM) ve bir sonraki cümle tahmini (SCT) kullanılan iki en yaygın yöntemdir. MDM, cümlelerden rastgele kelimeleri çıkarır ve modeli bağlama dayalı olarak eksik veya tam kelimeleri tahmin etmek için eğitmektedir. SCT, iki cümlelik çiftleri birlikte eğitir ve iki cümle girdi olarak kullanılır. SCT, ardından modeli ikinci cümlenin birinci cümlenin geçerli bir devamı olup olmadığını anlamaya eğitmektedir. Ön eğitimden sonra, BERT varyasyonları alan özel verilerde ince ayara tabi tutulur. Tüm parametreler uçtan uca ince ayarlanır. Ön eğitimli transformatör modeller, belirli bir akış aşağısı göreve yönelik olarak kullanılan metin dizilerini temsil etme konusunda iyi ve bağlam bağımlı yollar öğrenmektedir. Modellerin, belirli bir görevi gerçekleştirmek için temsil özelliklerini ince ayarlamaları yeterlidir. Ön eğitime kıyasla, ince ayarlama tekniği nispeten daha ucuz ve daha az maliyetlidir.

BERT (Çift Yönlü Kodlayıcı Temsilleri Transformatörler) (Devlin et al., 2018), dil modellemeye uygulanan çift yönlü eğitimli transformatör kodlayıcı mimarisini (Garcia, 2021) kullanan derin sinir ağı modelidir. BERT modeli, iki ön eğitim görevine dayanır: MDM ve SCT. BERT gibi transformer modelleri, kelime, alt kelime ve karakterlerden oluşan belirteçleri içeren sözcük parçaları

tokenleştirmesini kullanarak kelime sözlüğünü oluşturur. BERT-base modeli 12 katman, 768 gizli ünite veya özellik sayısı ve 12 öz-dikkat başına sahiptir. BERT-base, İngilizce Wikipedia ve BookCorpus üzerinde ön eğitimli 110 milyon parametre içeren sinir ağı mimarisidir.

Dil modellerini ön eğitme ve ince ayarlama paradigması, büyük, etiketlenmemiş derlemlerden etkili bir şekilde yararlanma ve sonrasında modeli, gereken etiketli veri miktarı açısından nispeten kaynak verimli bir süreçte belirli bir görevi gerçekleştirmek için özelleştirme avantajı sağlar. Dil modelleri, genellikle genel alanlarda derlem kullanarak ön eğitim alır, ör. Wikipedia. Bununla birlikte genel dil modellerinin özel alanlarda kullanılması önemli alan farklılıkları nedeniyle (Lewis et al., 2019; Gururangan et al., 2020) en uygun olmayabilir. Sonuç olarak, SciBERT (Beltagy et al., 2019) ve BioBERT (Lee et al., 2020) gibi alan özelinde dil modelleri geliştirmeye yönelik birçok çalışma ortaya çıkmıştır. Alan özelinde dil modelleri geliştirmek için farklı yaklaşımlar vardır, bunlar arasında sıfırdan alan içi verilerle bir dil modelini ön eğitmek ve mevcut genel bir dil modelini alan uyumlu ön eğitim olarak bilinen bir süreçte alan içi verilerle ön eğitmeye devam etmek yer almaktadır. Şekil 3.4, çok etiketli sınıflandırma görevi için kullanılan bir BERT-tabanlı model örneğini sunmaktadır.



Şekil 3.4: BERT

Genel BERT modellerini biyomedikal alanına uyarlamaya yönelik erken girişimlerden biri olan BioBERT, performansını artırmak için sürekli ön eğitim kullanmıştır. Özellikle, BioBERT, genel BERT sürümünün ön yüklemesiyle başlatılmıştır ve ardından PubMed özetleri ve tam metin makaleler üzerinde eğitime devam ettirilmiştir. Geliştirilen model adlandırılmış öğe tanıma, ilişki çıkarma ve 15 açık biyomedikal veri kümesinde soru cevaplama gibi DDİ görevlerinde genel alandaki BERT'e göre daha iyi performans sağlamıştır. Benzer şekilde, ClinicalBERT (Alsentzer et al., 2019), MIMIC verileri ile sürekli ön eğitim kullanarak oluşturulan alan özgü bir dil modelidir. BERT ve BioBERT'ten başlatılan modellere kıyasla, MIMIC verilerini sürekli ön eğitim ile kullanarak klinik görev performansını artırmanın etkili olduğunu göstermiştir. Biyomedikal dil modelleri oluşturmak için sürekli ön eğitim yöntemini inceleyen diğer bir çalışmada SciBERT (Beltagy et al., 2019) modelidir. SciBERT büyük bir bilimsel metin derlemesinde BERT'in sürekli ön eğitim ile devam etmesiyle oluşturulmuştur. Bu yöntem ile biyomedikal NLP görevlerinde adlandırılmış öğe tanıma, ilişki çıkarma ve öğe bağlama gibi görevlerde performansı diğer çalışmalara kıyasla arttırmıştır. İkinci olarak, BlueBERT

(Peng et al., 2019), biyomedikal ve genel alan derlemelerinin bir kombinasyonunda BERT'in sürekli ön eğitim ile devam etmesiyle oluşturulmuştur. Yazarlar, biyomedikal derlemelerin biyomedikal DDİ görevlerinde performansı artırdığını raporlamışlardır, ancak etki, özel göreve bağlı olarak değişeceğini belirtmişlerdir. Biyomedikal alan genel modellerin transferinde ikinci yaklaşım, genel dil modeline kullanmadan sadece belirli bir alandaki verilerle dil modeli ön eğitimi yöntemidir. Baştan ön eğitim olarak bilinen bu yöntem biyomedikal dil modelleri oluşturmada etkili olduğu yapılan çalışmalarda kanıtlanmıştır. PubMedBERT (Gu et al., 2021), PubMed özetleri üzerinde baştan eğitilmiş bir model örneğidir. PubMedBERT modeli PubMed gibi büyük, kapsamlı ve yüksek kaliteli bir veri kümesinin sadece sözlük ve model oluşturma için kullanılması durumunda daha iyi sonuçlar elde edilebileceğini göstermiştir.

İki ön eğitim yöntemi karşılaştırıldığında, sürekli ön eğitim yöntemiyle modelin genel alandan biyomedikal veya klinik alana aktarılmasının daha başarılı olduğu çalışmalar da bulunmaktadır. Örneğin, bir çalışmada (Bressem et al., 2020) dört BERT modeli önerilmiştir; bunların ikisi Almanca'da 3,8 milyon radyoloji serbest metin raporundan ön eğitimli (FS-BERT, RAD-BERT) ve diğer ikisi açık kaynaklı modellere dayalı (MULTI-BERT, GER-BERT) sürekli eğitim modelleridir. Sıfırdan ön eğitim yaklaşımı kullanılarak geliştirilen FS-BERT modelinin performansı diğer modellere göre görece düşük olarak raporlanmıştır. Yapılan karşılaştırma sonuçlarına dayanarak yazarlar yalnızca alan özelindeki derlem kullanmanın uygun kelime gösterimlerini öğrenmek için yetersiz olabileceğini savunmuşlardır. Benzer şekilde başka bir çalışmada RadBERT, milyonlarca radyoloji raporu ve çeşitli dil modelleriyle ön eğitimli altı trasnformatör dil modeli geliştirilmiştir. Çalışmada, transformer temelli dil modellerini radyoloji için uyarlamalarının radyoloji DDİ uygulamalarında performansı iyileştirip iyileştiremeyeceğini araştırılmıştır. RadBERT'in her varyantı, anormal cümle sınıflandırması, rapor kodlama ve rapor özetleme içeren radyolojide üç DDİ görevi için ince ayar yapılmıştır. Genel alan veya biyomedikal modelden farklı ağırlık başlatma ile ön eğitimli alan özelindeki modeller, temel alınan modellerden (BERT-taban, BioBERT, Klinik-BERT,

BlueBERT ve BioMed-RoBERTa) daha iyi sonuçlar vermiştir.

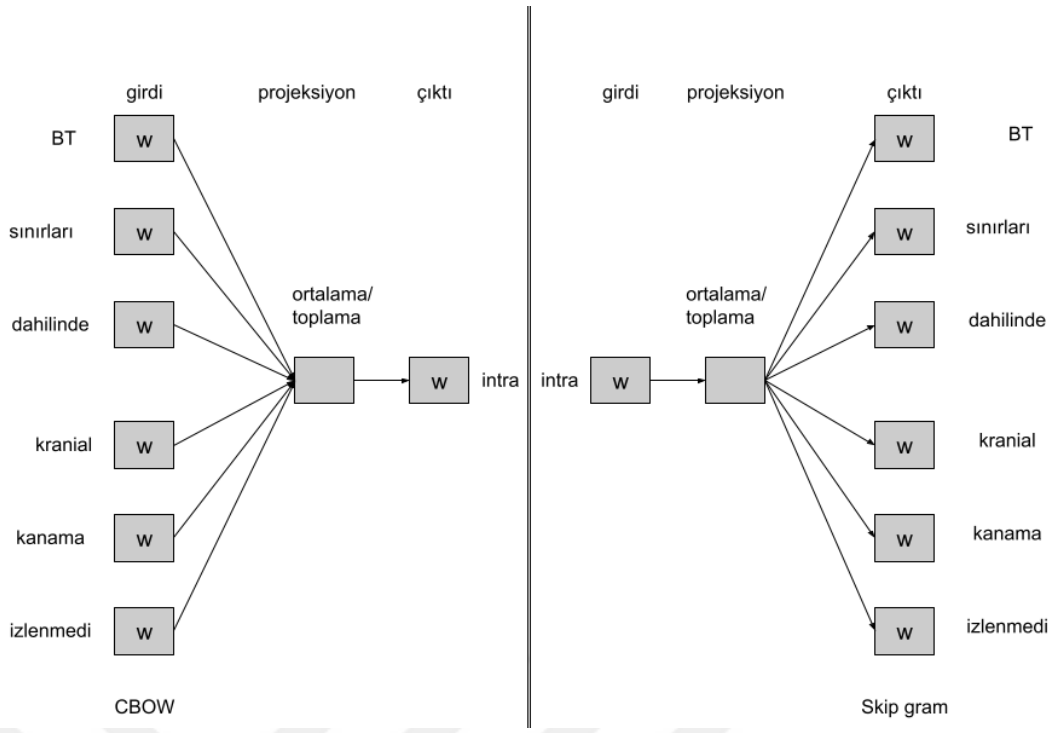
BERT modellerinin ön eğitimi çeşitli biyomedikal NLP görevlerinde performansı artırabilecek daha genel bir yaklaşım olmasına rağmen çeşitli derlemler seçerek yapılan söz konusu ön eğitim yaklaşımları ön eğitim için önemli miktarda alan özelinde veri gerektirmektedir. Genel alan verilerinin aksine biyomedikal metin verileri yaygın olarak kullanılabilir değildir ve çeşitli kaynaklara dağılmıştır. Ayrıca az sayıda web üzerinden erişilebilir tıbbi veritabanı İngilizce dışındaki dillerde yazılmıştır, bu da dil modellerini ön eğitmek için büyük biyomedikal derlem elde etmeyi zorlaştırmaktadır. Bu nedenle sınırlı dil modeli kaynaklarıyla bile iyi çalışabilen dil modelleri oluşturmak için etkili ve verimli tekniklere yüksek talep vardır. Bu probleme bir çözüm olarak, yapılan bir çalışmada önerilen Eşzamanlı ön eğitim tekniği (Wada et al., 2020), sınırlı alan özelindeki bir derlemi yukarı örnekleme ve daha büyük bir derlemle dengeli bir şekilde ön eğitim için kullanma yöntemidir. Çalışmada küçük Japonca tıbbi makale özetleri ve Japonca Wikipedia metni kullanarak, yazarlar eşzamanlı ön eğitilmiş bir BERT modeli olan ouBioBERT'i oluşturmuşlardır. Ayrıca yapılan çalışma Japonca tıbbi BERT modelinin, Japonca tıbbi belge sınıflandırma görevinde geleneksel modellerden ve diğer BERT modellerinden daha iyi performans gösterdiğini doğrulamıştır. Sınırlı kaynak probleminin engellerini aşmak için bir başka çözüm olarak birçok araştırmacı görev dağılımından çıkarılan daha küçük bir derlem üzerinde sürekli ön eğitimin faydalarını, görev-uyumlu ön eğitim olarak araştırmışlardır (Gururangan et al., 2020; Boudjellal et al., 2021; Schneider et al., 2020). Ancak bu çalışmalar, klinik DDİ görevleri için dil modellerini değerlendirmemiştir.

Bu tez çalışması, yukarıda belirtilen boşluğu ele alarak, az kaynaklı bir dil olan Türkçe'de, kısıtlı Türkçe medikal ve klinik derlem üzerinde farklı ön eğitim teknikleri ile önceden eğitilmiş BERT modelleri geliştirmeyi hedeflemiştir (Türkmen et al., 2022). Oluşturulan dil modelleri, Ege Üniversitesi Hastanesi'nden alınan Türkçe kafa BT raporlarını sınıflandırma amacıyla test edilmiştir. Elde edilen sonuçlar, alanla ilgili verilerin sınırlı olmasına rağmen, klinik görev performansını artırabileceğini göstermiştir.

3.4 Kelime Gösterimleri

Kelime gösterim teknikleri bir sırayı tahmin ederek V kelime dağılımı üreten ve beslemeli bir sinir ağı tarafından eğitilen ilk katmanın (gömülme katmanı) aktivasyonlarından elde edilen kelime özelliklerinden oluşmaktadır. Sinir ağı pahalı etiketleme işlemi gerektirmez ancak kolayca elde edilebilen büyük etiketsiz kümelerden türetilebilmektedir. Böylece önceden eğitilmiş kelime gösterimleri az miktarda etiketli veri kullanan DDİ görevlerinde kullanılarak performansı arttırmaktadır.

Kelime gösterimlerine önceden yaklaşımlar olsa da (Collobert and Weston, 2008), Word2vec (Mikolov et al., 2013) bu alanı popülerleştiren olmuştur. Word2vec, kelime anlamlarının bağlamıyla çıkarılabileceğine dair dağılımsal hipoteze dayanmaktadır. Word2vec, kelime temsillerini elde etmek için iki model mimarisiyle eğitilebilir: Sürekli Kelime Torbası (CBOW) ve Atla-gram modeli (Skip-gram). İlk model, çevresindeki kelime penceresi verildiğinde bir kelime tahmin etmeyi öğrenir ve kelime sırası dikkate alınmaz. İkinci model mimarisi, sürekli projeksiyon katmanıyla log-lineer bir sınıflandırıcı kullanarak belirli bir hedef kelime öncesi ve sonrasını tahmin etmeyi öğrenmektedir. Hedef kelimeye daha uzak olan kelimeler, genellikle hedef kelimeyle daha az ilgili oldukları için daha az ağırlık verilmektedir. Örneğin, Şekil 3.5'te gösterildiği gibi CBOW modeli, "BT sınırları dahiilinde" ve "kranial kanama izlenmedi" intra vektörünü tahmin ederken, Skip-gram modeli intra kelimesini "BT sınırları dahiilinde" ve "kranial kanama izlenmedi" kelimelerini tahmin etmek için kullanmaktadır.



Şekil 3.5: Word2vec mimarisi

3.5 Değerlendirme Ölçüleri

Tez kapsamında geliştirilen dil modellerinin sonuçları kesinlik, duyarlılık ve F1-skoru kullanılarak değerlendirilmiştir. Aynı zamanda ayrıntılı analiz için iki sınıflama problemi için her bir etiketin performansı ayrı bir şekilde değerlendirilmiştir. Kesinlik, duyarlılık ve F1-skorunun yanı sıra, t testi (Dietterich, 1998) modeller arasında istatistiksel farklılıklar olup olmadığını belirlemek için kullanılmıştır. Sonuçların istatistiksel olarak anlamlı olduğunu kabul etmek için eşik değer 0,05 olarak belirlenmiştir. Kullanılan değerlendirme metriklerinin ayrıntılı açıklaması aşağıda verilmiştir.

3.5.1 F1-skoru

Çok sınıflı sınıflandırma veya çok etiketli sınıflandırmanın model ve snınf seviyesindeki değerlendirme için F- ölçütü veya F1-skoru birincil değerlendirme ölçütü olarak kullanılmaktadır.

F1 skoru şu şekilde hesaplanır:

$$P = \frac{\text{gerçek pozitif oran}}{\text{gerçek pozitif oran} + \text{yanlış pozitif oran}} \tag{3.18}$$

$$R = \frac{\text{gerçek pozitif oran}}{\text{gerçek pozitif oran} + \text{yanlış negatif oran}} \tag{3.19}$$

$$F1 = 2 \times \frac{P \times R}{P + R} \tag{3.20}$$

Burada P hassasiyeti, R ise duyarlılığı ifade eder.

Çoklu etiket sınıflandırmada, bu durum daha karmaşıktır çünkü bu ölçümler örnek veya sınıf düzeyinde hesaplanabilir. Bu nedenle, literatürde genellikle iki ortalama yöntemi önerilir - makro ve mikro ortalamalar.

Mikro-ortalama puanlar örnek örneği j için hesaplanırken makro-ortalama puanlar sınıf m için hesaplanır ve tüm sınıflar k üzerinde ortalama alınır, aşağıdaki denklemlerde görüldüğü gibidir.

$$\text{Prec}_{\text{mac}} = \frac{1}{k} \sum_{m=1}^{k} \text{Prec}_m \tag{3.15}$$

$$\text{Prec}_{\text{mic}} = \frac{\sum_{j=1}^{i} \text{TP}j}{\sum_{j=1}^{i} (\text{TP}_j + \text{FP}_j)} \tag{3.16}$$

$$\text{Rec}_{\text{mic}} = \frac{\sum_{j=1}^{i} \text{TP}j}{\sum_{j=1}^{i} \text{TP}j + \sum_{j=1}^{i} \text{FN}_j} \tag{3.17}$$

$$\text{Rec}_{\text{mac}} = \frac{1}{k} \sum_{m=1}^{k} \text{Rec}_m \tag{3.18}$$

Mikro ve makro ortalamalı F ölçüt puanları, bu mikro ve makro puanları F1 skor formülüne yerleştirerek hesaplanır.

Bu ölçümler arasındaki temel fark, makro-ortalama F1'in tüm etiketlere eşit davranması ve mikro-ortalama F1'in her belgedeki sınıflandırıcının aldığı kararlara eşit ağırlık vermesidir. Dengesiz bir veri kümesinde, daha büyük sınıfların daha büyük bir etkisi vardır ve bu durumda makro-ortalama F1 ölçütünde dikkate alınmaz.

Tez çalışmasındaki oluşturulan veri kümesi sınıf bazında dengesiz dağılım gösterdiği sonuçlar sunulurken mikro-ortalama F1 skoru kullanılmıştır.

3.5.2 Farklılıkların istatistiksel değerlendirmesi

Belirli bir görev için makine öğrenimi yöntemlerinin performansını karşılaştırmak ve nihai bir yöntem seçmek, uygulamalı makina öğrenmesinde yaygın bir işlemdir. Performans karşılaştırmasında kullanılan istatistiksel anlamlılık testleri, modellerin performansını karşılaştırmak ve aynı dağılımdan çekildikleri varsayımı göz önüne alındığında gözlemlenen performans puanı örneklerinin olasılığını ölçmek için tasarlanmıştır. Bu varsayım veya boş hipotez reddedilirse, başarı puanlarındaki farkın istatistiksel olarak anlamlı olduğunu gösterir. Makine öğrenmesi modellerinin performansını karşılaştırmak için kullanılan en yaygın istatistiksel hipotez testi eğitim veri kümesinin rastgele alt örnekleri aracılığıyla birleştirilen eşleştirilmiş Student t-testidir. Bu testteki boş hipotez, uygulanan iki ML modelinin performansı arasında fark olmadığıdır. Başka bir deyişle boş hipotez her iki ML modelinin de aynı performansı gösterdiğini varsayar. Öte yandan alternatif hipotez, uygulanan iki makina öğrenmesi modelinin farklı performans gösterdiğini varsayar.
Eşleştirilmiş student t-testi (Dietterich, 1998), varyasyonları neredeyse normal dağılan bir dizi rastgele örnek için popülasyon ortalamaları arasındaki farkın hipotez incelemesini verir. Denekler genellikle bir önce-sonra durumunda veya mümkün olduğunca benzer deneklerle test edilir. Eşleştirilmiş t-testi, iki gözlem arasındaki farkların sıfır olduğuna dair bir testtir.

Eşleştirilmiş t-testinin formülü şu şekilde verilir:

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} \tag{3.21}$$

Burada, t test istatistiğini, $\bar{d}$ eşleştirilmiş farkların ortalamasını, s_d eşleştirilmiş farkların standart sapmasını ve n eşleştirilmiş gözlem sayısını gösterir.

3.6 İlgili Çalışmalar

3.6.1 Klinik Doğal Dil İşleme ve Radyoloji

Radyoloji alanında yapılan çalışmalar incelendiğinde üç perspektiften çeşitlendiği görülmüştür. Bunlardan ilk doğal dil işlemenin seviyesi (doküman, cümle ve sözcük), ikincisi seviyeye ait farklı etiketler ile radyolojiye özgü problemlerin uygulanması ve son olarak da dil modellerinin domaine transfer edilme yöntemleri. Bu bölümde radyolojideki farklı seviyelerdeki doğal dil işleme yöntemlerinin dil modelleri yöntemiyle uygulayan çalışmalar özetlenmiştir. David ve ark. (Wood et al., 2020) derin öğrenme medikal görüntü işleme uygulamalarında etiketli verilerin oluşturulmasındaki zorluklara vurgu yaparak radyoloji metinlerinden otomatik olarak görüntünün etiketlenmesini gerçekleyen ALARM adında bir uygulama geliştirmişlerdir. Magnetik rezonans görüntüleme raporlarını transformer temelli bir sınıflandırıcı ile "normal" ve "abnormal" olarak ve ek olarak daha detaylı bir etiket seviyesi olarak "abnormal" etiketinin beş kategorisini sınıflandırmışlardır. Biobert (Lee et al., 2020) modelinin üzerine özel bir sınıflandırıcı ekleyerek yeni bir model geliştirmişlerdir. Ve bu modelin doğruluğu global etiket kümesinde %99 ile spesifik etiketli veri kümesinde ise %90 doğruluk ile uzmanlardan daha başarılı sonuç vermiştir. Tekrardan görüntüleme uygulamalarındaki etiketleme maliyeti probleminden yola çıkarak Ignat ve ark. (Drozdov et al., 2020) göğüs radyografisi raporlarını "anormal", "abnormal" ve "unclear" olarak sınıflandırmışlardır. Andreas ve ark. (Grivas et al., 2020) beyin görüntüleme rapor metnindeki felç, felç alt tipleri ve tümörler ve küçük damar hastalığı gibi diğer ilgili bulguları tanıyan transformer mimarisi tabanlı bir sistem geliştirmişlerdir. Etiketler sözcük seviyesindedir ve klinik varlık tanıma uygulamasıdır. Radyoloji muayenelerindeki acil klinik bulguların sevk eden doktorlara iletilmesindeki gecikmeler, her yıl önemli olumsuz hasta sonuçlarından sorumludur. Aynı zamanda radyologlar ve sevk eden doktorlar arasında hızlı iletişim gerektiren radyolojik bulgulara ilişkin net kılavuzların olmaması bu gecikmenin ana nedenlerinden biridir. Xing ve ark. (Meng et al., 2020) bu problemden yola çıkarak, bir

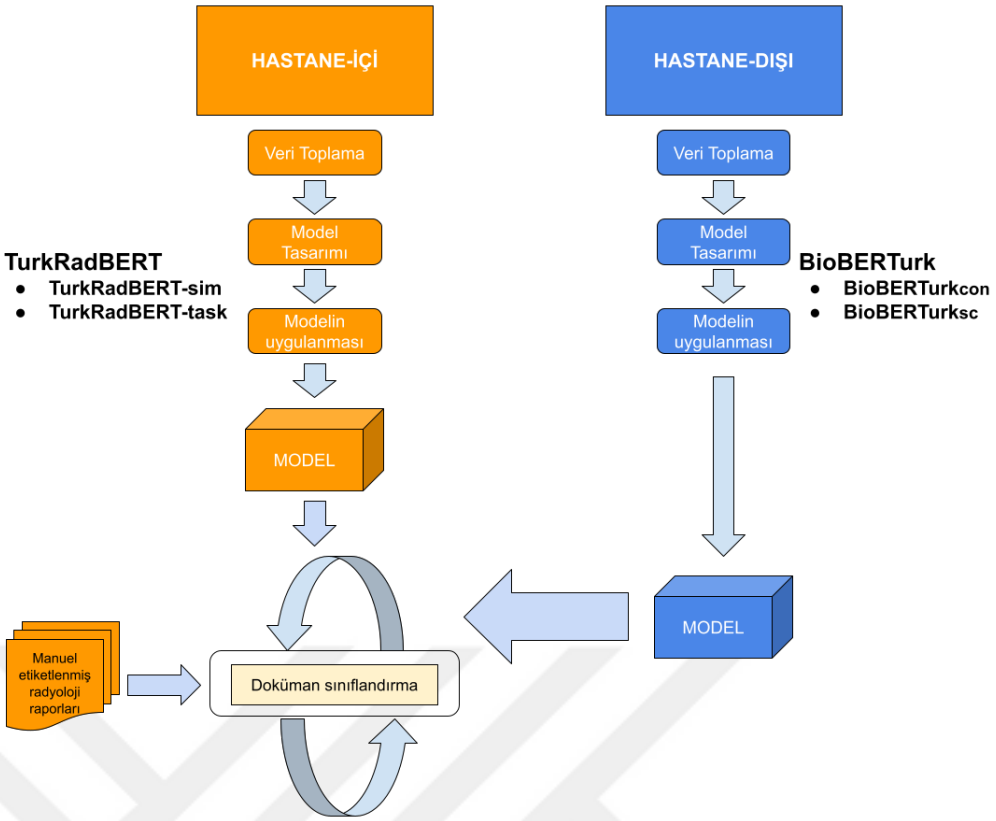
radyoloğun sevk eden doktorla hızlı iletişim kurmasını gerektiren radyoloji raporlarının tanımlanmasını iyileştirmeyi amaçlayan BERT mimarisini geniş bir radyoloji raporu külliyatında önceden eğiterek ve ince ayar (fine-tuning) gerçeklemiştir. Yapılan çalışmalarda Radyologlar rutin olarak hastaların takip görüntüleme önerileri yayınlamasına rağmen, takip için hastaların uyum oranlarının düşük olduğu bildirilmiştir. Ayrıca, takip görüntülemesi almayan hastalar, olumsuz sonuçlar ve komplikasyonlar açısından risk altındadır. Radyoloji alanında transformer mimarisini kullanan bir diğer çalışma ise Surabhi ve arkadaşları tarafından (Datta et al., 2020) geliştirilen ve semantik rol etiketleme temelinde bir etiket kümesi ile çalışan sınıflayıcıdır. Radyoloji metinlerinden bilgi çıkarımını yapabilmek için LSTM modelini BERT mimarisi ile melezleyerek kullanmışlardır. Medikal alanda alan transferinin en geniş kullanıldığı çalışma ise Akshay ve ark. (Smit et al., 2020) tarafından sunulan CheXbert sistemidir. Bu çalışmada, hem mevcut kural tabanlı sistemlerin ölçeğini hem de uzman etiketlerinin kalitesini kullanan tıbbi görüntü raporu etiketlemesine BERT tabanlı bir yaklaşım geliştirmişlerdir. Biyomedikal olarak önceden eğitilmiş bir BERT modelinin üstün performansını, önce kural tabanlı bir etiketleyicinin etiketleri üzerinde eğitilmiş ve daha sonra otomatik geri çeviri ile zenginleştirilmiş küçük bir uzman etiket kümesi üzerinde ince ayarı yapılmış bir BERT modelinin performansının göğüs röntgeni veri kümesinde yapılmış çoğu çalışmadan daha iyi olduğunu göstermişlerdir.

Yukarıda bahsedilen çalışmaların tümü ingilizce dilinde gerçeklenmiştir. Ancak az da olsa almanca (Bressem et al., 2020), ibranice (Barash et al., 2020) ve ispanyolca (López-Ubeda et al., 2020) gibi radyoloji metinleri üzerinde dil modelelri ile doğal dil işleme uygulamasını geliştiren çalışmalar da vardır.

4 DİL MODELİ GELİŞTİRİMİ: DENEYSEL ÇERÇEVE

Büyük dil modellerindeki (BDM'ler) son gelişmeler, GPT-3 Rae et al. (2021) gibi parametre boyutları yüz milyarı aşan ve devasa veri kümeleri üzerinde önceden eğitilmiş modellerin geliştirilmesine yol açmasına rağmen sınırlı kaynaklarla karakterize edilen özel alanlardaki BDM mimarilerine odaklanan araştırma sayısı azdır. Literatürde sınırlı dil kaynakları sorununu ele almak için dil modelleri geliştirmeye yönelik, çeşitli ön eğitim teknikleri Wada et al. (2020); Gururangan et al. (2020) mevcuttur. Ön eğitim tekniğinin seçimi belirli görev verilerine ve mevcut kaynaklara bağlıdır ancak ön eğitimde sınırlı verilerin en iyi şekilde kullanılmasının belirlenmesi ve ön eğitim yöntemleri için en uygun verilerin seçilmesi açık sorular olmaya devam etmektedir. Tez çalışmasında, düşük kaynak ortamları olan Türkçe medikal ve klinik alanında bir BDM türü olan BERT modellerinin ön eğitimi için sınırlı metin derlemi kullanarak farklı teknikleri değerlendirmeyi ve karşılaştırmayı amaçlayan iki farklı dil BERT dil modelleri ailesi, BioBERTurk ve TurkRadBERT dil modelleri ailesi geliştirilmiştir.

Bu bölüm, bir dil modelinin geliştirilmesi sürecinin genel bir çerçevesini sunarak, sınırlı kaynağa sahip olan Türkçe'de BERT dil modelinin geliştirmenin aşamalarına ve bu aşamaların nasıl bir deneysel çerçevede uygulanabileceğine odaklanmıştır. Bu genel çerçevenin ardından tezin ilerleyen bölümlerinde bu deneysel süreçlerin gerçek dünyadaki uygulamalarına odaklanılmıştır. Tez kapsamında geliştirilen ve iki farklı Türkçe dil modeli ailesi olan BioBERTurk ve TurkRadBERT'in geliştirme süreçleri bu bölümdeki deneysel çerçeve içinde tezin ayrı bir şekilde iki farklı bölüm şeklinde ele alınmış ve incelenmiştir. Geliştirilen Türkçe dil modellerinin geliştirilme sürecinde kullanılan deneysel prensipler ve süreçler bu bölümde sunulan genel çerçevenin uygulamalı bir örneğini oluşturacaktır.

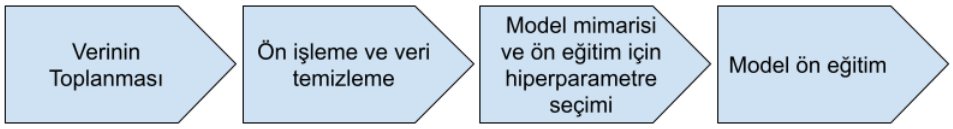


Şekil 4.1: Sistem mimarisi

Figür 4.2'de görüldüğü doktora tezi çalışmasında, radyoloji raporlarının sınıflandırılması görevine odaklanan ancak eğitimde kullandıkları veri kaynaklarına göre ayrılan iki ayrı BERT dil modeli ailesi oluşturulmuştur: BioBERTurk ve TurkRadBERT. Her iki model ailesi de ön eğitim yöntemlerini yansıtmak üzere iki farklı varyasyona sahiptir. BioBERTurk, hastane dışından toplanan genel veriler kullanılarak eğitilmiştir. Bu genel veri kümeleri, modelin dil modelleme ve genelleştirme yeteneklerini geliştirmek için kullanılmıştır. Öte yandan, TurkRadBERT modeli alana özgü bir duruma uyarlanmıştır: hastane içinden gelen veriler kullanılarak eğitilmiştir. Hastane içi veriler, modelin radyoloji raporlarındaki belirli terminolojiyi, yapıyı ve içeriği anlamasını sağlamak için kullanılmıştır.

Dil modellemesi konusunda, bir dil modelinin geliştirilme süreci, deneysel evreler serisini içerir ve bu sürecin anlaşılması, modelin yeteneklerini ve başarısını doğru bir şekilde değerlendirebilmek için hayati öneme sahiptir.

BERT dil modelinin geliştirilme süreci, veri toplama, ön işleme ve veri temizleme, model mimarisinin belirlenmesi, ön eğitim için hiperparametre seçimi ve modelin ön eğitimini gerçekleştirme gibi dört temel bileşenden oluşur. Tezdeki tüm BERT tabanlı dil modellerinin geliştirilme süreci Figür 4.2'de özetlenmiştir. Aşağıda her bir bileşen bu çerçeve dahilinde detaylandırılmıştır.



Şekil 4.2: BERT dil modeli geliştirim süreci

4.1 Veri Toplama

Küçük ölçekli dil modelleri ile karşılaştırıldığında, BERT mimarilerinin ön eğitimi için yüksek kaliteli verilere daha güçlü bir ihtiyacı vardır. Ayrıca model kapasiteleri büyük ölçüde ön eğitim derlemesine ve nasıl önceden işlendiğine bağlıdır (Zhao et al., 2023). Yetenekli bir dil modeli geliştirmek için çeşitli veri kaynaklarından büyük miktarda doğal dil derlemesi toplamak önemlidir. Mevcut dil modelleri, ön eğitim derlemesi olarak çeşitli araştırmacıların kullanıma farklı alanlara ait (web, biyomedikal vb) açık metin veri kümelerinin bir karışımını kullanmaktadır. Ön eğitim derlemesinin kaynağı genel olarak iki türe ayrılabilir: genel veri ve özel veri.

BERT dil modelleri, genellikle web sayfaları, kitaplar ve konuşma metinleri gibi büyük, çeşitli ve kolayca erişilebilir olan genel verilerle beslenir. Bu tür veriler dil modelleme ve genelleştirme yeteneklerinin geliştirilmesinde oldukça etkili olduğu konusunda birçok çalışmada belirtildiği üzere BERT mimarilerinin eğitiminde kullanımı için idealdir (Chowdhery et al., 2022). Ancak BERT dil modellerinin belirli Doğal Dil İşleme (DDİ) görevlerinde daha yüksek performans göstermeleri, sadece genel veri kullanımıyla sınırlı olmayabilir.

Özel veri kümelerinin kullanımı belirli bir dilde yazılmış metinlerden sağlık alanına özgü metinlere kadar belirli bir konu veya alanla ilgili metinlerin entegrasyonunu içerebilir. Bu, modelin ilgili alanda daha derinlemesine bir bilgi birikimi geliştirmesini sağlar. Örneğin, BLOOM ve PaLM gibi dil modelleri, çokdilli veri toplama stratejilerini kullanarak ön eğitim korpularını genişletmiş ve bu sayede çok dilli görevlerde etkileyici sonuçlar elde etmişlerdir.

Bilimsel metinlerin kullanımı da dil modellerinin bilimsel bilgi anlama kapasitelerini genişletme konusunda hayati bir rol oynar. Bilimsel bir korpusun ön eğitim aşamasında modelle entegre edilmesi dil modelinin bilimsel ve mantıksal görevlerde mükemmel performans sergilemesini sağlar. Bilimsel korpuslar arXiv makaleleri, bilimsel ders kitapları ve matematik web sayfaları gibi kaynaklardan derlenir ve bu verilerin karmaşık yapısı dil modellerinin işleyebileceği bir formata dönüştürmek için özel belirteçlere dönüştürme ve ön işleme tekniklerini gerektirir. Bu tez çalışmasında Türkçe özel alanda BERT modelleri geliştirmek için farklı kaynaklardan metin verileri toplanmıştır. Tez kapsamında geliştirilen Türkçe dil modellerinin ön eğitiminde kullanılan ve hazırlanan veri kümeleri aşağıdaki tabloda özetlenmiştir.

Tablo 4.1 'de görüldüğü gibi Türkçe BERT modeli eğitiminde bir genel veri kümesi üç özel alanda veri kümesi olmak üzere 4 farklı metin kümesi toplanmıştır. trW olarak ifade edilen ve Wikipedia gibi çeşitli internet sitelerinden toplanarak oluşturulan Türkçe web külliyatı genel alanda BERTurk [1] dil modelinin ön eğitimi için oluşturulmuştur. Tez kapsamında bu veri kümesi hazır olarak kullanılmıştır. trM, trR ve trK ile ifade edilen metin derlemleri ise tez kapsamında Türkçe medikal ve klinik alanda BERT modellerini geliştirilmek için çeşitli kaynaklardan toplanarak oluşturulmuştur.

BERT dil modeli geliştirimi için ilk oluşturulan derlem Türkçe Medikal derlemdir ('trM'). PubMed'deki Türkçe özetler dil modeli eğitiminde çok sınırlı olmasından kaynaklı daha anlamlı büyüklükte bir medikal alanında derlem oluşturabilmek için Dergipark sitesi kullanılmıştır. Dergipark, Ulakbim (Türkiye Akademik Ağ ve Bilgi Merkezi) tarafından geliştirilmiş ve yönetilen,

[1] https://github.com/stefan-it/turkish-bert/tree/master

Tablo 4.1: Türkçe dil modellerinin ön eğitiminde kullanılan metin derlemleri

Korpus	**Kelime öğesi sayısı**	**Boyut (GB)**	**Kaynak**	**Domain**
(trW)Türkçe Web Derlem	4,404,976,662	35	Wikipedia gibi Web kaynakları	Genel
(trM)Türkçe Medikal derlem	60,318,554	0,48	www.dergipark.com.tr (DergiPark)	Biyomedikal
(trR)Türkçe Radyoloji tezleri	15,268,779	0,11	www.tez.yok.gov.tr (YÖK)	Radyoloji
(trK)Kafa BT Raporları	0,036	4,177,140	Ege Üniversitesi Nöroloji ve Acil Servis Departmanı	Klinik Radyoloji

periyodik hakemli dergilere erişim sağlayan bir ağ geçididir. Bir yazılım ile Dergipark sitesi altındaki tüm medikal dergiler ziyaret edilerek dergilerde yayınlanan tüm tam metin pdf makaleleri toplanmıştır. Daha sonra toplanan pdf dokümanlarını ABioNER'e (Boudjellal et al., 2021) benzer bazı sezgisel kurallara dayanarak gerekli olan metin ayrıştırılmıştır. Örneğin, yazılan bir sezgisel kural ile Türkçe makalelerde başlangıç kısmı "özet" (abstract) ve bitiş kısmı "referanslar" (references) arasındaki metin verileri yayınlardan çekilmiştir. Yapılandırılmamış pdf'lerden metin çıkarmak için tüm kuralları tanımlamak zor ve zaman alıcıdır. Gerekli metni bu şekilde elde ettikten sonra ham metin verilerine özel adımlar içeren bir temizleme işlemlerinden oluşan adımlar uygulanmıştır. İlk olarak tüm veriler satır başına bir cümle olacak şekilde tek bir büyük metin dosyasında birleştirilmiştir. Daha sonra dil algılama ve manuel yazılmış sezgisel kurallar ile metin işlenmiştir. Bu kurallar çok yüksek rakam veya noktalama işareti oranı, Türkçe olmayan alfabe karakterleri veya düşük ortalama sözcük sayıları gibi kalıpları tanımlayarak metnin gürülütüden filtrelenmesini sağlmaıştır. Son olarak tekrar eden içerikten kaçınmak için kalan derlemler tekilleştirilmiştir.

trR ile ifade edilen ikinci derlem, radyoloji raporlarının sınıflandırılması görevi ile ilgili metnin etkisini değerlendirmeyi amaçlamaktadır. Değerlendirme derleminin oluşturulması için radyoloji ile ilgili açık alan metinlerini araştırılmıştır ve Yükseköğretim Kurulunun tüm açık doktora tezlerini aramak ve erişmek için bir web sitesi kullanılmıştır. Site üzerinde Tıp fakültelerinin radyoloji bölümlerinde yapılan tüm tezleri filtrelenerek tek bir klasör altında toplanmıştır. Daha sonra toplanan tüm tezler tekrardan tek bir metin dosyasında birleştirilmiş ve trM derleminde uygulanan temizleme işlemleri uygulanarak radyoloji tezlerinden oluşan derlem oluşturulmuştur. Ayrıca bu çalışma, doktora tezlerini dil modelimi geliştiriminde görevle ilgili alan derlemi olarak kullanmaya yönelik literatürdeki ilk girişimdir.

Tez kapsamında oluşturulan son derlem ise sınıflandırma görevi ile aynı veri dağılımını içeren Türkçe Kafa BT raporlarıdır (trK). Bu veri kümesi ile sınırlı görev odaklı dil kaynağının dil modeli geliştiriminde nasıl daha

etkili kullanılacağının araştırması yapılmıştır. 'trK' veri kümesi için Ocak 2016 ile Haziran 2018 tarihleri arasında Ege Üniversitesi Hastanesi nöroloji ve acil servislerinden 8 yaş ve üzeri hastalara ait bilgisayarlı tomografi (BT) incelemelerine ilişkin 40.306 doğrulanmış Türk radyoloji raporu toplanmıştır. Veri analizinden önce, 100 karakterden az karakter içeren raporlar hariç tutulmuş ve tutarlılık için satırsonu karakterleri ve radyolojiye özgü kodlamalar kaldırılmıştır. Son olarak tüm metin verileri kimliksizleştirme ve tekillleştirme işlemlerinden geçirilmiştir. Hasta gizliliği nedenlerinden dolayı bu veri kümesi araştırmacılara açılmamıştır.

4.2 Ön işleme

Tokenizasyon, ham metni özel sözcük dizilerine ayırmayı amaçlayan ve dil modeli eğitimi için kritik bir adım olan bir ön işleme sürecidir. Bu işlem, dil modellemesinde kullanılan tüm model tipleri için ortak bir adımdır ve özellikle karmaşık yapıya sahip dillerde oldukça önem taşır. Türkçe gibi sondan eklemeli ve morfolojik açıdan zengin dillerde tokenizasyon süreci ekstra bir öneme sahip olabilir. Türkçe'nin bu zengin morfolojik yapısı, belirli bir kökten çok sayıda farklı anlam içerebilecek kelimeler üretir. Dil işleme perspektifinden bakıldığında, bu, kelime dışı sorunlarına yol açabilir ve modelin eğitim doğruluğunu düşürebilir. Wordpiece algoritması, bu tür dillerde kelime dışı problemi hafifletmek için güçlü bir araç olarak kabul edilir ve birçok çalışmada Türkçe'nin dil işleme görevlerinde en yüksek performansa ulaştığı kanıtlanmıştır Toraman et al. (2022). Tez kapsamında geliştirilen tüm dil modellerinde, Türkçe'nin bu özel ve zengin morfolojik yapısına uygun bir çözüm olarak Wordpiece algoritması kullanılmıştır. Bu seçim, özellikle dilin karakteristik yapısını dikkate alarak daha doğru ve etkili bir dil modeli eğitimi sağlamayı amaçlamaktadır.

Dil modelinin ön eğitim tekniğine göre de Wordpiece kelime sözlüğünün kullanımı değişebilir. Özel bir alana transfer edilen bir modelde genellikle o alana özgü sıfırdan bir kelime sözlüğü üretilirken, genel alandaki bir dil modelinin etkinliğini kullanmak istenirse, hazır bir kelime sözlüğü doğrudan kul-

lanılabilir. Bu seçim dil işleme görevinin iyi anlaşılmasını gerektirir çünkü her yaklaşımın modelin başarımı üzerinde önemli etkileri olabilir. Ön işleme metotu dil her bir BERT dil modellerinin oluşturulmasında ayrı bir şekilde dikkatle ele alınmıştır böylece en uygun tokenizasyon stratejisinin seçilmesi sağlanmıştır.

4.3 Model mimarisi ve Ön eğitim için Hiperparametre Seçimi

Tez kapsamında geliştirilen tüm BERT ön eğitim stratejileri için 12 katmanlı dönüştürücü blok, 12 dikkat başlığı ve 110 milyon parametreden oluşan BERT tabanlı mimari kullanılmıştır. Tüm modeller adil bir karşılaştırılma yapılabilmesi için orjinal BERT mimari ile aynı ön eğitim hiperparametreler kullanılarak oluşturulmuştur(Tablo 4.2).

Hiperparametre	**Değer**
Learning rate	1e-4
Batch size	256
Optimizer	Adam
β_1	0.9
β_2	0.999
Warmp up steps	10000
Max sequence length	512
Max prediction per seq	76
Masked MLM probability	0.15
epoch	1000000

Tablo 4.2: BioBERTurk dil modelleri ön eğitim parametre ayarları

4.4 Model ön eğitimi

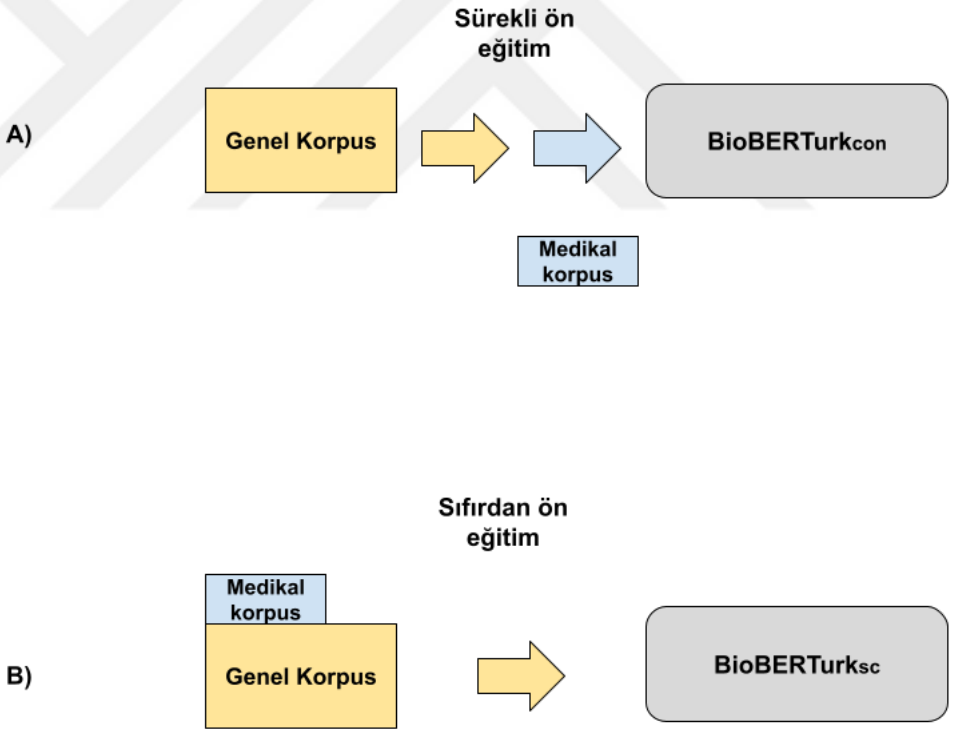
Bu bölümde, BERT dil modelinin tezde uygulanan ön eğitim tekniklerini ayrıntılı bir şekilde incelenmiştir. BERT modeli, kendisinden önceki dil modellerinden farklı olarak gelişmiş dil anlama yetenekleri için bidirectional (iki

yönlü) bir yaklaşım kullanır. Bu özellik BERT'in bir kelimenin hem solundaki hem de sağındaki bağlama bakarak daha iyi bir dil anlaması oluşturmasına olanak sağlar. BERT modelinin ön eğitim süreci büyük çaplı metin verileri üzerinde gerçekleştirilir ve genellikle iki ana görevi içerir: maskelenmiş dil modellemesi (MDM) ve bir sonraki cümle tahmini (SCT). MLM görevlerinde, girdi belirteçlerinin belirli bir yüzdesi rastgele maskelenir ve model, Cloze görevinde (Taylor, 1953) açıklandığı gibi bir cümledeki maskelenmiş belirteçleri tahmin etmektedir. SCT için model, ikinci cümlenin veri kümesindeki ardışık bir cümleyi takip edip etmediğini tahmin etmektedir. Detaylı anlatımı tezin 3. Bölümünde verilmiştir. Bu bölümde tezde kullanılan dil modeli eğitiminde, özellikle BERT gibi transformer tabanlı modellerin ön eğitim tekniklerini odaklanılmıştır. Biyomedikal doğal dil işleme görevlerinde, BERT modellerinin ön eğitimi performansı artırabilir ancak ön eğitimde alana özgü spesifik verilere ihtiyaç duyulmaktadır. Bu alandaki önemli problemlerden birisi biyomedikal metin verileri sınırlı bulunması ve farklı kaynaklara dağılmış olmasıdır. Üstelik halka açık tıbbi veritabanlarının çoğu İngilizce dışında yazılmamıştır. Bu, Türkçe gibi sınırlı dil kaynağı olan dillerde etkili teknikler geliştirmek için yüksek bir talep oluşturmaktadır. Tez çalışmasında, sınırlı dil eğitim kaynağı özgül soruna yanıt olarak, dört farklı BERT ön eğitim tekniği başarıyla uygulanmıştır. Bu uygulama çerçevesinde, çalışma BERT gibi transformer tabanlı modellerin eğitiminde kullanılan dört önemli yaklaşımın incelenmesine odaklanmaktadır: Sürekli ön eğitim, sıfırdan ön eğitim, görev odaklı ön eğitim ve eş zamanlı ön eğitim tekniği. Ön eğitim teknikleriyle uygulanan dil modelleri aşağıda detaylandırılmıştır.

BioBERTurk, iki ayrı versiyonu ile sunulmaktadır: BioBERTurk$_{con}$ ve BioBERTurk$_{sc}$. BioBERTurk$_{con}$, sürekli ön eğitim (contiunal pretraining) yöntemi ile geliştirilmiştir. Bu yaklaşım, modelin geniş bir veri seti üzerinde sürekli olarak eğitilerek ve genelleştirme yeteneklerinin geliştirilmesi hedeflenmiştir. Öte yandan, BioBERTurk$_{sc}$, 'sıfırdan eğitim' (pretraining from scratch) yöntemi kullanılarak geliştirilmiştir. Bu yaklaşımda, modelin sıfırdan öğrenme yeteneğini ve genel amaçlı bir dil modeli oluşturma kapasitesini test

etmek için bir başlangıç noktası oluşturulmuştur. Figür 4.3 A'da sürekli ön eğitim gösteriliyorken, B'de ise sıfırdan ön eğitim gösterilmiştir.

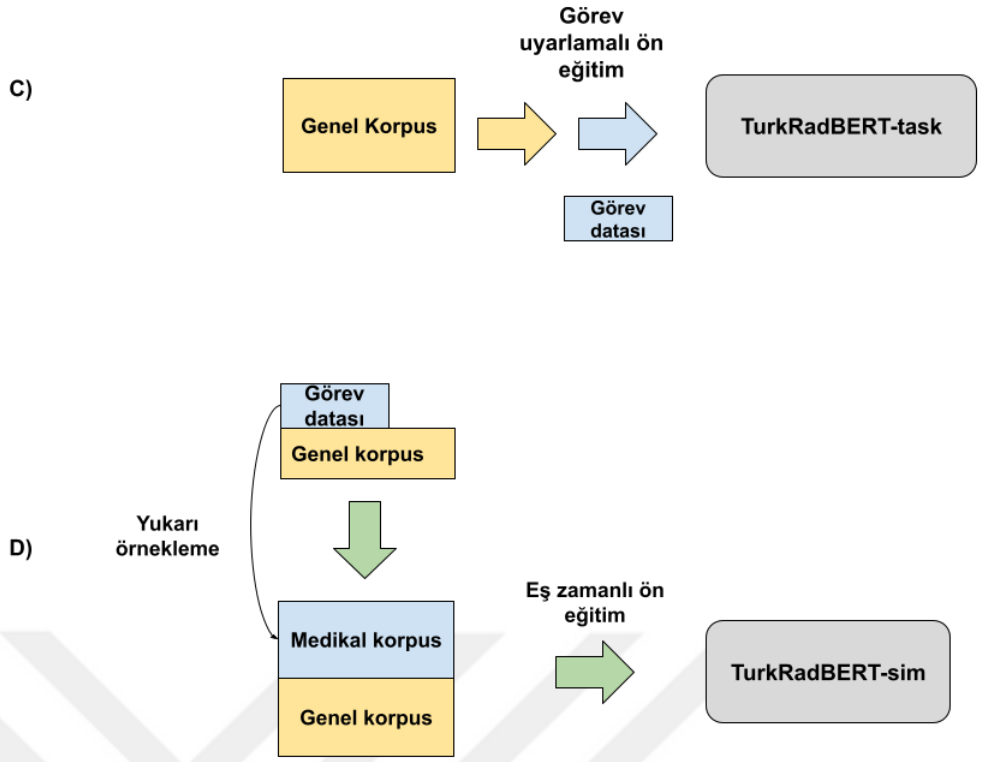
- Sürekli Ön eğitim: Bu teknik, genel veri setlerinden belirli bir göreve veya alan özgü verilere doğru ön eğitim sürecinin genişletildiği bir yaklaşımdır. Burada model, belirli bir uygulama için genel dil anlama becerileri üzerine sürekli olarak daha fazla eğitim alır. Sürekli ön eğitimde modelin genel dil anlama becerilerini korurken, belirli bir görev veya alanla ilgili bilgi ve becerileri geliştirmesini sağlar.
- Sıfırdan ön eğitim: Bu yaklaşımda, model sıfırdan eğitilir yani önceden eğitilmiş bir modelin ağırlıkları kullanılmaz. Tüm model parametreleri rastgele başlatılır ve model belirli bir görevi çözmek için tüm bilgi ve becerilerini verilerden öğrenir.



Şekil 4.3: BioBERTurk ön eğitim yöntemleri

Tez kapsamında geliştirilen bir diğer dil modelleri ailesi olan TurkRadBERT modeli de iki farklı versiyonla geliştirilmiştir: TurkRadBERT-task ve TurkRadBERT-sim. TurkRadBERT-task, görev odaklı ön eğitim (task-adaptive pretraining) yöntemi ile oluşturulmuştur. Bu model, belirli bir göreve daha etkin bir şekilde uyarlanabilmek için hastane içi veriler üzerinde eğitilmiştir. Diğer yandan, TurkRadBERT-sim, eş zamanlı ön eğitim (simultaneous pretraining) yaklaşımı ile oluşturulmuştur. Bu versiyon, görev datasıyla aynı dağılıma sahip kısıtlı veri ile eş zamanlı ön eğitimi kapsamaktadır. Figür 4.4 C'da görev odaklı ön eğitim gösteriliyorken, D'de ise eş zamanlı ön eğitim gösterilmiştir.

- Görev odaklı Ön eğitim: Görev uyarlamalı ön eğitim , belirli bir görev için etiketlenmemiş eğitim seti üzerinde ön eğitim anlamına gelmektedir. Yapılan başka çalışmalarda yöntemin etkinliği kanıtlanmıştır (Howard and Ruder, 2018). BioBERTurk dil modelleri ailesinde kullanılan sürekli ön eğitim ve sıfırdan ön eğitim yöntemleri ile karşılaştırıldığında görev uyarlamalı yaklaşımı daha farklı bir denge kurmaktadır: çok daha küçük bir ön eğitim külliyatı kullanır ancak bu külliyat eğitim setinin görevin yönlerini iyi temsil ettiği varsayımı altında görevle çok daha ilgilidir.
- Eş zamanlı ön eğitim: Alana özgü derlemin genel alan derleminden ufak boyutlu olduğu durumlarda kullanılan bir diğer ön eğitim yöntemi eş zamanlı eğitimdir. Eş zamanlı ön eğitim, dil modeli eğitiminde uygulanan transfer öğrenme yönteminin bir çeşididir ve boyutu dengelemek için veri artırma (yeni sentetik örnekler oluşturma) yöntemi ile az boyuttaki veri büyük boyuttaki veriye eşitlenmesi fikrine dayanmaktadır. Daha sonra eğitim aşamasında genel alan ve alana özgü derlemler eş zamanlı olarak kullanılır. Bu yaklaşımın arkasındaki amaç genel alan derlemindeki dil kalıplarının geniş kapsamından yararlanırken aynı zamanda alana özgü derlemde kapsanan alana özgü bilgiyi de içeren bir model oluşturmaktır



Şekil 4.4: TurkRadBERT ön eğitim yöntemleri

Özetle, BERT dili geliştirilmesinde seçilecek olan yöntemler büyük ölçüde gerçekleştirilecek olan doğal dil işleme görevine ve eldeki kaynaklara (cihaz kaynağı, veri kaynağı vb) bağlıdır. Bu aşama deneysel ve iteratif bir süreç olarak tanımlanabilir ve ayrıca dil modelinin gelişimini doğru bir şekilde gözlemlemek için çeşitli metrikler ve teknolojiler kullanılmaktadır.

5 BioBERTurk: TÜRKÇE BİYOMEDİKAL DİL MODELİ AİLESİ

Bu bölümde, büyük dil modeli olan Türkçe biyomedikal dil modeli ailesi üyelerinin, BioBERTurk, geliştirilmesi ve bunların çeşitli ön eğitim yöntemleriyle karşılaştırılması ele alınmaktadır. Ayrıca, bu modellerin DDİ görevlerindeki performansı ve uygulanabilirliği incelenerek alanında uzmanlar için değerli içgörüler sunulacaktır.

5.1 BioBERTurk dil modelleri ailesi geliştirme süreci

Tez kapsamında iki dil modelleri ailesi bu bölümde detaylandırılmıştır. BioBERTurk dil modelleri ailesi, BioBERTurk$_{con}$, BioBERTurk$_{sc}$ adlı iki modelden oluşmaktadır. Burada, *con* sürekli ön-eğitim (contiunal pretraining) yöntemini, *sc* ise sıfırdan ön-eğitim (pretraining from scratch) yöntemini temsil etmektedir. Her iki model, spesifik eğitim stratejileri kullanılarak geliştirilmiştir ve oluşturlan Türkçe biyomedikal derlemler üzerinde uygulanmıştır. Geliştirilen modeller, Github üzerinden araştırmacıların kullanımına açılmış olup, bu alandaki çalışmaları desteklemek amacıyla tasarlanmıştır.

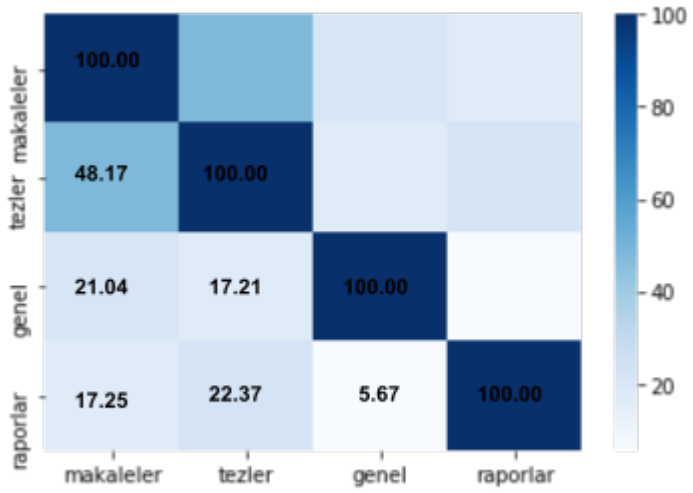
5.1.1 Kullanılan Veri kaynakları

- (trM)Türkçe Medikal derlemi: Türkçe medikal alana özgü metinlerden oluşmakta olup, BioBERTurk$_{con}$(+trM), BioBERTurk$_{con}$(+trM+trR), ve BioBERTurk$_{sc}$(+trW+trM+trR) dil modellerinde kullanılmıştır.
- (trR)Türkçe Radyoloji tezleri derlemi: Türkçe radyoloji tezlerinden oluşan spesifik derlemdir.BioBERTurk$_{con}$(+trR), BioBERTurk$_{con}$(+trM +trR), ve BioBERTurk$_{sc}$(+trW+trM+trR) modellerinin geliştirilmesinde kullanılan bir kaynaktır.
- (trW)Türkçe Web Derlemi: BioBERTurk$_{sc}$(+trW+trM+trR) modelinin sıfırdan eğitiminde kullanılmıştır, aynı zamanda BERTurk'ün sıfırdan ön eğitiminde de kullanılmış olan bir hazır derlemdir.

Bu derlemlerin oluşturulma süreci, detayları ve kullanım amaçları, Tezin 4. Bölümünde daha ayrıntılı olarak ele alınmıştır. Her bir derlem özelleştirilmiş dil modeli eğitiminin farklı yönlerini yansıtmak üzere tasarlanmıştır.

5.1.2 Alan Benzerliğinin Ölçülmesi

Dil modelleri ön eğitime tabi tutulmadan önce daha detaylı analiz için toplanan metin verilerinin ile hedef görev etki alanı (klinik radyoloji) arasındaki benzerliği hesaplanmıştır. Etki alanlarının benzerliğini, etki alanı kelime dağarcıklarının kesişim oranını hesaplanarak değerlendirilmiştir. Bu yaklaşımın altında yatan varsayım, etki alanları arasında paylaşılan kelime sayısının ne kadar benzer olduklarını göstermesidir (Dai et al., 2019a). Durdurma sözcükleri, noktalama işaretleri ve sayılar çıkarıldıktan sonra en sık kullanılan 10 bin unigramı içeren alan sözlüklerini dikkate alınmıştır. Ayrıca, kelime dağarcığı oluşturmak için her bir BERT alanının derlemindeki belgelerin rastgele örneklerinden 100 bin cümle kullanılmıtşır. Görev sözlüğü için, çok daha kısa oldukları için 50 bin radyoloji raporu izlenimi kullanılmıştır. Şekil 5.1, etki alanları arasındaki paylaşılan kelime oranını göstermektedir. Hesaplanan ölçütler, Türkçe tıbbi makaleler alanının Türkçe radyoloji tezleri ile güçlü bir kelime örtüşmesine sahip olduğunu göstermektedir. Makalelerin etki alanı tezlerden daha kapsamlı olmasına rağmen benzerdir çünkü benzer bir tenora sahip olduğu görülmüştür. Ayrıca, hedef alanın radyoloji alanı nedeniyle tez alanına en çok benzeyen (%22,37) ve genel alana en az benzeyen (%5,67) alan olduğunu gözlemlenmiştir. Dolayısıyla, toplanan tez derlemi, raporlarla tamamen benzer bir jargona sahip olmasa bile, ön eğitimli dil modeli oluşturmada görevle ilgili derlem kullanımının etkisini gözlemlemek için diğerlerine kıyasla daha uygun (%22,37) görünmektedir.



Şekil 5.1: Alanlar arasında kelime dağarcığı örtüşme oranı (%)

5.1.3 Ön işleme

Tez kapsamında geliştirilen BioBERTurk aileinden BioBERTurk$_{con}$'un ön eğitim ve ince ayar girdilerini önceden işlemek için BERTurk'ten Wordpiece algoritması kullanılmıştır. Öte yandan, BioBERTurk$_{sc}$'nin ön işlemesi için yeni bir Wordpiece kelime sözlüğü oluşturmuştur. Ayrıca kelime sözlüğü oluşturulma sında HuggingFace[2]'daki tokenizer kütüphanesi kullanılmış ve BERTurk yapılan dırma dosyasında tanımlanan boyutu ayarlamak için sözcük sözlüğünün boyutunu 32k olarak ayarlanmıştır. Daha sonra tüm ham BERT girdisini yapılandırılmış tensorflow örneklerine dönüştürmek için Google AI Research ekibi tarafından sağlanan resmi *create_pretraining_data.py* python kodu kullanılmıştır.

5.1.4 Model mimarisi ve ön-eğitim için hiperparametre seçimi

Tez kapsamında geliştirilen BioBERTurk dil modelleri, Bölüm 4'te detaylandırıldığı üzere BERT tabanlı mimari üzerine inşa edilmiştir. BioBERTurk$_{sc}$ modeli, karma derlemleri kullanarak sıfırdan ön eğitim sürecine tabi tutulurken, BioBERTurk$_{con}$ varyantları, BERTurk kontrol noktalarının TensorFlow

[2]https://huggingface.co/docs/tokenizers/python/latest/

versiyonu kullanılarak sürekli ön eğitime başlatılmıştır. Eğitim parametreleri, daha önce Bölüm 4'te açıklandığı şekilde, maksimum 512 dizi uzunluğu ve 128 parti boyutuyla, toplamda 1 milyon adım süresince belirlenmiştir. Google Cloud Compute Services'ten temin edilen 8 çekirdekli V3 TPU'ları kullanarak gerçekleştirilen bu eğitim, resmi BERT GitHub deposunda bulunan açık kaynaklı eğitim komut dosyalarıyla yönetilmiştir. BioBERTurk$_{con}$ versiyonların eğitim yaklaşık 3 gün sürerken, BioBERTurk$_{sc}$ modeli yaklaşık 1 hafta sürmüştür. BERT mimarisi dışında ayrıca, karşılaştırmalı sınıflandırma performansını göstermek için Barash et al. (2020)'da sunulduğu gibi bir temel model oluşturulmuştur. Bu model, bu tez çalışmasındaki deneylerde kullanılan aynı görev olan ancak İngilizce olmayan bir dilde (İbranice) kafa BT incelemelerinin radyoloji raporlarını sınıflandırmak için kullanılmıştır. Model bir dikkat katmanı ve bunun üzerinde tam bağlantılı bir katmanla birleşitirilmiş bir LSTM katmanına sahiptir. Girdi olarak oluşturulan türkçe biyomedikal derleminden elde edilen word2vec kelime vektörlerini almaktadır. Tez çalışmasının devamında oluşturulan temel model LSTM-attn-wvc olarak adlandırılmıştır.

5.1.5 Geliştirilen dil modelleri

Tez kapsamında geliştirilen dil modelleri aşağıda listelenmiştir:

- **BioBERTurk$_{con}$(+trM):** Yalnızca Türkçe biyomedikal metin kullanarak, mevcut genel Türkçe BERTurk ağırlıklarıyla başlatılarak sürekli eğitim yaklaşımı uygulanmıştır. Bu model için ön eğitim ve ince ayar girdilerini önceden işlemek üzere BERTurk'ten Wordpiece sözlüğü kullanılmıştır.
- **BioBERTurk$_{con}$(+trR):** Sürekli ön eğitim için yalnızca radyoloji tezleri derlemi kullanılarak oluşturulmuştur. Bu model için de BERTurk Wordpiece sözlüğü kullanılmıştır.
- **BioBERTurk$_{con}$(+trM+trR):** Türkçe biyomedikal metne radyoloji tezlerinden oluşan bir derlem eklenerek ve sürekli ön eğitim yoluyla

oluşturulmuştur. BERTurk Wordpiece sözlüğü kullanılmıştır.

- **BioBERTurk$_{sc}$(+trW+trM+trR):** Türkçe biyomedikal ve radyoloji tez derlemleri ile genel alan derleminden oluşan karma bir derlem kullanılarak sıfırdan eğitilmiştir. Bu model için yeni bir Wordpiece kelime sözlüğü oluşturulmuştur.

Bu modellerin geliştirilmesi, bölüm 4'te açıklanan metodoloji ve hiperparametre seçimleri doğrultusunda gerçekleştirilmiştir.

5.2 Model Performansı ve Değerlendirme

Oluşturulan modellerin performanslarının değerlendirilmesinde Bölüm 2'de bahsedilen BT raporlarının bulugular ve izlenim olarak iki kümeye bölünmüş ve "normal", "abnormal" ve "seri dışı" olarak sınıflandığı veri kümeleri kullanılmıştır. Veri kümelerinin detaylı açıklaması için Bölüm 2'ye bakınız.

5.2.1 İnce-ayar yöntemi için hiperparametre seçimi

Dil modellerinin ince ayarı, Devlin et al. (2018) ile aynı mimari ve optimizasyon yöntemi kullanılarak yapılmıştır. Her model için öğrenme hızı ϵ {2e-4, 3e-5, 5e-5}, maksimum dizi uzunluğu ϵ {128, 256, 512}, yığın boyutu ϵ {16, 32} ve eğitim epok sayısı ϵ {3, 4, 5} değerleri için hiperparametre araması gerçekleştirilmiştir. Modeller için seçilen parametreler aşağıdaki Tablo 4.2 'da verilmiştir. Bellek sınırlamaları nedeniyle 64 batch boyutu kullanılmamıştır. Tüm deneylerde Adam optimizer kullanılmıştır. İnce ayar yöntemi NVIDIA Quadro RTX 8000 grafik kartları ile gerçekleştirilmiş ve her deney yaklaşık 10 dakika sürmüştür.

Temel modele girdi olarak, Barash et al. (2020) çalışmasında belirtildiği gibi 200 boyutlu word2vec kelime vektörleri kullanılmıştır. Vektörler ayrıca CBOW mimarisi kullanılarak Gensim çerçevesi tarafından eğitilmiştir Mikolov et al. (2013). Temel model için bir dizi parametre kombinasyonunu

değerlendirilerek, maksimum uzunluğu 128, yığın boyutunu 16 ve eğitimi 25 epok seçilmiştir.

5.2.2 Karşılaştırmalı performans sonuçları ve analizi

Deneyler izlenim ve bulgular veri kümeleri üzerinde ayrı gerçekleştirilerek tüm metrikler her model için en iyi hiperparametre ayarlarında verilmiştir. Tablo 5.1, izlenim veri kümesi için on çalıştırma üzerinden ortalama F1 puanlarını sunmaktadır. Sonuçlara göre, tüm BERT varyantları temel modelden önemli ölçüde daha iyi performans göstermiştir. Ayrıca, alan içi model BioBERTurk$_{con}$+(trR), BERTurk (P değeri 2.46e-05), BioBERTurk$_{sc}$ (P değeri 1.77e-11) ve çok dilli BERT'ten (mBERT) (P değeri 9.11e-07) istatistiksel olarak daha yüksek F1-skoru elde etmiştir. BioBERTurk$_{con}$+(trR), BioBERTurk$_{con}$+(trM)'den daha iyi performans göstermesine rağmen bu modeller arasında istatistiksel bir fark yoktur (P değeri 0,59). Ayrıca, dilin radyoloji raporu metin sınıflandırması üzerindeki etkisini ölçmek için tüm Türkçe BERT modellerini mBERT ile karşılaştırılmıştır. Bazı çalışmalar mBERT'in belirli görevler için tek dilli BERT modellerinden daha sağlam performans gösterdiğini ortaya koymuş olsa da Schneider et al. (2020); Muller et al. (2021), tez çalışması BioBERTurk$_{con}$ varyantlarının ve BERTurk modelinin Türkçe radyoloji raporlarının sınıfını daha doğru bir şekilde belirlediğini göstermektedir.

Ayrıntılı analiz için sınıf başına F1-skorları da Tablo 5.2'te raporlanmıştır. "Normal" ve "Seri Dışı" sınıfları için en yüksek F1-skoru BioBERTurk$_{con}$+(trR) tarafından elde edilirken, "Anormal" sınıfı için BioBERTurk$_{con}$+(trM) tarafından elde edilmiştir. Ayrıca en iyi model olan BioBERTurk$_{con}$+(trR)'nin diğerlerine göre daha yüksek hassasiyet ve duyarlılık skoru elde ettiğini gözlemlenmiştir (Tablo 5.2).

Model	Hassasiyet	Duyarlılık	F1 skor
BERTurk *+trW(c)*[1]	91.88%	91.87%	91.86%
BioBERTurk$_{con}$ *+trM(c)*	93.00%	93.02%	92.99%
BioBERTurk$_{con}$ *+(trM+trR)(c)*	92.74%	92.77%	92.75%
BioBERTurk$_{con}$ *+trR(c)*	**93.13**%	**93.14**%	**93.13%**
BioBERTurk$_{sc}$ *+(trW+trM+trR)(u)*[2]	89.52%	89.51%	89.48%
mBERT(c)	91.45%	91.43%	91.42%
LSTM-attn-wvc	80.80%	82.00%	80.72%

Tablo 5.1: İzlenim veri kümesinin test setine dayalı radyoloji raporu sınıflandırma deneylerinin hassasiyet, duyarlılık ve F1-skoru

Model	Normal	Abnormal	Seri Dışı
BERTurk *+trW(c)*[1]	94.24%	90.61%	85.39%
BioBERTurk$_{con}$ *+trM(c)*	94.97%	**93.02**%	85.91%
BioBERTurk$_{con}$ *+(trM+trR)(c)*	94.80%	92.84%	85.29%
BioBERTurk$_{con}$ *+trR(c)*	**95.11%**	92.17%	**87.59%**
BioBERTurk$_{sc}$ *+(trW+trM+trR)(u)*	92.13%	89.33%	80.33%
mBERT(c)	93.63%	91.06%	84.15%
LSTM-attn-wvc	88.40%	84.71%	47.75%

Tablo 5.2: İzlenim veri kümesinin test setinde etiket başına F1 skoru

Tablo 5.3, bulgular veri kümesi için on çalıştırma üzerinden ortalama F1 puanlarını göstermektedir. Bu deneylerin ilk net gözlemi tüm modellerin bulgular verilerinde daha kötü performans gösterdiğidir. Benzer sonuçlar İngilizce Gundogdu et al. (2021) verisinde de gözlemlenmiştir. Alınan sonuçlar, bulgular verisinin izlenim verisine göre daha uzun ve sınıflandırma açısından daha az bilgilendirici olması nedeniyle beklenen bir durum olduğuna karar verilmiştir. Bulgular veri kümesinde, BioBERTurk$_{con}$+(trM) %89,97'lik en yüksek F1 skoru ile iyi performans göstermiş ve onu istatistiksel olarak anlamlı bir fark olmaksızın BioBERTurk$_{con}$+(trM+trR) izlemiştir (P değeri

0,02). Ayrıca tüm BERT varyantları temel modele göre önemli ölçüde daha iyi performans göstermiştir. Tablo 5.4'teki diğer metrikleri incelediğimizde, BioBERTurk$_{con}$+(trM) modeli "Normal" sınıfı için çok etkili bir performans sergilemiş ancak şaşırtıcı bir şekilde "Anormal" ve "Seri Dışı" sınıflarında başarılı olamamıştır.

Tablo 5.3: Bulgular veri kümesinin test setine dayalı radyoloji raporu sınıflandırma deneylerinin hassasiyet, duyarlılık ve F1-skoru

Model	**Hassasiyet**	**Duyarlılık**	**F1 skor**
BERTurk *+trW(c)*	89.00%	88.55%	88.60%
BioBERTurk$_{con}$ *+(trM)*	**90.34%**	**89.98%**	**89.97%**
BioBERTurk$_{con}$ *+(trM+trR)(c)*	88.93%	89.35%	89.38%
BioBERTurk$_{con}$ *+trR(c)*	88.61%	88.76%	88.75%
LSTM-attn-wvc	82.49%	83.01%	82.61%

Tablo 5.4: Bulgular veri kümesinin test setinde etiket başına F1 skoru

Model	**Normal**	**Abnormal**	**Seri Dışı**
BERTurk *+trW(c)*	89.55%	91.57%	76.22%
BioBERTurk$_{con}$ *+(trM)*	**92.75%**	91.17%	77.94%
BioBERTurk$_{con}$ *+(trM+trR)(c)*	89.85%	**92.25%**	**78.17%**
BioBERTurk$_{con}$ *+trR(c)*	89.17%	91.88%	76.69%
LSTM-attn-wvc	84.71%	88.49%	57.83%

5.2.3 Tartışma

Deneyleri genel olarak değerlendirildiğinde çalışmadan aşağıdaki sonuçlar çıkarılmıştır.

İlk olarak, sonuçlar tüm BioBERTurk$_{con}$ varyantlarının her iki veri kümesinde de mevcut genel BERT modelinden ve geleneksel temel modelden daha iyi sonuçlar verdiğini göstermektedir. Benzer sonuçlar, alan içi modellerin genel modellerden daha iyi performans gösterdiği İngilizce'de de gözlemlenmiştir

Alsentzer et al. (2019); Lee et al. (2020). Ancak, bu tez çalışmasında, genel derlemle karşılaştırıldığında çok küçük boyutlu bir alan içi derlemle ön eğitime devam etmek, klinik görevi sınıflandırmada hala etkili olduğu görülmüştür. Benzer sonuçları tıbbi makaleler ve radyoloji tezleri derlemlerinde de gözlemlenmiştir ancak tez çalışmasında oluşturulan buvderlemler PubMed özetlerinden daha uzun ve daha gürültülü veriler içermektedir

Bir diğer kritik gözlem de tez derleminin sürekli ön eğitimdeki etkisidir. Tez derlemi, tıbbi makale derlemine kıyasla çok küçük olmasına rağmen (0,11 GB'a karşı 0,48 GB) sadece tezler derlemi ile sürekli eğitilen BioBERTurk$_{con}$+(trR) modeli istatistiksel olarak diğer modellerle benzer performans göstermiştir. Bu sonuçlar görev alanına az ölçüde benzeyen bir derlemin küçük boyutlu, gürültülü ve uzun metin verilerinde bile etkili olabileceğini göstermektedir. Ayrıca BioBERTurk$_{con}$+(trR) öodeli, gösterim veri kümesinde "Seri Dışı" etiketini sınıflandırmada diğer modellerden daha iyi performans göstermiştir. Radyoloji tezleri derlemi tıbbi makalelerle birleştirildiğinde, sonuç modeli (BioBERTurk$_{con}$+(trM+trR)) özellikle bulgular veri kümesindeki "Anormal" ve "Seri Dışı" etiketlerini sınıflandırmada etkili bir performans sergilemiştir. Bu nedenle, görevle ilgili küçük verilerle sürekli ön eğitimin, Türk radyoloji raporlarının düşük frekanslı etiket (Seri Dışı) sınıflandırması için daha iyi doğruluk sağladığı sonucuna varılmıştır.

BERT ön eğitim tekniğini araştırmak için BioBERTurk$_{sc}$ modeli diğer modeller ile karşılaştırıldığında temel model hariç her iki veri kümesi için de kötü sınıflandırma doğruluğu vermektedir. Dolayısıyla çok küçük etki alanı verilerini büyük genel verilerle birleştirmek, en azından sıfırdan Türkçe etki alanı odaklı ön eğitimli model üretiminde etkili bir yaklaşım değildir.

Yukarıdaki sonuçlar ışığında, hedef alanın genel alandan önemli ölçüde farklı olması durumunda (Türkçe genel alan ile Türkçe klinik alanın benzerliği %9'dur), sürekli ön eğitim tekniğinde görevle ilgili verilerin kullanılmasının sınıflandırma performansını artırabileceğini gösterilmiştir.

Düşük kaynaklı diller için önemli sıfır atışlı (zero-shot) diller arası transfer yeteneklerine sahip olan mBERT'i uygulayan daha önce de bir çalışma

yapılmıştır Wu and Dredze (2019). mBERT'in F1skoru diğer modellerle karşılaştırıldığında, sürekli ön eğitim kullanılarak geliştirilen tek dilli modeller Türkçe klinik görevinde daha iyi sonuçlar vermektedir. Özetle alan içi modellerimizin başarısı, biyomedikal makalelerin sürekli ön eğitiminin, mevcut dil kaynakları kısıtlı olsa bile, Türkçe klinik bir görevde model performansını artırabileceğini göstermektedir.

Son olarak ve en önemlisi, tez çalışması ile ilk Türkçe biyomedikal kaynakları tanıtılmış ve kaynaklar DDİ topluluğunun kullanımına sunulmuştur.

Çalışmanın bazı kısıtları da bulunmaktadır. Türkçe'de sınıflandırma için tıp alannında bir DDİ görevi bulunmadığından alan içi modeller Türkçe'deki tek bir klinik görev için değerlendirilmiştir. İkinci olarak, BERT modelinin gerektirdiği girdi boyutu sınırı olan 512 karakter olmasından dolayı daha uzun raporlar girdi olarak kullanılamamıştır.

6 TurkRadBERT: TÜRKÇE KLİNİK DİL MODELİ AİLESİ

Tezin Bu bölümünde, klinik alandaki Türkçe metinlerin derinlemesine analizi ve işlenmesi için özel olarak tasarlanmış bir dil modeli olan TurkRadBERTuk'un gerçek dünya uygulaması ele alınmaktadır. Bölüm 4'te metodolojik olarak incelendiği üzere, TurkRadBERT dil modeli ailesi, sınırlı kaynaklı bir alandaki zorluklara özel çözümler sunma amacıyla dil modellemesi araştırmaları için detaylı karşılaştırmalar sunmaktadır. açmaktadır.

6.1 TurkRadBERT dil modelleri ailesi geliştirme süreci

TurkRadBERT dil modelleri, eşzamanlı ön eğitim ve görev uyarlamalı ön eğitim olmak üzere iki ön eğitim stratejisi ile dört Türkçe klinik dil modelini içermektedir. TurkRadBERT-sim ailesi olarak adlandırılan iki model (TurkRadBERT-sim v1 ve TurkRadBERT-sim v2), farklı kelime sözlüğünü kullanırken genel, biyomedikal ve klinik görev derlemlerini birleştiren eşzamanlı ön eğitim teknikleri kullanılarak geliştirilmiştir. Eş zamanlı ön eğitim, modelin büyük ölçekli metinler üzerinde eğitim alarak dil temsillerini öğrenmesini sağlamaktadır. Ancak bu yaklaşım, çok fazla miktarda veri içermesi nedeniyle pahalı olabilir. Son olarak, yalnızca küçük klinik görev verileri kullanarak göreve uyarlanabilir ön eğitim yöntemini (Gururangan et al., 2020) uygulanarak TurkRadBERT-task dil modelleri oluşturulmuştur. Bu teknik diğerlerine kıyasla daha az kaynak yoğun bir yöntemdir. Aşağıdaki bölümlerde TurkRadBERT dil modelleri ailesinin geliştirilmesindeki süreçler Bölüm 4'de anlatılan çerçevede detaylandırılmıştır.

6.1.1 Kullanılan veri kaynakları

Çeşitli dil modellerinin geliştirilmesinde modellerin belirli bir alana ve eldeki göreve uygun olmasını sağlamak için birden fazla derlem kullanılmıştır. Uygun derlemlerin seçimi, alana özgü dil kalıplarını, yapılarını ve sözcük

dağarcıklarını anlamalarını doğrudan etkilediği için dil modellerinin performansı açısından çok önemlidir. Aşağıda TurkRADBERT dil modellerinin geliştirilmesinde kullanılan veri kaynakları özetlenmiştir.

- Türkçe karma derlem:(trM)Türkçe Medikal derlemi: Türkçe medikal derlem (trM), Türkçe Elektronik Radyoloji Tezleri (trR) ve teknik sözdizimi öğrenmek için kullanılan daha özelleştirilmiş bir veri kümesi olan Türkçe Kafa BT Raporlarından oluşan derlemi (trK) içeren karma veri kümesidir. TurkRadBERT-sim v1 ve v2 dil modellerinin oluşturulmasında kullanılmıştır. Veri kaynaklarının oluşturulması Bölüm 4'de özetlenmiştir.
- Türkçe Kafa BT Raporları (trK): Özellikle kafa BT raporlarının dil anlamasını ve özgül jargonunu öğrenmek için kullanılan 30 MB'lık bir veri kümesidir. Bu veri kümesi TurkRadBERT-task v1 ve v2 dil modelerinin ön eğitim için tek başına kullanılmıştır.

6.1.2 Ön işleme

Eş zamanlı ön eğitimde BERT mimarisinde girdi olarak verilecek veriler diğer BERT ön eğitim yöntemlerinden farklıdır. Bu yöntemin açık kaynak kodu olmadığından ve Japonca üzerine yapılan çalışmada (Wada et al., 2020) sistem GPU cihaza göre tasarlandığından dolayı, BERT'e beslenecek tüm mühendislik süreçleri Google Cloud TPU'lar için tekrardan tasarlanmış ve CPU çekirdeği i8 kullanılarak uygulanmıştır. Bu ön eğitim tekniğinde küçük tıbbi derlemi ve büyük genel derlemi eşit boyutta daha küçük belgelere bölünerek yapılandırılmış girdiler oluşturmak için birleştirilmiştir. Bu yaklaşım, küçük tıbbi veriler içeren MDM örnekleri için ön eğitim sıklığını artırarak sınırlı veri boyutundan kaynaklanan potansiyel aşırı uyumu azaltmaktadır. Uygulamada gerçekleşitirilen bu adımların sözde kodu aşağıda Figür 6.1 gösterilmiştir. BERT mimarisinde sözde kodda gösterilen modifikasyonlar Google AI Research ekibi [3] tarafından sağlanan *create_pretraining_data.py* komut dosyası üzerinde gerçekleştirilmiştir ve Git-

[3]https://github.com/google-research/bert

hub üzerinden kullanıcılara açılmıştır [4]. Ayrıca (Wada et al., 2020) tarafından yapılan aynı çalışmaya uygun olarak, TurkRadBERT-sim v1 modeli için güçlendirilmiş kelime sözlüğü olarak adlandırılan alana özgü bir kelime sözlüğü oluşturmak için alana özgü oluşturulmuş metin ve Wordpiece algoritması kullanılmıştır. TurkRadBERT-sim v2, TurkRadBERT-task v1 ve TurkRadBERT-task v2 modelleri için BERTurk modelinin Wordpiece kelime sözlüğü kullanılmıştır. Adil bir karşılaştırma için sıfırdan oluşturulan kelime sözlüğünün sözlük yapılandırma dosyası BERTurk ile aynıdır. BERT'in kelime sözlüğünün eşzamanlı ön eğitim ve sıfırdan ön eğitimde oluşturmak için Huggignface [5] kütüphanesi uygulanmıştır. Görev odaklı ön eğitim ile oluşturulan TurkRadBERT-task modelleri için işlenmiş metin belgelerini TPU cihazlarıyla uyumlu TensorFlow örneklerine dönüştürmekde tekrardan Google AI Research ekibi [6] tarafından sağlanan *create_pretraining_data.py* komut dosyası kullanılmıştır.

Algorithm 1: Eş zamanlı ön-eğitim sözde kodu

Result: output_files = tf_records examples

```
Scorpus = SmallCorpus;
Lcorpus = LargeCorpus;
tokenizer = BertWordPieceTokenizer();
tokenizer.train( corpus=(Scorpus * int(Lcorpus.filesize / Scorpus.filesize)+
  Lcorpus);
split_Sdocs = split_corpus(corpus=Scorpus, each_file_size=10MB);
split_Ldocs = split_corpus(corpus=Lcorpus, each_file_size=10MB);
for _ in range (n.corpus) do
    docs = (random.sample(split_Sdocs,2)+ random.sample(split_Ldocs,2);
    merged_files = merging (docs);
    instance = create_training_instance(merged_files, tokenizer,
      max_seq_length, masked_lm_prob, max_predictions_per_seq);
    write_instance_to_example_files(instance, tokenizer, max_seq_length,
      max_predictions_per_seq, output_files)
end
```

Şekil 6.1: Eş zamanlı ön eğitim sözde kodu

[4] https://github.com/hazalturkmen/BERT-pretraining-techniques-on-TPU-pods

[5] https://huggingface.co/docs/tokenizers/python/latest/

[6] https://github.com/google-research/bert

6.1.3 Model mimarisi ve ön-eğitim için hiperparametre seçimi

Bölüm 4'te belirtildiği gibi, tezde oluşturulan TurkRadBERT dil modelleri, BERT temelli bir yapıyı temel alarak geliştirilmiştir. Eğitim parametreleri de tüm modeller ile aynı olan Bölüm 4'deki ayarlamalardır.

Eş zamanlı ön eğitim ile oluşturulan TurkRadBERT-sim modelleri rasgele eğitime başlatılırken, görev odaklı ön eğitimle oluşuturulan TurkRadBERT-task modelleri BERTurk modelinin ve BioBERTurk modelinin tensorflow kontrol noktasıyla eğitime başlatılmıştır. Ayrıca TurkRadBERT modellerinin tümü BioBERTurk modellerinden farklı olarak Google Cloud Compute Services [7] tarafından sağlanan 32 çekirdekli V3 TPU pod cihazları kullanılarak resmi BERT GitHub deposunda bulunan açık kaynaklı eğitim komut dosyalarıyla eğitilmiştir. Google Cloud TPU-pod, Google'ın özel olarak tasarladığı ve yüksek hızlı ağ arayüzleri ile bağlanarak dağıtık bir yapı oluşturan Tensör İşlem Birimleri (TPU) cihazlarıdır.

6.1.4 Geliştirilen dil modelleri

Eşzamanlı ön-eğitim tekniği, küçük bir alan içi derlemi kullanarak uygulanan ilk ön-eğitim yöntemidir. Tez kapsamında kelime sözlüğü kullanımına göre farklı TurkRadBERT-sim modelleri üretilmiştir. İki model arasındaki temel fark, kelime sözlüğü kullanımlarında yatmaktadır; ilk model güçlendirilmiş, alana özgü bir kelime sözlüğünden yararlanırken, ikincisi BERTurk kelime sözlüğünü kullanmaktadır.

TurkRadBERT-sim v1 oluşturulmasında BERTurk'ü geliştirmek için kullanılan büyük bir Türkçe genel derlemin (35 GB) yanı sıra daha küçük olarak karma bir derlem kullanmıştır (Türkçe biyomedikal derlem, Türkçe Elektronik Radyoloji Tezleri ve Türkçe Kafa BT Raporları). Model eşzamanlı ön eğitim için oluşturulan derlemden oluşturulan güçlendirilmiş alan içi bir kelime sözlüğünü kullanmıştır.

TurkRadBERT-sim v2 Eş zamanlı olarak ön eğitim ile oluşturulmuştur.

[7] https://cloud.google.com/

Model, ön eğitim sırasında v1 ile aynı derlemi kullanmıştır. İki model arasındaki fark, alana özgü kelime sözlüğünün etkisini gözlemlemek için genel alan kelime sözlüğünü kullanılmasıdır.

Son ön eğitim yöntemi, Bölüm 4'de ayrıntılı bahsedilen Kafa BT raporları veri kümesi (trK) üzerinde göreve uyarlanabilir ön eğitimdir. Model başlatma noktasına göre iki farklı BERT modeli geliştirilmiştir.

TurkRadBERT-task v1, model başlatma için Türkçe için genel bir alan dil modeli olan BERTurk'ü kullanmıştır ve ardından göreve uyarlanabilir bir ön eğitim yöntemi gerçekleştirmiştir. Kelime sözlüğü ise BERTurk dil modelinden kullanılmıştır.

TurkRadBERT-task v2 modeli ön yükleme için Türkçe bir biyomedikal BERT modeli olan BioBERTurk variant(Turkmen et al., 2022) kullanmıştır. Bu Türkçe biyomedikal BERT, radyoloji raporlarını sınıflandırmada en iyi skoru elde ettiği için seçilmiştir (Turkmen et al., 2022). Kelime sözlüğü için tekrardan BERTurk modelinden kullanılmıştır.

6.2 Model Performansı ve Değerlendirme

Model performansının değerlendirilmesi için Bölüm 2'de bahsedilen 2000 Türkçe kafa BT raporunu kullanarak oluşturulan çok etiketli veri kümesi kullanılmıştır. Veri kümesiyle oluşturulan rapor sınıflandırma görevinin amacı serbest metin formatında sunulan bir radyoloji raporunda klinik olarak önemli gözlemlerin varlığını belirlemektir. Bunlar 'İntraventriküler', 'Gliosis', 'Epidural', 'Hidrosefali', 'Ensefalomalazi', 'Kronik iskemik değişiklikler', 'Lakuna', 'Lökoaraiosis', 'Mega sisterna magna', 'Meningioma', 'SAK', 'Subdural', 'Bulgu Yok'. Sınıflandırma süreci rapordaki cümlelerin gözden geçirilmesini ve olumlu ya da olumsuz olmak üzere iki sınıftan birine dahil edilmesini içermektedir. 13. gözlem olan "Bulgu Yok", herhangi bir bulgunun olmadığını göstermektedir. Uzmanlar tarafından şemadaki 12 etiket kontrast öncesi kraniyal bilgisayarlı tomografi (BT) incelemesinde tespit edilebilecek başlıca ve nispeten yaygın klinik patolojileri belirtmek üzere seçilmiştir. Ayrıca,

çalışmada kullanılan 12 etiket de belirsiz radyolojik bulgular değil, kesin klinik patolojilerdir. Bu nedenle, radyoloji uzmanları veri kümesini bu açıklama şemasına göre doküman düzeyinde etiketlemiştir.

6.2.1 İnce-ayar yöntemi için hiperparametre seçimi

Tüm ön eğitimli modellerin ince ayarı daha önce (Devlin et al., 2018) çalışmasında kullanılan aynı mimari ve optimizasyon yöntemleri kullanılarak birbirinden bağımsız olarak gerçekleştirilmiştir. İnce ayar sürecinde amaç oluşturulan rapor sınıflandırma görevinde mevcut son teknoloji performansını aşmak değil Türkçe klinik dil modelleri geliştirmek için ön eğitim tekniklerini değerlendirmek ve karşılaştırmaktır. Bu nedenle hiperparametrelerin kapsamlı bir araştırması yapılmamıştır. TurkRadBERT-sim ve TurkRadBERT-task modelleri için kullanılan konfigürasyonlar sırasıyla aşağıda Tablo 6.1 ve Tablo 6.2'da gösterilmektedir.

Önceden eğitilmiş farklı BERT modellerinin klinik çok etiketli sınıflandırma görevi üzerindeki etkinliğ en uygun hiperparametre ayarları kullanılarak on çalıştırma boyunca ortalama kesinlik, duyarlılık ve F1 skoru hesaplanarak değerlendirilmiştir.

Parametre	**TurkRadBERT-sim**
Learning rate	5e-5
Batch size	32
Optimizer	Adam
Max sequence length	512
epoch	20

Tablo 6.1: TurkRadBERT-sim dil modelleri ailesi için ince ayar konfigürasyonu

Parametre	TurkRadBERT-task
Learning rate	3e-5
Batch size	32
Optimizer	Adam
Max sequence length	512
epoch	15

Tablo 6.2: TurkRadBERT-task dil modelleri ailesi için ince ayar konfigürasyonu

6.2.2 Karşılaştırmalı performans sonuçları ve analizi

Türkçe klinik çok etiketli sınıflandırmada BERTurk, TurkRadBERT-task v1, TurkRadBERT-task v2, TurkRadBERT-sim v1 ve TurkRadBERT-sim v2 olmak üzere beş farklı modelin performansını değerlendirilmiştir. Ortalama kesinlik, duyarlılık ve F1 skoru açısından modellerin on çalıştırma üzerinden performanslarını karşılaştırılmıştır. Ayrıca, en başarılı iki modelin (BERTurk, TurkRadBERT-task v1) hasta bulgularının performansının F1 skorlarına göre daha detaylı analiz gerçekleştirilmiştir. Sonuçlar 6.3 ve 6.4 tablolarında sunulmuştur.

Tablo 6.3, BERTurk'ün 0,9738 kesinlik ve 0,9456 duyarlılık ile 0,9562 F1 puanı elde ettiğini göstermektedir. TurkRadBERT-task v1, 0,9556 ile biraz daha düşük bir F1 skoruna sahiptir. Her iki model de sınıflandırma görevinde güçlü performans gösterirken, BERTurk genel F1 puanı açısından TurkRadBERT-görev v1'den biraz daha iyi performans göstermiştir. BERTurk, TurkRadBERT-görev v1'den daha iyi performans göstermiş olsa da, bu modeller arasında istatistiksel bir fark yoktur (P değeri 0,255). TurkRadBERT -task v2, TurkRadBERT-sim v1 ve TurkRadBERT-sim v2 modelleri, BERTurk ve TurkRadBERT-task v1'e kıyasla daha düşük genel performans göstermektedir.

Model	Kesinlik	Duyarlılık	F1 skor
BERTurk	**0.9738**	0.9456	**0.9562**
TurkRadBERT-task v1	0.9736	**0.9462**	0.9556
TurkRadBERT-task v2	0.9643	0.9352	0.9470
TurkRadBERT-sim v1	0.8613	0.7969	0.8149
TurkRadBERT-sim v2	0.8170	0.7863	0.7879

Tablo 6.3: Her model için hassasiyet, duyarlılık ve F1-skoru

Bununla birlikte, modellerin güçlü ve zayıf yönlerinin daha derin bir şekilde anlaşılmasını sağlamak için modellerin performansını her bir etiket için değerlendirmek önemlidir. Tablo 6.4, BERTurk ve TurkRadBERT-task v1 için her kategorinin F1 puanlarını göstermektedir. Sonuçlar modellerin performansının kategoriler arasında değiştiğini ve bazı etiketlerin iki model arasında F1 skorlarında belirgin bir fark gösterdiğini ortaya koymaktadır. BERTurk aşağıdaki kategorilerde TurkRadBERT-task v1'den daha iyi performans göstermektedir: İntraventriküler, Gliosis, Epidural, Leukoaraiosis, Mega cisterna magna ve Bulgu Yok. Buna karşılık, TurkRadBERT-task v1 Hidrosefali, Ensefalomalazi, Kronik iskemik değişiklikler, SAK ve Subdural kategorilerinde BERTurk'ten daha iyi performans göstermektedir. Lacuna ve Meningioma için F1 skorları her iki model için de aynıdır.

Category	BERTurk	TurkRadBERT-task v1
İntraventriküler	0.4815	0.4000
Gliosis	0.8580	0.8155
Epidural	0.9012	0.9000
Hidrosefali	0.9458	0.9673
Ensefalomalazi	0.9622	0.9633
Kronik iskemik değişiklikler	0.9918	0.9921
Lakuna	0.9655	0.9655
Lökoaraiosis	0.8995	0.8762
Mega cisterna magna	0.6000	0.4500
Meningioma	1.0000	1.0000
SAK	0.9281	0.9544
Subdural	0.9666	0.9757
Bulgu Yok	0.9455	0.9311

Tablo 6.4: Her model için Kesinlik, hassasiyet ve F1-skoru

6.2.3 Tartışma

Deneyleri bir bütün olarak değerlendirildiğinde aşağıdaki sonuçlara ulaşılmaktadır.

Eş zamanlı ön eğitim ile göreve uyarlanabilir ön eğitim karşılaştırıldığında görev verileri ile genel veriler arasındaki boyut farkı nedeniyle, sınırlı alana özgü verilerin büyük genel alan verileri tarafından gölgede bırakılabildiği ve modelin göreve özgü özelliklerden ziyade genel özellikleri öğrenmeye daha fazla odaklanmasına neden olduğu gözlemlenmiştir. Bu olgu modelin özel alanın nüanslarını etkili bir şekilde yakalamasını sağlamak için ön eğitim sürecinde genel ve alana özgü verilerin dikkatlice dengelenmesinin önemini vurgulamaktadır.

BERTurk ve TurkRadBERT-task v1 modellerinin performansı oldukça yakındır çünkü her iki model de ön eğitim sırasında geniş genel alan derleminden elde edilen bilgiden yararlanmaktadır. BERTurk doğrudan bu büyük

derlem üzerinde ön eğitime tabi tutulurken TurkRadBERT-task v1, BERTurk'ün ağırlıkları ile başlatılır ve ardından daha küçük bir klinik derlem üzerinde göreve uyarlanabilir ön eğitim kullanılarak ince ayar yapılır. Bu ince ayar TurkRadBERT-task v1'in genel alan verisinde bulunmayan alana özgü kalıpları, yapıları ve terminolojileri yakalamasını sağlamıştır.

Bununla birlikte göreve uyarlanabilir ön eğitimde kullanılan göreve özgü küçük derlem, modelin alana özgü bilgiyi öğrenmesini sınırlayabilir. Sonuç olarak, göreve uyarlanabilir ön eğitimin faydalarına rağmen TurkRadBERT-task v1 (bu yaklaşımı kullanan) BERTurk'ten biraz daha düşük performansa sahiptir. Sınırlı veri senaryolarında göreve uyarlanabilir ön eğitim yaklaşımı, özellikle göreve özgü küçük bir derlem üzerinde ön eğitim yapıldığında aşırı uyum sağlamaya eğilimli olabilmektedir. Model, eğitim verilerine aşırı özelleşebilir ve görülmeyen örnekler üzerinde iyi genelleme yapamayabilir (Zhang et al., 2022).

Performans açısından, TurkRadBERT-task v1, TurkRadBERT-task v2'den (0,9470) biraz daha yüksek bir F1 puanına (0,9556) sahiptir. Bu durum, BioBERTurk'teki daha özelleşmiş biyomedikal bilgiye rağmen genel alan BERTurk modelinin bu spesifik klinik görevde göreve uyarlanabilir ön eğitim için daha sağlam bir temel sağladığını göstermiştir.

Bu çalışmada ulaşılan bir diğer sonuç ise TurkRadBERT-sim v1 ve v2 arasındaki karşılaştırmanın, alana özgü kelime dağarcığının model performansı üzerindeki etkisine dair içgörüler sunmasıdır. Oluşturulan derlemden oluşturulan güçlendirilmiş bir kelime sözlüğü kullanan TurkRadBERT-sim v1, genel alan kelime sözlüğünü kullanan TurkRadBERT-sim v2'den daha iyi performans göstermiştir. Bu bulgu ön eğitim sırasında alana özgü bir kelime sözlüğü kullanmanın, modelin alana özgü dil kalıplarını yakalama ve anlama yeteneğini artırabileceğini ve sonuçta klinik DDİ görevlerinde daha iyi performansa yol açabileceğini göstermiştir.

Tablo 6.4'daki her bir etiket için F1 puanlarının incelenmesi, en başarılı iki modelin performansı hakkında daha ayrıntılı bir bakış açısı sağlamaktadır. BERTurk, İntraventriküler, Gliozis, Epidural, Lökoaraiozis, Mega sisterna

magna ve Bulgu Yok gibi belirli etiketlerde TurkRadBERT-task v1'den daha iyi performans göstermektedir. BERTurk'ün belirli etiketlerdeki daha yüksek performansı, ön eğitim sırasında edindiği genel alan bilgisine atfedilebilir; bu da belirli kategoriler için, özellikle de göreve özgü derlemde daha düşük frekansa sahip olanlar için daha iyi kapsama alanı sağlayabilir. BERTurk'ün daha geniş ön eğitim verilerine maruz kalması TurkRadBERT-task v1 BERTurk ile başlatılmış olsa bile göreve özgü derlemde daha düşük temsile sahip belirli etiketlerle uğraşırken TurkRadBERT-task v1 gibi modellere göre potansiyel olarak bir avantaj sağlayabilir. Bu durum genel alan bilgisi ve göreve özgü ince ayar kombinasyonunun farklı kategorilerde optimum performans için kritik olabileceğini göstermektedir. Öte yandan TurkRadBERT-task v1, Hidrosefali, Ensefalomalazi, Subaraknoid Kanama ve Subdural gibi etiketler için üstün performans sergilemektedir. Bu, göreve uyarlanabilir ön eğitimin modele alana özgü bilgiler üzerinde ince ayar yaparak bazı durumlarda performans artışı sağlayabileceğini göstermektedir. Bununla birlikte iki model arasındaki genel performans farklarının nispeten küçük olduğunu belirtmek gerekir; bu da bu modellerde hem genel alana hem de göreve özgü bilgiden yararlanmanın önemini vurgulamaktadır.

7 SONUÇ VE GELECEK ÇALIŞMALAR

Bu tez boyunca Türkçe klinik hasta raporlarının dil modelleri ile sınıflandırılmasının potansiyel katkılarını açılayarak başlanmış ve Türkçe dilinde geliştirilen ilk biyomedikal ve klinik BDM olma özelliğini taşıyan BioBERTurk ve TurkRADBERT dil modelleri ailesinin gelişimi ve etkinliği gösterilmiştir. Ayrıca tezde BERT mimarisi için çeşitli ön eğitim metodolojilerini araştırılmış, etkileri gözlemlenmiş ve sınırlı veri senaryolarıyla ilişkili zorluklar tartışılmıştır. Sunulan modeller, radyoloji raporlarının çok etiketli sınıflandırılması gibi görevlerde performansı artırarak alana özgü verilerden yararlanmayı başarıyla başarmıştır. Aşağıda geliştirilen iki dil modeli ailesinin radyoloji raporlarını sınıflandırmasında çıkarılan sonuçlardan bahsedilmiştir.

- BioBERTurk: Bu çalışmanın merkezinde, Türkçe radyoloji raporlarının sınıflandırılması için önceden eğitilmiş dört biyomedikal dil modelinin sunulduğu BioBERTurk ailesi yer almaktadır. Yapılan analizler radyoloji alanına özgü küçük ölçekli bir veri kümesiyle önceden eğitilmiş BioBERTurk$_{con}$ modelin Türkçe radyoloji raporlarının sınıflandırılmasın da genel BERT dil modelinden daha iyi bir performans sergilediğini göstermektedir.
- TurkRadBERT: Bu çalışma, radyoloji raporu sınıflandırma görevindeki performanslarına göre BERTurk, TurkRadBERT-task v1 ve v2, TurkRadBERT-sim v1 ve v2 modellerini kapsamlı şekilde karşılaştırmak tadır. Bulgular, BERTurk modelinin genel performans bakımından en yüksek başarıyı gösterdiğini, hemen ardından TurkRadBERT-task v1 modelinin geldiğini göstermektedir. Bu sonuçlar, karmaşık görevlerde en yüksek performansı elde etmek için, önceden eğitim sırasında elde edilen genel-domain bilgisinin ve görev özgü ince ayarın birlikte kullanılmasının önemini vurgulamaktadır.

Türkçe dilinde radyoloji raporlarının analizi için geliştirdiğimiz dil modelleri, büyük miktarda verinin hızlı ve doğru bir şekilde işlenmesini ve radyoloji

raporlarının otomatik analizini sağlamıştır. Sonuçlar, geliştirdiğimiz modelin karmaşık tıbbi terminolojiyi ve serbest metin formatındaki radyoloji raporlarını anlamlandırmakta ve bu bilgileri klinisyenlere kullanılabilir formatta sunmakta etkili olduğunu göstermiştir.

Tez kapsamında geliştirilen modeller, sağlık profesyonellerinin retrospektif çalışmalarda daha fazla veriye hızlı bir şekilde erişmesini sağlayarak, daha doğru ve bilgilendirilmiş klinik kararlar verilmesine yardımcı olabilecek doğruluktadır. Ayrıca, modeller, klinik uygulamalar ve araştırmalar için değerli bir araç olmuştur.

7.1 Gelecek Çalışmalar

Bu tezde elde edilen bulgular ve gözlemler, dil modellemesinde ön eğitim stratejilerinin ve genel ile alana özgü bilgilerin dengesinin önemini vurgulamaktadır. Bu bulguların ışığında, gelecekteki araştırmalar Türkçe gibi kaynak kısıtlı diller için büyük veri kümelerinin oluşturulmasına ve bu verilerin yeni nesil dil modellerinin geliştirilmesine odaklanabilir.

Bir gelecek çalışma önerisi, dil modellemesi alanında daha büyük veri kümelerinin oluşturulması yoluyla Türkçe için daha etkili ve geniş kapsamlı dil modellerinin geliştirilmesidir. Bu büyük dil modellerini kullanarak açık İngilizce metin verilerinin Türkçe'ye çevirisini yaparak gerçekleştirilebilir. Gerçekleştirelecek olan çeviri süreci daha geniş ve çeşitli bir Türkçe derlem oluşturabilir. Bu tür bir derlem, dil modellemesinde geniş çerçevede kullanılabilir ve Türkçe için daha güçlü ve daha kapsamlı dil modellerinin oluşturulmasına yardımcı olabilir. Özellikle böyle bir çalışma Türkçe'nin düşük kaynaklı bir dil olması nedeniyle dil modellemesi araştırmalarında önemli bir adım olabilir.

Bununla birlikte çeviri ile oluşturulan büyük veri kümelerinden transformer mimarilerinden farklı olarak yeni nesil Türkçe dil modelleri oluşturulurken çeviri sürecinin model performansı üzerindeki etkisinin de incelenmesi gerekmektedir. Çeviri süreci, dilin bazı nüanslarını ve kültürel bağlamlarını kaybedebilir ve bu durum modelin performansını ve genelleştirme yeteneğini

etkileyebilir(Lyu et al., 2023). Örneğin yapılan bir çalışmada radyologların ChatGPT dil modelinin kullanımıyla Fransızca'dan İngilizce'ye çeviride dil nüanslarının ve kültürel referansların farklı olabileceğini vurgulayarak ana dili İngilizce olan bir kişi tarafından yapılan çevirinin ikinci bir bakış ile kontrol edilmesinin iyi olabileceğini vurgulamıştır (Lecler et al., 2023). Bu nedenle, çeviri sürecinin ve oluşturulan derlemlerin kalitesinin, modelin genel performansı ve etkinliği üzerindeki etkilerini ölçmek ve anlamak için kapsamlı bir analiz ve değerlendirme yapılması gerekmektedir. Yapılacak olan byöle bir çalışma, dil modellemesi alanında daha derin bir anlayış sağlamaya ve Türkçe dil modellemesi için daha etkili stratejiler geliştirmeye yardımcı olabilir.

KAYNAKLAR DİZİNİ

Allahyari, M., **Pouriyeh, S.**, **Assefi, M.**, **Safaei, S.**, **Trippe, E.D.**, **Gutierrez, J.B.**, and **Kochut, K.**, 2017, Text summarization techniques: a brief survey, *arXiv preprint arXiv:1707.02268*.

Alsentzer, E., **Murphy, J.R.**, **Boag, W.**, **Weng, W.H.**, **Jin, D.**, **Naumann, T.**, and **McDermott, M.**, 2019, Publicly available clinical BERT embeddings, *arXiv preprint arXiv:1904.03323*.

Bahdanau, D., **Cho, K.**, and **Bengio, Y.**, 2014, Neural machine translation by jointly learning to align and translate, *arXiv preprint arXiv:1409.0473*.

Barash, Y., **Guralnik, G.**, **Tau, N.**, **Soffer, S.**, **Levy, T.**, **Shimon, O.**, **Zimlichman, E.**, **Konen, E.**, and **Klang, E.**, 2020, Comparison of deep learning models for natural language processing-based classification of non-English head CT reports, *Neuroradiology*, 62(10), 1247–1256.

Bates, J., **Fodeh, S.J.**, **Brandt, C.A.**, and **Womack, J.A.**, 2016, Classification of radiology reports for falls in an HIV study cohort, *Journal of the American Medical Informatics Association*, 23(e1), e113–e117.

Beltagy, I., **Lo, K.**, and **Cohan, A.**, 2019, SciBERT: A pretrained language model for scientific text, *arXiv preprint arXiv:1903.10676*.

Bengio, Y., **Ducharme, R.**, and **Vincent, P.**, 2000, A neural probabilistic language model, *Advances in neural information processing systems*, 13.

Boudjellal, N., **Zhang, H.**, **Khan, A.**, **Ahmad, A.**, **Naseem, R.**, **Shang, J.**, and **Dai, L.**, 2021, ABioNER: a BERT-based model for Arabic biomedical named-entity recognition, *Complexity*, 2021.

Bressem, K.K., **Adams, L.C.**, **Gaudin, R.A.**, **Tröltzsch, D.**, **Hamm, B.**, **Makowski, M.R.**, **Schüle, C.Y.**, **Vahldiek, J.L.**, and **Niehues, S.M.**, 2020, Highly accurate classification of chest radiographic reports using a

deep learning natural language model pre-trained on 3.8 million text reports, *Bioinformatics*, 36(21), 5255–5261.

Carter, E.J., **Pouch, S.M.**, and **Larson, E.L.**, 2014, The relationship between emergency department crowding and patient outcomes: a systematic review, *Journal of Nursing Scholarship*, 46(2), 106–115.

Chowdhery, A., **Narang, S.**, **Devlin, J.**, **Bosma, M.**, **Mishra, G.**, **Roberts, A.**, **Barham, P.**, **Chung, H.W.**, **Sutton, C.**, **Gehrmann, S.**, **et al.**, 2022, Palm: Scaling language modeling with pathways, *arXiv preprint arXiv:2204.02311*.

Collobert, R. and **Weston, J.**, 2008, A unified architecture for natural language processing: Deep neural networks with multitask learning, in *Proceedings of the 25th international conference on Machine learning*, pages 160–167.

Cover, T. and **Hart, P.**, 1967, Nearest neighbor pattern classification, *IEEE transactions on information theory*, 13(1), 21–27.

Dai, X., **Karimi, S.**, **Hachey, B.**, and **Paris, C.**, 2019a, Using similarity measures to select pretraining data for NER, *arXiv preprint arXiv:1904.00585*.

Dai, Z., **Yang, Z.**, **Yang, Y.**, **Carbonell, J.**, **Le, Q.V.**, and **Salakhutdinov, R.**, 2019b, Transformer-xl: Attentive language models beyond a fixed-length context, *arXiv preprint arXiv:1901.02860*.

Datta, S., **Si, Y.**, **Rodriguez, L.**, **Shooshan, S.E.**, **Demner-Fushman, D.**, and **Roberts, K.**, 2020, Understanding spatial language in radiology: Representation framework, annotation, and spatial relation extraction from chest X-ray reports using deep learning, *Journal of biomedical informatics*, 108, 103473.

Devlin, J., **Chang, M.W.**, **Lee, K.**, and **Toutanova, K.**, 2018, Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805*.

Dietterich, T.G., 1998, Approximate statistical tests for comparing supervised classification learning algorithms, *Neural computation*, 10(7), 1895–1923.

Drozdov, I., **Forbes, D.**, **Szubert, B.**, **Hall, M.**, **Carlin, C.**, and **Lowe, D.J.**, 2020, Supervised and unsupervised language modelling in Chest X-Ray radiological reports, *Plos one*, 15(3), e0229963.

Dunnick, N.R. and **Langlotz, C.P.**, 2008, The radiology report of the future: a summary of the 2007 Intersociety Conference, *Journal of the American College of Radiology*, 5(5), 626–629.

DURMUŞ, E. and **GÜNEYSU, F.**, 2020, Acil Servis Hekimlerinin Bilgisayarlı Tomografi Tetkikine Yaklasımlarının Degerlendirilmesi, *Klinik Tıp Bilimleri*, 8(1), 1–10.

Garcia, M., 2021, Embeddings in Natural Language Processing: Theory and Advances in Vector Representations of Meaning, *Computational Linguistics*, 47(3), 699–701.

Ginde, A.A., **Foianini, A.**, **Renner, D.M.**, **Valley, M.**, and **Camargo, Jr, C.A.**, 2008, Availability and quality of computed tomography and magnetic resonance imaging equipment in US emergency departments, *Academic emergency medicine*, 15(8), 780–783.

Graves, A., 2013, Generating sequences with recurrent neural networks, *arXiv preprint arXiv:1308.0850*.

Grivas, A., **Alex, B.**, **Grover, C.**, **Tobin, R.**, and **Whiteley, W.**, 2020, Not a cute stroke: analysis of rule-and neural network-based information extraction systems for brain radiology reports, in *Proceedings of the 11th international workshop on health text mining and information analysis*, pages 24–37.

Gu, Y., **Tinn, R.**, **Cheng, H.**, **Lucas, M.**, **Usuyama, N.**, **Liu, X.**, **Naumann, T.**, **Gao, J.**, and **Poon, H.**, 2021, Domain-specific language model

pretraining for biomedical natural language processing, *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1), 1–23.

Gundogdu, B., **Pamuksuz, U.**, **Chung, J.H.**, **Telleria, J.M.**, **Liu, P.**, **Khan, F.**, and **Chang, P.J.**, 2021, Customized Impression Prediction from Radiology Reports Using BERT and LSTMs, *IEEE Transactions on Artificial Intelligence*.

Gururangan, S., **Marasović, A.**, **Swayamdipta, S.**, **Lo, K.**, **Beltagy, I.**, **Downey, D.**, and **Smith, N.A.**, 2020, Don't stop pretraining: Adapt language models to domains and tasks, *arXiv preprint arXiv:2004.10964*.

Hartung, M.P., **Bickle, I.C.**, **Gaillard, F.**, and **Kanne, J.P.**, 2020, How to create a great radiology report, *RadioGraphics*, 40(6), 1658–1670.

He, J., **Baxter, S.L.**, **Xu, J.**, **Xu, J.**, **Zhou, X.**, and **Zhang, K.**, 2019, The practical implementation of artificial intelligence technologies in medicine, *Nature medicine*, 25(1), 30–36.

Hochreiter, S. and **Schmidhuber, J.**, 1997, Long short-term memory, *Neural computation*, 9(8), 1735–1780.

Howard, J. and **Ruder, S.**, 2018, Universal Language Model Fine-tuning for Text Classification.

Hsu, E., **Malagaris, I.**, **Kuo, Y.F.**, **Sultana, R.**, and **Roberts, K.**, 2022, Deep learning-based NLP data pipeline for EHR-scanned document information extraction, *JAMIA open*, 5(2), ooac045.

Joachims, T., 2005, Text categorization with support vector machines: Learning with many relevant features, in *Machine Learning: ECML-98: 10th European Conference on Machine Learning Chemnitz, Germany, April 21–23, 1998 Proceedings*, pages 137–142, Springer.

Johnson, K.D. and **Winkelman, C.**, 2011, The effect of emergency department crowding on patient outcomes: a literature review, *Advanced emergency nursing journal*, 33(1), 39–54.

Kanwal, S., **Khan, F.**, **Alamri, S.**, **Dashtipur, K.**, and **Gogate, M.**, 2022, COVID-opt-aiNet: A clinical decision support system for COVID-19 detection, *International Journal of Imaging Systems and Technology*, 32(2), 444–461.

Lecler, A., **Duron, L.**, and **Soyer, P.**, 2023, Revolutionizing radiology with GPT-based models: Current applications, future possibilities and limitations of ChatGPT, *Diagnostic and Interventional Imaging*, 104(6), 269–274.

Lee, J., **Yoon, W.**, **Kim, S.**, **Kim, D.**, **Kim, S.**, **So, C.H.**, and **Kang, J.**, 2020, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, *Bioinformatics*, 36(4), 1234–1240.

Lewis, M., **Liu, Y.**, **Goyal, N.**, **Ghazvininejad, M.**, **Mohamed, A.**, **Levy, O.**, **Stoyanov, V.**, and **Zettlemoyer, L.**, 2019, Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, *arXiv preprint arXiv:1910.13461*.

Liu, F., **Weng, C.**, and **Yu, H.**, 2012, Natural language processing, electronic health records, and clinical research, *Clinical Research Informatics*, pages 293–310.

López-Ubeda, P., **Díaz-Galiano, M.C.**, **López, L.A.U.**, **Martín-Valdivia, M.T.**, **Martín-Noguerol, T.**, and **Luna, A.**, 2020, Transfer learning applied to text classification in Spanish radiological reports, in *Proceedings of the LREC 2020 Workshop on Multilingual Biomedical Text Processing (MultilingualBIO 2020)*, pages 29–32.

Lyu, C., **Xu, J.**, and **Wang, L.**, 2023, New trends in machine translation using large language models: Case examples with chatgpt, *arXiv preprint arXiv:2305.01181*.

Maron, M.E., 1961, Automatic indexing: an experimental inquiry, *Journal of the ACM (JACM)*, 8(3), 404–417.

Meng, X., **Ganoe, C.H.**, **Sieberg, R.T.**, **Cheung, Y.Y.**, and **Hassanpour, S.**, 2020, Self-supervised contextual language representation of radiology reports to improve the identification of communication urgency, *AMIA Summits on Translational Science Proceedings*, 2020, 413.

Mikolov, T., **Chen, K.**, **Corrado, G.**, and **Dean, J.**, 2013, Efficient estimation of word representations in vector space, *arXiv preprint arXiv:1301.3781*.

Mujtaba, G., **Shuib, L.**, **Idris, N.**, **Hoo, W.L.**, **Raj, R.G.**, **Khowaja, K.**, **Shaikh, K.**, and **Nweke, H.F.**, 2019, Clinical text classification research trends: systematic literature review and open issues, *Expert systems with applications*, 116, 494–520.

Muller, B., **Elazar, Y.**, **Sagot, B.**, and **Seddah, D.**, 2021, First align, then predict: Understanding the cross-lingual ability of multilingual BERT, *arXiv preprint arXiv:2101.11109*.

Névéol, A., **Dalianis, H.**, **Velupillai, S.**, **Savova, G.**, and **Zweigenbaum, P.**, 2018, Clinical natural language processing in languages other than English: opportunities and challenges, *Journal of biomedical semantics*, 9(1), 1–13.

Nguyen, H. and **Patrick, J.**, 2016, Text mining in clinical domain: Dealing with noise, in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 549–558.

Peng, Y., **Yan, S.**, and **Lu, Z.**, 2019, Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets, *arXiv preprint arXiv:1906.05474*.

Pineda, A.L., **Ye, Y.**, **Visweswaran, S.**, **Cooper, G.F.**, **Wagner, M.M.**, and **Tsui, F.R.**, 2015, Comparison of machine learning classifiers for influenza detection from emergency department free-text reports, *Journal of biomedical informatics*, 58, 60–69.

Qu, W., **Balki, I.**, **Mendez, M.**, **Valen, J.**, **Levman, J.**, and **Tyrrell, P.N.**, 2020, Assessing and mitigating the effects of class imbalance in machine learning with application to X-ray imaging, *International journal of computer assisted radiology and surgery*, 15, 2041–2048.

Rae, J.W., **Borgeaud, S.**, **Cai, T.**, **Millican, K.**, **Hoffmann, J.**, **Song, F.**, **Aslanides, J.**, **Henderson, S.**, **Ring, R.**, **Young, S.**, **et al.**, 2021, Scaling language models: Methods, analysis & insights from training gopher, *arXiv preprint arXiv:2112.11446*.

Raja, A.S., **Ip, I.K.**, **Sodickson, A.D.**, **Walls, R.M.**, **Seltzer, S.E.**, **Kosowsky, J.M.**, and **Khorasani, R.**, 2014, Radiology utilization in the emergency department: trends of the last two decades, *AJR. American journal of roentgenology*, 203(2), 355.

Rani, G.J.J., **Gladis, D.**, and **Mammen, J.**, 2015, Classification and prediction of breast cancer data derived using natural language processing, in *Proceedings of the Third International Symposium on Women in Computing and Informatics*, pages 250–255.

Rumelhart, D.E., **Hinton, G.E.**, and **Williams, R.J.**, 1986, Learning representations by back-propagating errors, *nature*, 323(6088), 533–536.

Sackett, D.L., **Rosenberg, W.M.**, **Gray, J.M.**, **Haynes, R.B.**, and **Richardson, W.S.**, 1996, Evidence based medicine: what it is and what it isn't.

Schneider, E.T.R., **de Souza, J.V.A.**, **Knafou, J.**, **e Oliveira, L.E.S.**, **Copara, J.**, **Gumiel, Y.B.**, **de Oliveira, L.F.A.**, **Paraiso, E.C.**, **Teodoro, D.**, and **Barra, C.M.C.M.**, 2020, Biobertpt-a portuguese neural language model for clinical named entity recognition, in *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 65–72.

Shah, N.H. and **Tenenbaum, J.D.**, 2012, The coming age of data-driven medicine: translational bioinformatics' next frontier.

Shah, S.M. and **Khan, R.A.**, 2020, Secondary use of electronic health record: Opportunities and challenges, *IEEE access*, 8, 136947–136965.

Shamshad, F., **Khan, S.**, **Zamir, S.W.**, **Khan, M.H.**, **Hayat, M.**, **Khan, F.S.**, and **Fu, H.**, 2023, Transformers in medical imaging: A survey, *Medical Image Analysis*, page 102802.

Shin, B., **Chokshi, F.H.**, **Lee, T.**, and **Choi, J.D.**, 2017, Classification of radiology reports using neural attention models, in *2017 international joint conference on neural networks (IJCNN)*, pages 4363–4370, IEEE.

Smit, A., **Jain, S.**, **Rajpurkar, P.**, **Pareek, A.**, **Ng, A.Y.**, and **Lungren, M.P.**, 2020, CheXbert: combining automatic labelers and expert annotations for accurate radiology report labeling using BERT, *arXiv preprint arXiv:2004.09167*.

Stewart, W.F., **Shah, N.R.**, **Selna, M.J.**, **Paulus, R.A.**, and **Walker, J.M.**, 2007, Bridging The Inferential Gap: The Electronic Health Record And Clinical Evidence: Emerging tools can help physicians bridge the gap between knowledge they possess and knowledge they do not., *Health Affairs*, 26(Suppl1), w181–w191.

Taylor, W.L., 1953, "Cloze procedure": A new tool for measuring readability, *Journalism quarterly*, 30(4), 415–433.

Toraman, C., **Yilmaz, E.H.**, **Şahinuç, F.**, and **Ozcelik, O.**, 2022, Impact of Tokenization on Language Models: An Analysis for Turkish, *arXiv preprint arXiv:2204.08832*.

Tsoumakas, G., **Katakis, I.**, and **Vlahavas, I.**, 2010, Mining multi-label data, *Data mining and knowledge discovery handbook*, pages 667–685.

Turkmen, H., **Dikenelli, O.**, **Eraslan, C.**, and **Callı, M.C.**, 2022, Bioberturk: Exploring Turkish Biomedical Language Model Development Strategies in Low Resource Setting.

Türkmen, H., **Dikenelli, O.**, **Eraslan, C.**, **Çalli, M.C.**, and **Ozbek, S.S.**, 2022, Developing Pretrained Language Models for Turkish Biomedical Domain, in *2022 IEEE 10th International Conference on Healthcare Informatics (ICHI)*, pages 597–598, IEEE.

Vaswani, A., **Shazeer, N.**, **Parmar, N.**, **Uszkoreit, J.**, **Jones, L.**, **Gomez, A.N.**, **Kaiser, Ł.**, and **Polosukhin, I.**, 2017, Attention is all you need, *Advances in neural information processing systems*, 30.

Wada, S., **Takeda, T.**, **Manabe, S.**, **Konishi, S.**, **Kamohara, J.**, and **Matsumura, Y.**, 2020, Pre-training technique to localize medical bert and enhance biomedical bert, *arXiv preprint arXiv:2005.07202*.

Weng, C., **Appelbaum, P.**, **Hripcsak, G.**, **Kronish, I.**, **Busacca, L.**, **Davidson, K.W.**, and **Bigger, J.T.**, 2012, Using EHRs to integrate research with patient care: promises and challenges, *Journal of the American Medical Informatics Association*, 19(5), 684–687.

Widyassari, A.P., **Rustad, S.**, **Shidik, G.F.**, **Noersasongko, E.**, **Syukur, A.**, **Affandy, A.**, **et al.**, 2022, Review of automatic text summarization techniques & methods, *Journal of King Saud University-Computer and Information Sciences*, 34(4), 1029–1046.

Wood, D.A., **Lynch, J.**, **Kafiabadi, S.**, **Guilhem, E.**, **Al Busaidi, A.**, **Montvila, A.**, **Varsavsky, T.**, **Siddiqui, J.**, **Gadapa, N.**, **Townend, M.**, **et al.**, 2020, Automated Labelling using an Attention model for Radiology reports of MRI scans (ALARM), in *Medical Imaging with Deep Learning*, pages 811–826, PMLR.

Wu, S. and **Dredze, M.**, 2019, Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT, *arXiv preprint arXiv:1904.09077*.

Zhang, A., **Xing, L.**, **Zou, J.**, and **Wu, J.C.**, 2022, Shifting machine learning for healthcare from development to deployment and from models to data, *Nature Biomedical Engineering*, pages 1–16.

Zhao, W.X., **Zhou, K.**, **Li, J.**, **Tang, T.**, **Wang, X.**, **Hou, Y.**, **Min, Y.**, **Zhang, B.**, **Zhang, J.**, **Dong, Z.**, **et al.**, 2023, A survey of large language models, *arXiv preprint arXiv:2303.18223*.

Zuccon, G., **Wagholikar, A.S.**, **Nguyen, A.N.**, **Butt, L.**, **Chu, K.**, **Martin, S.**, and **Greenslade, J.**, 2013, Automatic classification of free-text radiology reports to identify limb fractures using machine learning and the snomed ct ontology, *AMIA Summits on Translational Science Proceedings*, 2013, 300.

TEŞEKKÜR

Ege Üniversitesi Tıp Fakültesi Radyoloji Bölümün'den Prof. Dr. Süha Süreyya Özbek, Prof. Dr. Cem Çallı ve Doç. Dr Cenk Erarslan hocalarıma verinin oluşturulmasındaki destekleri ve katkılarından dolayı teşekkürler.
Ayrıca tez çalışmasında sağladıkları teknik destek ile uygulamanın gelişmesine katkı sunan Google TRC ekibine teşekkür ederim.

.

12/07/2023

İmzası

Hazal TÜRKMEN

ÖZGEÇMİŞ

Kişisel Bilgiler

Ad ve Soyad: Hazal TÜRKMEN

Eğitim

Doktora: Ege Üniversitesi, Bilgisayar Mühendisliği (2016−2023)

Yüksek Lisans: Kocaeli Üniversitesi, Bilgisayar Mühendisliği (2013−2016)

Lisans: Kocaeli Üniversitesi, Bilgisayar Mühendisliği (2009−2013)

Akademik Tecrübeler

Araştırma Görevlisi, Ege Üniversitesi, Bilgisayar Mühendisliği (2018−)