**T.C.**

**YILDIZ TEKNİK ÜNİVERSİTESİ**

**FEN BİLİMLERİ ENSTİTÜSÜ**

# TÜRKÇE ALT KATEGORİLİ VARLIK İSMİ TANIMA İLE OTOMATİK ETİKETLEME

**Lola KHUDOYBERDİEVA**

YÜKSEK LİSANS TEZİ

Bilgisayar Mühendisliği Anabilim Dalı

Danışman

Prof. Dr. Banu DİRİ

Mayıs, 2024

**T.C.**

**YILDIZ TEKNİK ÜNİVERSİTESİ**

**FEN BİLİMLERİ ENSTİTÜSÜ**

# TÜRKÇE ALT KATEGORİLİ VARLIK İSMİ TANIMA İLE OTOMATİK ETİKETLEME

Lola KHUDOYBERDİEVA tarafından hazırlanan tez çalışması 03.07.2023 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı **YÜKSEK LİSANS** olarak kabul edilmiştir.

Prof. Dr. Banu Diri
Yıldız Teknik Üniversitesi
Danışman

**Jüri Üyeleri**

Prof. Dr. Banu DİRİ, Danışman
Yıldız Teknik Üniversitesi _____

Dr. Öğr. Üyesi Hafıza İrem TÜRKMEN ÇİLİNGİR, Üye
Yıldız Teknik Üniversitesi _____

Doç. Dr. Aysun GÜRAN, Üye
Doğuş Üniversitesi _____

Danışmanım Prof. Dr. Banu DİRİ sorumluluğunda tarafımca hazırlanan "Türkçe Alt Kategorili Varlık İsmi Tanıma ile Otomatik Etiketleme" başlıklı çalışmada veri toplama ve veri kullanımında gerekli yasal izinleri aldığımı, diğer kaynaklardan aldığım bilgileri ana metin ve referanslarda eksiksiz gösterdiğimi, araştırma verilerine ve sonuçlarına ilişkin çarpıtma ve/veya sahtecilik yapmadığımı, çalışmam süresince bilimsel araştırma ve etik ilkelerine uygun davrandığımı beyan ederim. Beyanımın aksinin ispatı halinde her türlü yasal sonucu kabul ederim.

Lola KHUDOYBERDİEVA

İmza

*Aileme*

*ve*

*sevgilime*

# TEŞEKKÜR

Bu çalışmanın gerçekleştirilmesinde her zaman bana destek veren, ne zaman zorlansam elimden tutup yol gösteren, bana güven sağlayan güler yüzlü, saygıdeğer danışman hocam Prof. Dr. Banu Diri'ye sonsuz teşekkür ve sevgilerimi sunuyorum.

Yüksek Lisans sürecinde bana bilgi kazandıran, yardımlarını sakınmayan diğer hocalarıma teşekkürlerimi sunuyorum. Çalışmamda desteğini ve bana olan güvenini benden esirgemeyen Ali Tiner'e ve son olarak canım ablama teşekkürlerimi sunuyorum.

Lola KHUDOYBERDİEVA

# İÇİNDEKİLER

**TEŞEKKÜR** **V**

**KISALTMA LİSTESİ** **VIII**

**ŞEKİL LİSTESİ** **IX**

**TABLO LİSTESİ** **X**

**ÖZET** **XI**

**ABSTRACT** **XII**

**1 GİRİŞ** **13**

1.1 Literatür Taraması ..... 13

1.1.1 Türkçe'de Varlık İsmi Tanıması ..... 14

1.1.2 Alt Kırılımlı Varlık İsmi Tanıması ..... 15

1.1.3 Türkçede Alt Kırılımlı Varlık İsmi Tanıması ..... 16

1.2 Tezin Katkıları ..... 17

1.3 Tezin Ana Hatları ..... 18

**2 VERİ SETİ** **19**

2.1 Veri Seti Oluşturma ..... 19

2.1.1 Etiketleme İşlemi ..... 21

2.2 Varlıkların Açıklaması ..... 22

2.2.1 Alt Kırılımlı Varlıklar ..... 22

2.3 Veri Seti İstatistikleri ..... 26

**3 DERİN ÖĞRENME MODELLERİ** **29**

3.1 BERT ..... 29

3.2 DistilBERT ..... 30

3.3 ConvBERT ..... 31

3.4 XLM-RoBERTa ..... 31

**4 DENEYSEL SONUÇLAR** **32**

4.1 Değerlendirme Metrikleri ..... 32

4.2 Kullanılan Kütüphaneler ..... 33

4.3 BERTurk İle Yapılan Deneyler ..... 33

4.3.1 Üst Varlık ..... 33

4.3.2 Alt Kırılımlı Varlık ..... 33

4.4 DistilBERT İle Yapılan Deneyler ..... 38

4.4.1 Üst Varlık ..... 38
4.4.2 Alt Kırılımlı Varlık ..... 38
4.5 ConvBERTurk İle Yapılan Deneyler ..... 43
4.5.1 Üst Varlık ..... 43
4.5.2 Alt Kırılımlı Varlık ..... 43
4.6 XLM-RoBERTa İle Yapılan Deneyler ..... 48
4.6.1 Üst Varlık ..... 48
4.6.2 Alt Kırılımlı Varlık ..... 48
4.7 Sonuçların Özeti ..... 53
**5 SONUÇ** **56**
**KAYNAKÇA** **57**
**A EK** **61**
**TEZDEN ÜRETİLMİŞ YAYINLAR** **62**

# KISALTMA LİSTESİ

| | |
|---|---|
| DDİ | Doğal Dil İşleme |
| VİT | Varlık İsmi Tanıma |
| MUC | Message Understanding Conferences |
| CoNLL | Conference on Natural Language Learning |
| BIO | Beginning, Inside, Outside |
| BERT | Bidirectional Encoder Representations from Transformers |
| CRF | Conditional Random Fields |
| BiLSTM | Bidirectional Long Short-Term Memory |
| XLM | Cross Language Models |
| RAM | Random Access Memory |
| GPT | Generative Pre-trained Transformers |
| MLM | Masked Language Modelling |
| NSP | Next Sentence Prediction |

# ŞEKİL LİSTESİ

**Şekil 1.1 Varlık isimleri ağaç yapısı ..... 13**
**Şekil 2.1 Süreç blok diyagramı ..... 19**
**Şekil 2.2 Vikipedi bilgi kutusu web ve markdown formatı ..... 20**
**Şekil 3.1 BERT modeli ön eğitim ve ince ayar yapma süreci ..... 30**
**Şekil 4.1 BERTurk ile LOC alt kırılımlı VİT hata matrisi ..... 36**
**Şekil 4.2 BERTurk ile ORG alt kırılımlı VİT hata matrisi ..... 37**
**Şekil 4.3 BERTurk ile PER alt kırılımlı VİT hata matrisi ..... 37**
**Şekil 4.4 DistilBERTurk ile LOC alt kırılımlı VİT hata matrisi ..... 41**
**Şekil 4.5 DistilBERTurk ile ORG alt kırılımlı VİT hata matrisi ..... 42**
**Şekil 4.6 DistilBERTurk ile ORG alt kırılımlı VİT hata matrisi ..... 42**
**Şekil 4.7 ConvBERTurk ile LOC alt kırılımlı VİT hata matrisi ..... 46**
**Şekil 4.8 ConvBERTurk ile ORG alt kırılımlı VİT hata matrisi ..... 47**
**Şekil 4.9 ConvBERTurk ile PER alt kırılımlı VİT hata matrisi ..... 47**
**Şekil 4.10 XLM-R ile LOC alt kırılımlı VİT hata matrisi ..... 51**
**Şekil 4.11 XLM-R ile ORG alt kırılımlı VİT hata matrisi ..... 52**
**Şekil 4.12 XLM-R ile PER alt kırılımlı VİT hata matrisi ..... 52**
**Şekil 4.13 Üst varlıkların ortalama başarı sonuçları ..... 53**
**Şekil 4.14 Alt kırılımlı varlıkların ortalama başarı sonuçları ..... 54**
**Şekil 4.15 Varlık sayısı ve başarı puanı arasındaki ilişki ..... 54**
**Şekil 4.16 Sol: veri seti etiketi, sağ: bağlama bağlı etiket tahmini ..... 55**
**Şekil 4.17 Sol: Veri setindeki eksik etiketleme, sağ: bağlama bağlı etiket tahmini ..... 55**
**Şekil 4.18 Sol: Veri setindeki doğru etiketleme, sağ: hatalı etiket tahmini ..... 55**

# TABLO LİSTESİ

**Tablo 2.1 Etiketsiz veri seti formatı ve örnekler ..... 21**
**Tablo 2.2 LOC alt kategorileri varlıkların açıklaması ..... 23**
**Tablo 2.3 ORG alt kategorileri varlıkların açıklaması ..... 24**
**Tablo 2.4 PER alt kategorileri varlıkların açıklaması ..... 25**
**Tablo 2.5 Üst varlık ile etiketli veri istatistikleri ..... 26**
**Tablo 2.6 DATE ve LOC alt kırılımlı varlık ile etiketli veri istatistikleri ..... 27**
**Tablo 2.7 ORG alt kırılımlı arlık ile etiketli veri istatistikleri ..... 28**
**Tablo 2.8 PER alt kırılımlı varlık ile etiketli veri istatistikleri ..... 28**
**Tablo 4.1 BERTurk ile üst varlık tahmin sonuçları ..... 34**
**Tablo 4.2 BERTurk ile DATE ve LOC alt kırılımlı varlık tahmin sonuçları 34**
**Tablo 4.3 BERTurk ile ORG alt kırılımlı varlık tahmin sonuçları ..... 35**
**Tablo 4.4 BERTurk ile PER alt kırılımlı varlık tahmin sonuçları ..... 35**
**Tablo 4.5 DistilBERTurk ile üst varlık tahmin sonuçları ..... 38**
**Tablo 4.6 DistilBERTurk ile DATE ve LOC alt kırılımlı varlık tahmin sonuçları ..... 39**
**Tablo 4.7 DistilBERTurk ile ORG alt kırılımlı varlık tahmin sonuçları ..... 40**
**Tablo 4.8 DistilBERTurk ile PER alt kırılımlı varlık tahmin sonuçları ..... 40**
**Tablo 4.9 ConvBERTurk ile üst varlık tahmin sonuçları ..... 43**
**Tablo 4.10 ConvBERTurk ile DATE ve LOC alt kırılımlı varlık tahmin sonuçları ..... 44**
**Tablo 4.11 ConvBERTurk ile ORG alt kırılımlı varlık tahmin sonuçları ..... 45**
**Tablo 4.12 ConvBERTurk ile PER alt kırılımlı varlık tahmin sonuçları ..... 45**
**Tablo 4.13 XLM-R ile üst varlık tahmin sonuçları ..... 48**
**Tablo 4.14 XLM-R ile DATE ve LOC alt kırılımlı varlık tahmin sonuçları .. 49**
**Tablo 4.15 XLM-R ile ORG alt kırılımlı varlık tahmin sonuçları ..... 50**
**Tablo 4.16 XLM-R ile PER alt kırılımlı varlık tahmin sonuçları ..... 50**

# ÖZET

## Türkçe Alt Kategorili Varlık İsmi Tanıma ile Otomatik Etiketleme

Lola KHUDOYBERDİEVA

Bilgisayar Mühendisliği Anabilim Dalı

Yüksek Lisans Tezi

Danışman: Prof. Dr. Banu DİRİ

Varlık İsmi Tanıma, metin içerisindeki varlık isimlerini bulma ve etiketleme işlemidir. İlişki çıkarma, bilgi tabanı oluşturma ve soru cevaplama gibi farklı uygulama alanlarında varlıkların alt kategorilerinin de kullanılması ayrıca önemlidir. Ancak, bu kapsamda Türkçe çalışmalar yetersiz kalmaktadır. Bu eksikliği gidermek adına çalışmamızda Türkçe için büyük ölçekli 23 farklı alt kategoriye sahip bir veri kümesi oluşturulmuştur. Bu alt kategoriler kişi, kurum ve yer varlıklarının alt kırılımlarından oluşmaktadır. Çalışmada Türkçe ile eğitilen dönüştürücü tabanlı modeller ile denemeler yapılmış ve üst varlık isimlerinin tespit edilmesinde 93,34 F1 puanı alınmıştır. Bu puan Türkçe alınmış başarılara yakın bir sonuçken, alt kırılımlı kategoriler için başarı performansı 79,82 olmuştur.

**Anahtar Kelimeler:** Varlık İsmi Tanıması; Alt Kategorili Varlıklar; Doğal Dil İşleme.

# ABSTRACT

## Automatic Tagging With Turkish Fine-Grained Named Entity Recognition

Lola KHUDOYBERDIEVA

Department of Computer Engineering

Master of Science

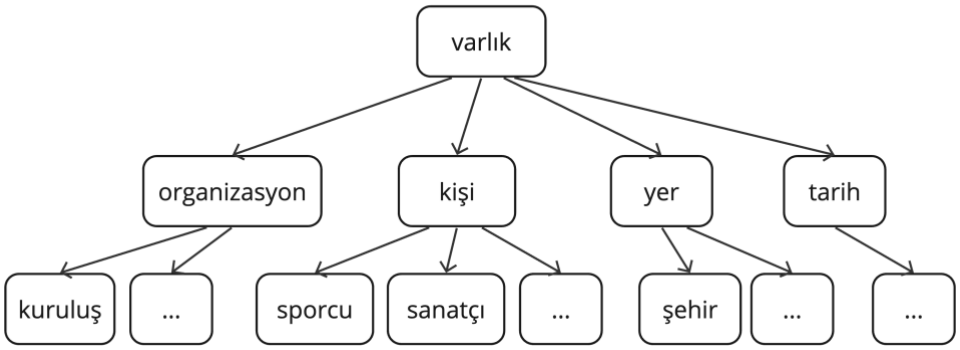
Supervisor: Prof. Dr. Banu DIRI

Named Entity Recognition is the process of finding and labelling entity names in a text. It is also important to use fine-grained entities in different application areas such as relation extraction, knowledge base building and question answering. However, studies in Turkish are insufficient for this context. In order to overcome this deficiency, we created a large-scale dataset with 23 different fine-grained entities for Turkish. These fine-grained entities consist of subcategories of people, institutions and place entities. In this study, experiments were conducted with the transformer-based models in Turkish. While a F1 score of 93.34, which is close to the state-of-the-art for Turkish, was obtained for general entity names, the performance for subcategories was 79.82.

**Keywords:** Named Entity Recognition; fine-grained labels; Natural Language Processing.

# 1
# GİRİŞ

Varlık İsmi Tanıma (VİT) bir bilgi çıkarımı problemi olup, metindeki insan, yer, organizasyon gibi varlıkları otomatik olarak elde etmeyi hedeflemektedir. VİT Doğal Dil İşleme (DDİ) alanındaki soru cevaplama, metin özetleme gibi konularda kullanışlıdır. Üst varlık isimleri Kişi (PER), Yer (LOC), Organizasyon (ORG) gibi tüm alanlarda kullanılabilen isimlerdir. Daha özel alanlarda kullanılabilmesi için alt kategorili varlıklar oluşturulmuştur. Alt kategorili (kırılımlı) varlık isimleri üst varlıklardan türetilmiş olup, onların altındaki kategoride de alt gruplanmış varlık isimleri vardır. Örneğin, Kişi üst varlık ismi olup, Sanatçı, Sporcu gibi alt dallara ayrılabilir (Şekil 1.1).




**Şekil 1.1** Varlık isimleri ağaç yapısı

Türkçe'de üst varlık ismi tanıması üzerine birçok çalışma yapılmıştır ve güncel başarı puanları elde edilmiştir. Literatürde alt kırılımlı varlık ismi veri setleri finans ve genetik gibi özel alanlarda oluşturulmuş olup, bu tip alanlarda geniş kapsamlı veri seti eksiği vardır. Varlık isimlerini alt kategorilere ayırmanın önemi, soru cevaplama, ilişki çıkarma gibi birçok DDİ alanında giderek artmaktadır.

## 1.1 Literatür Taraması

Geleneksel VİT sınırlı sayıdaki kategorilerden oluşturulmuş etiketlerle çalışmaktadır. MUC [1] ve CoNLL [2] gibi konferanslardaki çalışmalarda, sınırlı sayıdaki varlık sınıfları tanımlanmış ve etiketlenmiştir. Ancak, bu yaklaşımlarla genellikle en yaygın varlık türleri ele alınmış ve daha spesifik veya ince ayrıntılı

sınıflandırmalar yapılmamıştır. Alt kırılımlı VİT, üst etiketler altında gruplanabilen, çok sayıdaki etiketlerden oluşmaktadır. Metinlerden daha karmaşık ve çeşitli bilgileri çıkarabilmek için, varlıkları alt kategorilere ayırarak daha net bir sınıfsal ayrım yaratmak önemlidir.

VİT veri seti etiketlenmesinde son yıllarda zayıf gözetimli ve gözetimsiz yaklaşımlar öne çıkmaktadır. Çalışmalarda çoğunlukla Derin Öğrenme ve Koşullu Rastgele Alanlar (Conditional Random Fields – CRF) yöntemlerinden yararlanılmaktadır. Bu bölümde Türkçe VİT çalışmaları ve alt kırılımlı VİT çalışmaları incelenecektir.

### 1.1.1 Türkçe'de Varlık İsmi Tanıması

Son yıllarda üst VİT alanında yapılan çalışmalarda yapay sinir ağı modelleri kullanılmış ve yüksek başarım sonuçları alınmıştır. Şeker ve Eryiğit [3] çalışmalarında düzgün metinler için özelleştirilmiş Koşullu Rasgele Alanlar (CRF) [4] temelinde Türkçe için bir VİT modeli önermişlerdir. Türkçe haberlerden ve Web 2.0'dan toplanan veriler ile oluşturulan iki veri setinden sırasıyla 92 ve 65'lik bir F1 başarısı elde etmişlerdir. Yazarların bir diğer çalışmasında [5] ise haber metinleri, Twitter, konuşmadan metne çeviri ve Donanım forumlarından oluşturulan üç farklı veri seti üzerinde çalışılmıştır. Üç tane farklı modelle yapılan deneylerde haber metinleri üzerinde en yüksek 91,64'lük bir başarı elde edilmiştir.

Güneş ve Tantuğ [6], BiLSTM [7] yapısını kullanarak bir VİT sistemi önermişlerdir. Yazarlar, kullandıkları veri setini sözcük vektörlerinden oluşacak bir dizi halinde kullanmışlardır. BiLSTM tabanlı 5 farklı model geliştirmişlerdir. Yazarlar, deneysel çalışmalarının sonucunda 93,7'lik bir başarı yakalamışlardır.

Safaya vd. [8] Türkçe dili için yedi farklı doğal dil işleme alt alanı belirlemiş ve farklı veri setleri üzerinde önerdikleri yöntemleri çalıştırarak, sonuçlarını referans çalışmalarla karşılaştırmışlardır. Bu alanlardan biri olan Varlık İsmi Tanımada WikiANN [9] ve MİLLİYET_NER [10] veri setleri ile BERT [11] ve BiLSTM tabanlı modelleri kombine ederek eğitmişler ve son katmanına CRF eklenmiş modelleri geliştirmişlerdir. BiLSTM-CRF ile WikiANN üzerinde 93,8'lik bir başarı elde etmişlerdir.

Özçelik vd. [12] Türkçe için hazırlanmış farklı veri setleri ve dönüştürücü modelleri kullanarak varlıkları tanımak için kapsamlı bir çalışma yapmışlardır. Çalışmada,

beş farklı veri seti üzerinde derin öğrenme, dönüştürücü tabanlı ve hibrit modeller ile deneyler yapılmıştır. Yazarlar WikiANN veri setini gürültülü ve tekrarlanan verilerden arındırarak kullanmışlardır. WikiANN Vikipedi metinlerinden oluşan, ENAMEX varlık türlerini kullanan çok dilli bir veri setidir. Yazarlar WikiANN üzerinde ConvBERTurk [13] modeli ile en yüksek 92,2'lik başarı elde etmişlerdir. Haber makalelerinden alınan veri setleri üzerinde ise ELECTRA [14] modeli ile en yüksek 96,1'lik başarı elde etmişlerdir. Bu çalışma, Türkçe'deki VİT çalışmalarını kapsamlı olarak ele alan ve karşılaştırmalar yapan en güncel çalışmadır.

### 1.1.2 Alt Kırılımlı Varlık İsmi Tanıması

İlk alt kırılımlı çalışma İngilizce'de Ling ve Weld [15] tarafından yapılmış olup, çalışmada 112 farklı etikete sahip veri seti üretilmiştir. Bu etiketler çok sayıdaki etiketleri barındıran Freebase veritabanından alınmıştır. Veri gürültüsünü azaltmak için, veri tabanından sadece daha çok tekrar eden etiketler çıkarılmıştır. Çok detaylı olan etiketler tek etiket altında birleştirilmiştir. Çalışmada çok sayıdaki farklı etiketin elle etiketlenmesinin zorluğu vurgulanmış ve bu aşama Vikipedi makalelerindeki linkler kullanılarak etiketleme, gözetimsiz (unsupervised) yöntemle yapılmıştır. Yazarlar etiketli verileri kullanarak CRF tabanlı bir tahmin modeli geliştirmişlerdir ve 93'lük bir F1 başarısı elde etmişlerdir.

Gillick vd. [16] bağlama bağlı alt kırılımlı etiketlerle etiketlenmiş veri seti hazırlamışlardır. Çalışmada bir varlığa verilebilecek etiket çeşidi, cümledeki anlamına bağlı olarak belirlenmiştir. Etiketler ağaç yapısında düzenlenmiştir. Örneğin, kişi varlığı person/artist/actor olarak üç farklı veya person/political-figure olarak iki seviyeye ayrılmıştır. Etiketleme işlemi elle yapılmış olup, 12000'den fazla token bağlama bağlı etiketlenmiştir.

Choi vd. [17] "Ultra Fine Entity" diye adlandırdıkları çalışmalarında 2000'in üzerinde serbest formlu etiket kullanarak bir veri seti oluşturmuşlardır. Etiketleme işlemi *crowdsourcing* kullanılarak elle yapılmıştır. Derin öğrenme yöntemleri kullanılarak yapılan deneylerde 32'lik bir F1 başarısı elde edilmiştir. Yazarlar ikinci deney olarak, Gillick [16] çalışmasında üretilen veri seti üzerine zayıf gözetim yöntemiyle alt kırılımlı varlıkları etiketlemişlerdir. Bu veri seti üzerinde yapılan deneyler ile 71,8'lik bir F1 başarısı elde edilmiştir.

Lee vd. [18] Çince'de elle etiketlenmiş bir korpus üretmişlerdir. Yazarlar, Choi vd. [17] çalışmasına benzer şekilde serbest formlu etiket tipi kullanmışlardır. Çalışmada 1,9 milyon sözcükten (token) oluşan ve 7100 farklı etiket tipi ile zayıf gözetim yöntemiyle etiketlenen büyük veri seti üretilmiştir. Ayrıca, çok sayıdaki alt kırılımlı varlık, 10 tane üst varlık ismi altında birleştirilmiştir. BERT tabanlı modellerle yapılan deneylerle üst kategorilerde 77,3'lük F1 başarısı ve alt kırılımlı VİT'de 24,9'luk bir F1 başarısı elde edilmiştir. Çalışma *kültürel miras* varlıklarının tahmininde başarısızlığa uğramış ve *yazar* gibi daha yaygın alt kırılımlı etikette ise 87'lik yüksek tahmin başarısı aldıklarını vurgulamışlardır.

Liqreina vd. [19] Arapça'da ilk geniş kapsamlı alt kategorili VİT çalışması yapmışlardır. Yazarlar veri setini daha önce yapılmış olan Arapça üst VİT veri seti çalışması [20] üzerine kurmuşlardır. Veri setindeki üst varlık etiketlerini 31 farklı alt kırılımlı varlıkla elle değiştirmişlerdir. BERT tabanlı modellerle yapılan deneylerle üst VİT için 92'lik, alt kırılımlı VİT için ise 88,5'lik bir F1 başarısı elde etmişlerdir.

### 1.1.3 Türkçede Alt Kırılımlı Varlık İsmi Tanıması

Alt kırılımlı varlıklar ile etiketlenmiş Türkçe veri setleri genelde genetik, finans, hukuk gibi alanlara özgü çalışmalarda oluşturulmuştur. Dinç [21], çalışmasında finansal haber metinlerinden oluşan bir veri kümesi üretmiştir. Veri kümesi 500 bin tane sözcükten oluşturulmuş olup, etiketleme yöntemine göre dört farklı şekilde sunulmuştur. Çalışma sonucunda LOTTERY, FIN-PRODUCT gibi yeni varlık isimleri eklenmiş olup, ORG varlık ismi FIN-ORG ve NON-FIN-ORG olarak ikiye ayrılarak ele alınmıştır. XLM, BERT tabanlı modellerle farklı deneyler yapılmıştır. mDeBERTaV3 modeli ile en yüksek başarıya erişilmiş olup, 90,65'lik bir F1 başarı puanı elde edilmiştir.

Çetindağ vd. [22] ise Türkçe hukuk metinleri için CRF ve BiLSTM temelli bir VİT mimarisi geliştirmişlerdir. Yazarlar üst varlıklara ek olarak alana özel Mevzuat, Mahkeme, Referans gibi altı kırılımlı varlıkları oluşturmuşlardır. GloVe ve Morph2Vec'in karakter düzeyinde özelliklerle birleştirildiği bir hibrit kelime temsilcisi ile 92,3'lük bir başarı yakalamışlardır.

Tanrısever vd. [23], Türk Savunma Sanayii Ürün Kataloğu'ndaki ürün açıklamalarını kullanarak bir veri seti oluşturmuşlardır. Yazarlar çalışmada ürün,

alakalı terim ve birim gibi özel varlıklar oluşturmuşlardır. Çalışma sonucunda CRF katmanlı bir BiLSTM modeliyle yapılan deneyler sonucunda 72,9'luk bir başarı alınmıştır.

Topcu ve El-Kahlout [24], Türkçe arama motorları için özel olarak seçilmiş varlıkları tanımaya odaklanan veri seti oluşturmuşlardır. Yazarların ürün_sanat_müzik olarak oluşturdukları varlık, film, şarkı, kitap gibi ürün alt kırılımlı varlıklarını kapsamaktadır. Çalışmadaki BERT ile yapılan eğitimler sonunda 90,41'lik bir başarı alınmıştır.

Kılıç [25] çalışmasında Türkçe spor haber metinlerinden yeni veri kümesi oluşturmuştur. Farklı BERT tabanlı modellerle eğitimler yapmış, modellerin yorumlanabilirliği üzerine de ek bir çalışma yapmıştır. Çalışmada XLM [26] ve BERT tabanlı modellerle beş farklı deney yapılmış ve 88,60'lık F1 başarısı elde edilmiştir.

## 1.2 Tezin Katkıları

Çalışmada öncelikli olarak Vikipedi metinlerinden üst ve seçilen alt kırılımlı varlıklar otomatik olarak etiketlenerek iki farklı veri seti oluşturulmuştur. Türkçe veri kümesi ile eğitilen BERT tabanlı modellere etiketli veri ile transfer öğrenme yapılarak yeni bir model geliştirilmiştir. Üst varlıklar üzerinden alınan sonuçlar farklı veri setleri üzerinden alınan sonuçlar ile karşılaştırılmış, alt kırılımlı varlıkların başarı sonuçları raporlanarak gelecek çalışmalar için önerilerde bulunulmuştur. Çalışmanın ilerleyen bölümlerinde sırası ile veri setinin oluşturulması ve istatistikleri hakkında bilgi verilmiş ve farklı modellerden alınan sonuçlar paylaşılmıştır.

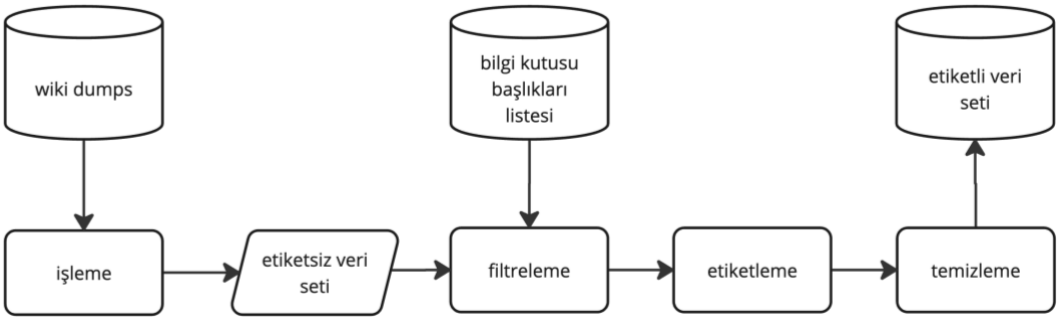
Türkçe varlık ismi tanımada alt kırılımlara sahip veri setinin eksikliğinden yola çıkılarak büyük ölçekli bir veri kümesinin oluşturulmasının bu alana önemli bir katkı sağlayacağına inanıyoruz. Veri kümesini oluştururken kişi, yer ve organizasyondan türetilen alt kırılımlı etiketlere odaklanılmış ve çalışmamızda 23 farklı alt kategori için etiketleme yapılmıştır. Etiketleme için yaygın olarak kullanılan BIO (Beginning, Inside, Outside) formatı seçilmiştir. Yer isminin bir alt kırılımı olan şehir için B-LOC-şehir, kişi isminin bir alt kırılımı olan sanatçı için B-PER-sanatçı olarak etiketleme yapılmıştır.

## 1.3 Tezin Ana Hatları

Tezin geri kalanı şu şekilde düzenlenmiştir: İkinci bölümde veri setini oluşturma, temizleme ve etiketleme aşamaları detaylı şekilde anlatmıştır. Çalışmada kullanılan dönüştürücü tabanlı derin öğrenme modelleri hakkında üçüncü bölümde detaylı olarak bilgi verilmiştir. Modellerin seçimine, ayarlara, modellerle yapılan deney ve sonuçlarına dördüncü bölümde yer verilmiştir.

# 2
# VERİ SETİ

Veri seti oluşturmak için Türkçe Vikipedi makaleleri kullanılmıştır. Veri seti oluşturmada ana unsur Vikipedi makalelerindeki bilgi kutusu olmuştur. Etiketleme işlemi Vikipedi makale formatından yararlanılarak zayıf gözetimli yöntem ile yapılmıştır. Şekil 2.1'de veri oluşturma ve etiketleme süreci blok diyagramı gösterilmiştir.




**Şekil 2.1** Süreç blok diyagramı

Veri seti oluşturma ve etiketleme işlemleri sonucunda 25 milyon üzeri sözcükten oluşan büyük çaplı bir veri seti oluşturulmuştur. Bu sözcükler BIO şemasındaki 23 farklı alt kırılımlı etiketlerle etiketlenmiştir. Alt kırılımlı varlıklar üst varlıklara geri dönüştürülerek aynı sözcüklerden dört üst varlıkla etiketlenmiş ikinci bir veri seti daha oluşturulmuştur.

## 2.1 Veri Seti Oluşturma

Metin kaynağı olarak Vikipedi'deki Türkçe makaleler kullanılmıştır. Metinler Şubat 2023 tarihinde indirilen Türkçe Vikipedi yedek paketinden elde edilmiştir. İndirilen dosya, üzerinde çalışılabilecek metin formatına dönüştürülmüştür. Metin formatı makale başlığı ve makale sütununa sahip .csv formatındadır.

Veri seti oluşturmadaki ana unsur Vikipedi bilgi kutusu özelliği olmuştur. Vikipedi'deki çoğu makalede bilgileri özetleyen, genellikle makalelerin sağ üst köşesine eklenen sabit formatlı bir bilgi kutusu bulunmaktadır. Her bilgi kutusu "şirket bilgi kutusu", "sporcu bilgi kutusu" gibi bir başlığa sahiptir. Bu başlıklar Vikipedi metinlerini filtrelemekte ve etiketleme işleminde önemli olmuştur. Metin

dosyasındaki makaleler bilgi kutusu, başlıklar, linkler ve diğer bilgilerin belli biçimlerde saklandığı "markdown" formatında tutulmaktadır. Şekil 2.2'de Sabiha Gökçen hakkındaki makaleye ait bilgi kutusunun web ve markdown hali gösterilmiştir.




```
{{Kişi bilgi kutusu
| ad                = Sabiha Gökçen
| resim             = Sabiha gokcen3.jpg
| resim_boyutu      =
| altyazı           =
| kendi_dilinde_adı =
| okunuş            =
| doğum_adı         =
| doğum_tarihi      = {{Doğum tarihi|1913|3|22}}
| doğum_yeri        = [[Bursa]], [[Osmanlı İmparatorluğu]]
| kaybolma_tarihi   =
| kaybolma_yeri     =
| kaybolma_durumu   =
| ölüm_tarihi       = {{ölüm tarihi ve yaşı|2001|3|22|1913|3|22}}
| ölüm_yeri         = [[Ankara]], [[Türkiye]]
| ölüm_sebebi       =
| bedeninin_bulunuşu =
| yattığı_yer       = [[Cebeci Askerî Şehitliği]], [[Ankara]]
| yattığı_yerin_koordinatları =
| anıtlar           =
| ikamet            =
| milliyet          =
| diğer_adları      =
| vatandaşlık       =
| eğitim            = [[Üsküdar Amerikan Lisesi]]
| alma_mater        =
| meslek            = Pilot ve yazar
| etkin_yılları     =
```

**Şekil 2.2** Vikipedi bilgi kutusu web ve markdown formatı

Vikipedi'de ayrıca tüm güncel bilgi kutusu başlıklarını listeleyen bir sayfa da vardır [27]. Bu sayfadan kişi, kuruluş, konum etiketleri altında kategorize edilebilen bilgi kutusu başlık listeleri çıkarılmıştır. Örneğin, "Filozof bilgi kutusu", "Kişi bilgi kutusu", "Kraliyet bilgi kutusu", "Makam sahibi bilgi kutusu", "Bilgisayar virüsü bilgi kutusu" listesinden "Bilgisayar virüsü bilgi kutusu" gibi kişi, organizasyon, konum başlıkları altında yer alamayan başlıklar filtrelenmiştir. İndirilen metin dosyasındaki makaleler, bu bilgi kutusu başlıklarından birini içerecek şekilde filtrelenmiştir. Elde edilen makaleler taranıp, bilgi kutusu başlığı ve varsa meslek bilgisi veri setinde ek sütunlar altında kaydedilmiştir. Etiketsiz veri seti formatı ve örnekler Tablo 2.1'de verilmiştir.

Makalelerdeki HTML etiketleri, yorumlar, bölüm başlıkları, dosya bilgileri, referans ve işaretleme parçaları gibi metin parçası olmayan kısımlar da temizlenmiştir. Bilgi kutusu bilgileri de makalelerden silinmiştir.

**Tablo 2.1** Etiketsiz veri seti formatı ve örnekler

| metinBaşlık | bilgiKutusuBaşlık | meslek | metin |
|---|---|---|---|
| Sabiha Gökçen | Kişi bilgi kutusu | ['pilot', 'yazar'] | … |
| Marie Curie | Bilim insanı bilgi kutusu | ['bilim insanı'] | … |
| Yıldız Teknik Üniversitesi | Üniversite bilgi kutusu | - | … |
| Ankara | Şehir bilgi kutusu | - | … |

### 2.1.1 Etiketleme İşlemi

Etiketleme işleminde metinlerin bilgi kutusu ve markdown özelliklerinden yararlanılmıştır. Öncelikle makaledeki başlıkla eşleşen sözcükler etiketleme işleminden geçmiştir. Kişi makalesinde sözcükler meslek bilgisi ile etiketlenmiştir. Bazı makalelerde meslek kısmında birden fazla mesleği içeren meslek dizisi bulunmaktadır. Bu durumda sözcükler, dizideki ilk meslek kategorisi ile etiketlenmiştir. Organizasyon ve yerleşim makalelerindeki sözcükler bilgi kutusu başlığındaki kategori ile etiketlenmiştir. Örneğin, Tablo 2.1'deki ilk örnekte makalede geçen Sabiha Gökçen sözcüğü *pilot* olarak, Yıldız Teknik Üniversitesi *üniversite* olarak etiketlenmiştir.

Bir sonraki aşamada makaledeki bağlantılı elemanlar ele alınmıştır. Çoğu makaleden diğer makalelere bağlantılar bulunmakta ve bu bağlantılar makalenin markdown formatında çift köşeli parantez içinde gösterilmekteydi. Başka bir makaleye bağlantılı linkler, o makalenin bilgi kutusu kategorisiyle etiketlenmiştir. Vikipedi makalesinden alınan *Sabiha Gökçen (22 Mart 1913, [[Bursa]] - 22 Mart 2001, [[Ankara]]), Türk pilot* cümlesinde Ankara ve Bursa linkli sözcüklerdir. Cümledeki *Ankara* sözcüğü, Tablo 2.1'deki format altında olan *Ankara* başlıklı makaleden alınan *şehir* kategorisi ile etiketlenmiştir.

Etiket seçiminde veri setinden elde edilen bilgi kutusu başlığı ve meslek bilgileri kullanılmıştır. Veri setinde tekrar etmeyen meslek tipleri ve farklı meslek bilgi kutusu başlıkları birleştirilip, daha geniş kategori altında gruplandırılmıştır. Bu kategorileri seçerken meslek gruplarının genelliğine ve çok özel olmamasına dikkat edilmiştir. Sporcu, basketbolcu, voleybolcu gibi spor alanlarına bağlı varlık isimleri *sporcu* altında gruplanmıştır. *Müzakere uzmanı* gibi başka gruplara dahil edilemeyen varlık isimleri ise *diğer* olarak değiştirilmiştir. Organizasyon ve yerleşim alt kategorileri ise bilgi kutusu başlıklarını gruplandırarak oluşturulmuştur. Üniversite, okul bilgi kutusu gibi başlıklara sahip olanlar *eğitim kurumu* altında birleştirilmiştir. Bu şekilde toplam 23 farklı alt kırılımlı varlık ismi elde edilmiştir.

Gruplanmış kategoriler BIO şema formatındaki uygun etiketlerle eşleştirilmiştir. Türkçe'de yapılan alt kırılımlı kategoriler, ait olduğu üst varlığın BIO şema sonuna eklenmiştir. Örneğin, "spor" etiketi B-ORG-spor olarak değiştirilmiştir. Aynı etiket altında birden fazla kelime varsa, ilki B-ORG-spor olarak ve sonrakiler I-ORG-spor olarak etiketlenmiştir. *Beşiktaş Jimnastik Kulübü* B-ORG-spor, I-ORG-spor, I-ORG-spor olarak etiketlenmiştir.

Alt kırılımlı varlık isimleri DATE, LOC, ORG ve PER üst varlıkları altında gruplanarak üst varlıklarla etiketlenmiş ikinci bir veri seti oluşturulmuştur. Bu veri setinde, yukarıda örnek olarak verilen *Beşiktaş Jimnastik Kulübü* sözcükleri B-ORG, I-ORG, I-ORG üst varlık ismi ile etiketlenmiştir. Veri setinde tekrar eden cümleler tespit edilip, kaldırılmıştır. İki veya daha az sözcüklerden oluşan kısa cümleler veri setinden çıkarılmıştır.

## 2.2 Varlıkların Açıklaması

Bu bölümde oluşturulan alt kırılımlı varlık isimlerinin açıklaması yapılmıştır. Alt kırılımlı varlık isimleri ile etiketlenen sözcüklerin bulunduğu cümle örneği ile gösterilmiştir.

### 2.2.1 Alt Kırılımlı Varlıklar

Alt kırılımlı kategoriler üst kategorileri farklı alanlara ayırarak oluşturulmuştur. Tablo 2.2'de LOC alt kırılımlı varlık isimleri açıklamaları ve örnekleri ile verilmiştir. LOC varlık ismi sekiz tane alt kategoriye ayrılmıştır.

**Tablo 2.2** LOC alt kategorileri varlıkların açıklaması

| **Varlık** | **Açıklama** | **Örnek** |
|---|---|---|
| LOC-diğer | Cadde, sokak, meydan | Şehir merkezine çok yakın **Kızılay Meydanı'a** 6 km bir konuma sahiptir. |
| LOC-doğaAlanı | Park, orman, nehir, ova | **Aras Nehri'nin** rejimi oldukça düzensizdir. |
| LOC-il | Eyalet ve iller | **Gümüşhane** il merkezine 69 km uzaklıktadır. |
| LOC-ilçe | İlin parçası olan yerleşimler | Özgür, Rize ilinin **Ardeşen** ilçesine bağlı bir köydür. |
| LOC-kıta | Avrupa, Güney Amerika gibi kıtalar | Hanya'nın Osmanlı İmparatorluğu tarafından fethi, **Avrupa'da** büyük etki uyandırdı. |
| LOC-yerleşim | Semt, komün yerleşimler, mahalle, bölge | İlk öğrenimine **Vezneciler'de** Sems'ül maârif mektebinde başladı. |
| LOC-ülke | Türkiye, Yeni Zelanda gibi ülkeler; imparatorluk, hanlıklar | Aile köken olarak **Karabağ Hanlığı** soyundan gelmektedir. |
| LOC-şehir | Büyük ve küçük şehirler | Irak'ın **Musul** şehrinden Batman'ın Gercüş ilçesine göç eden bir aileden gelmektedir. |

Tablo 2.3'te ORG alt kırılımlı varlık isimlerinin açıklaması ve cümlelerdeki örnekleri verilmiştir. ORG üst varlık ismi beş tane alt kırılımlı varlık ismine ayrılmıştır.

**Tablo 2.3** ORG alt kategorileri varlıkların açıklaması

| **Varlık** | **Açıklama** | **Örnek** |
|---|---|---|
| ORG-askeriKuruluş | Askeri kuruluşlar, örneğin Türk Silahlı Kuvvetleri | Volk 1956'da **Sovyet Hava Kuvvetleri'nde** pilot oldu. |
| ORG-diğer | Diğer kategorilere uymayan farklı kuruluşlar, örneğin Edinburgh festivali, Oscar ödülleri | 1988 yılında Hans Werner **Henze Festivalinde** En İyi Yorumcu ödülünü kazandı. |
| ORG-eğitimKurumu | Üniversite, Lise, Okul | **Lefkoşa Üniversitesi**, yine belediye sınırları içerisinde yer almaktadır. |
| ORG-kamusalYapı | Kütüphane, kule, hastane ve diğer yapılar | 1948-1993 yılları arasında **Beyoğlu Belediye Hastanesi** olarak hizmet verdi. |
| ORG-kuruluş | Siyasi parti, organizasyon, kuruluş | **Uluslararası Af Örgütü** tarafından vicdan mahkumu olarak kabul edildi. |
| ORG-spor | Spor kulüpleri ve organizasyonları | **UEFA**, **FIFA'ya** bağlı altı konfederasyon içerisindeki en büyük konfederasyondur. |

Tablo 2.4'te PER alt kırılımlı kategorilerin açıklaması ve cümlelerdeki örnekleri verilmiştir. PER kategorisi yedi alt kategoriye ayrılmıştır.

**Tablo 2.4** PER alt kategorileri varlıkların açıklaması

| Varlık | Açıklama | Örnek |
|---|---|---|
| PER-akademisyen | Profesör gibi akademide çalışan meslek grupları, bilim insanları | **Tzvetan Todorov** (1 Mart 1939, Sofya - 7 Şubat 2017), Fransız-Bulgar filozof, tarihçi. |
| PER-askeriPersonel | Komutan, asker, subaylar | 1520 yılında **Barbaros Hayreddin Paşa** donanmasına katılır. |
| PER-diğer | Diğer kategorilere uymayan farklı mesleklere sahip kişiler | **Turgut Toydemir** (d. 1938, İstanbul), Türk mimardır. |
| PER-kraliyetMensubu | Kraliyet mensubuna ait kişiler, örneğin kral, kraliçe, prenses | İmparatorluk, zirvesini 16. yüzyılda, özelikle **I. Süleyman** döneminde yaşadı. |
| PER-politikacı | Politikacılar, devlet memurları, liderler | **Mustafa Kemal Atatürk** Meclis-i Mebusan'ın İstanbul'da değil Anadolu'da toplanmasını istemekteydi. |
| PER-sanatçı | Müzisyen, oyuncu, ressam gibi sanatçılar | **Yakub Kolas** onuruna Minsk'in merkezindeki bir sokağa adı verilmiştir. |
| PER-sporcu | Sporcular, spor müsabakalarına katılmış olan kişiler | **Yalçın Ayhan** (d. 1 Mayıs 1982, İstanbul) Türk eski futbolcudur. |

Üst varlık isimlerinden biri olan DATE, alt kırılımlı kategorilere ayrılmamıştır. DATE kategorisi yıl, ay veya yüzyıl gibi tarih varlık isimlerini etiketlemede kullanılmıştır. *14 Eylül 1946 yılında doğmuştur* cümlesinde *14 Eylül 1946* sözcükleri DATE kategorisi ile etiketlenmiştir.

LOC, ORG ve PER üst varlıkları altında *diğer* kategorisi oluşturulmuştur. Başka kategori altında gruplanamayan varlık isimleri *diğer* kategorisi ile etiketlenmiştir. Köy gibi daha az bilinen yerleşimler ile ilgili makalelerdeki eksik bilgilerden dolayı, bazı küçük yerleşimler LOC-diğer olarak hatalı etiketlenmiştir.

## 2.3 Veri Seti İstatistikleri

Veri seti 25.000.000 sözcükten oluşmaktadır. Toplamda 1.456.561 farklı cümle vardır. Bu cümlelerden 1.052.271 tanesinde, yani %72,24'ünde en az bir tane etiketli varlık ismi mevcuttur.

Tablo 2.5'de görüldüğü gibi, veri setindeki tüm varlıkların %14,47'si üst etiketler ile etiketlenmiş olup, bunlardan sırasıyla en çok DATE, LOC, ORG, PER kategorisi ile etiketlenme gerçekleşmiştir.

**Tablo 2.5** Üst varlık ile etiketli veri istatistikleri

| **Varlık** | **Sayısı** | **Oranı %** |
|---|---|---|
| DATE | 1.453.327 | 5,40 |
| LOC | 866.827 | 3,40 |
| ORG | 658.314 | 2,58 |
| PER | 711.151 | 2,79 |
| **Toplam** | **3.689.619** | **14,47** |

Tablo 2.5'de DATE üst varlık ismi ve LOC alt kırılımlı varlık isimlerin toplam veriye oranı gösterilmiştir. Tablo 2.6'de ORG ve Tablo 2.7'de ise PER alt kırılımlı varlık isimlerin toplam veriye oranı gösterilmiştir. ORG-kuruluş, LOC-yerleşim

gibi geniş kapsamlı alt kırılımlı varlık isimlerin sayısı, ORG-askeriKuruluş gibi daha özel etiketlerin sayısına göre oldukça yüksektir. Alt kırılımı olmayan DATE varlık ismi %5,70 oranı ile en çok görünen varlık ismi olmuştur. Veri setinin %0,02'si LOC-doğaAlanı varlık ismi ile etiketlenmiştir, bu da en düşük orana sahip alt kırılımlı varlık ismi olmuştur.

**Tablo 2.6** DATE ve LOC alt kırılımlı varlık ile etiketli veri istatistikleri

| **Varlık** | **Sayısı** | **Oranı %** |
|---|---|---|
| DATE | 1.453.327 | 5,70 |
| LOC-diğer | 39.099 | 0,15 |
| LOC-doğaAlanı | 6224 | 0,02 |
| LOC-il | 62.880 | 0,25 |
| LOC-ilçe | 65.648 | 0,26 |
| LOC-kıta | 3053 | 0,01 |
| LOC-yerleşim | 250.889 | 0,98 |
| LOC-ülke | 323.574 | 1,27 |
| LOC-şehir | 115.460 | 0,45 |

Veri setindeki sözcükler, ORG üst varlık ismi altında en yüksek oranla ORG-spor, en düşük ORG-askeriKuruluş alt kırılımlı varlık ismi ile etiketlenmiştir. PER üst varlık ismi altında en yüksek oranla PER-sanatçı, en düşük PER-kraliyetMensubu alt kırılımlı varlık ismi ile etiketlenmiştir. LOC üst varlık ismi altında en yüksek oranla LOC-ülke alt kırılımlı varlık ismi ile etiketlenmiştir.

**Tablo 2.7** ORG alt kırılımlı arlık ile etiketli veri istatistikleri

| **Varlık** | **Sayısı** | **Oranı %** |
|---|---|---|
| ORG-askeriKuruluş | 12.512 | 0,05 |
| ORG-diğer | 17.138 | 0,07 |
| ORG-eğitimKurumu | 178.789 | 0,70 |
| ORG-kamusalYapı | 49.693 | 0,19 |
| ORG-kuruluş | 186.384 | 0,73 |
| ORG-spor | 213.798 | 0,84 |

**Tablo 2.8** PER alt kırılımlı varlık ile etiketli veri istatistikleri

| | | |
|---|---|---|
| PER-akademisyen | 65.096 | 0,25 |
| PER-askeriPersonel | 28.026 | 0,11 |
| PER-diğer | 155.932 | 0,61 |
| PER-kraliyetMensubu | 19.858 | 0,08 |
| PER-politikacı | 109.913 | 0,43 |
| PER-sanatçı | 243.323 | 0,95 |
| PER-sporcu | 89.003 | 0,35 |
| **Toplam** | **3.689.619** | **14,47** |

# 3
# DERİN ÖĞRENME MODELLERİ

Vaswani vd. [28] 2017 tarihli çalışmalarında yapay sinir ağları mimarisine dikkat (attention) mekanizması ekleyerek bir Dönüştürücü (Transformer) mimarisi oluşturmuşlardır. Günümüzdeki çoğu derin öğrenme modeli dönüştürücü mimarisi üzerinde geliştirilmiş olup, daha büyük boyuttaki derleme ve daha çok dil ile farklı DDİ işlemlerinde başarı göstermişlerdir.

Dönüştürücü mimarisi, kelime dizisinin bağlamını öğrenen ve bundan yeni veriler üreten bir sinir ağıdır. Tekrarlı Sinir Ağlarındaki (RNN) kodlayıcı-kod çözücü (encoder-decoder) mimarisinden esinlenmiştir. Fakat tekrarlama mantığı yerine dikkat mekanizmasını kullanarak, birbirinden uzak olan kelimeler arasındaki ilişkiyi iyi yakalamıştır.

Dönüştürücü mimarisi, BERT (Bidirectional Encoder Representations from Transformers) veya GPT (Generative Pre-trained Transformers) gibi önceden eğitilmiş modellerin gelişmesine yol açmıştır. BERT, dönüştürücü mimarisini iki taraflı kodlayıcı kullanarak eğiten model olup, kelimeyi hem öndeki hem arkadaki bağlamdan tahmin etmektedir. BERT tabanlı birçok açık kaynaklı model geliştirilmiş olup, duygu analizi, VİT gibi alanlarda yaygın olarak kullanılmaktadır.

BERT modelinin VİT üzerindeki yüksek başarısı sonucunda farklı dillerdeki veri setleri üzerinde eğitilen BERT tabanlı modeller geliştirilmiştir. BERTurk [29], Türkçe veri seti üzerinde eğitilen açık kaynaklı modeller kümesidir. BERTurk-128 [30], DistilBERTurk [31] ve ConvBERTurk [13] modelleri buna dahildir.

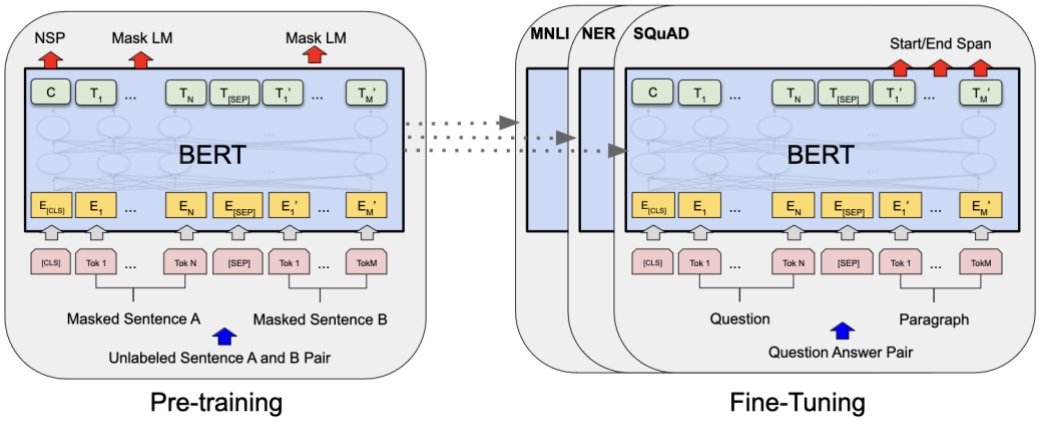
## 3.1 BERT

Devlin vd. dönüştürücü tabanlı BERT [11] modelini geliştirmişlerdir. BERT, bağlamı her iki taraftan ele alarak öğrenme üzerine yapılmıştır. Model eğitiminde Vikipedi metinleri dahil olan büyük korpus kullanılmıştır. BERT girdi için kelimeler yerine Wordpiece sözcüklerini (token) kullanmaktadır.

Genel kapsamda BERT, ön eğitim (pre-training) ve ince ayar (fine-tuning) sürecinden oluşmaktadır (Şekil 3.1). Her iki süreç benzer mimari yapı kullanmakta

olup, ince ayar yapma süreci BERT'in daha küçük veri setiyle soru cevaplama, VİT, duygu analizi gibi özel bir işlemde eğitim yapılması için ayrılmıştır.

BERT mimarisi maskeleme (Masked Language Modeling – MLM) ve bir sonraki cümleyi tahmin etme (Next Sentence Prediction – NSP) yöntemlerini kullanarak oluşturulmuştur. MLM ile bazı kelimeler maskelenir ve bu kelimelerin çevresindeki bağlamdan ne olabileceği tahmin edilir. NSP ile ise bir cümlenin bir önceki cümlenin devamı olup olmadığını bulmaya çalışır. BERT, bu iki yöntemi birlikte kullanarak kelimeler ve cümleler arasında ilişkiyi iyi kurmaya çalışır [32].




**Şekil 3.1** BERT modeli ön eğitim ve ince ayar yapma süreci

BERTurk-128 [30] veya BERTurk modeli 35 GB Türkçe veri seti üzerinde eğitilmiş BERT tabanlı ön eğitimli bir modeldir. BERTurk eğitim veri seti Vikipedi paketi (Wikipedia dumps), Kemal Oflazer [33] tarafından sağlanmış veri seti, OPUS [34] ve OSCAR [35] korpusundan oluşmaktadır. BERT'in Wordpiece sözcüklere (token) ayırma algoritmasını oluşturduğu veri seti üzerinde eğiterek kullanmışlardır. 128 bin sözlük boyutuna sahip olduğundan BERTurk-128 olarak isimlendirilmiştir.

## 3.2 DistilBERT

DistilBERT, BERT'in daha hafif ve daha hızlı şekilde çalışmasını sağlayan modeldir. Ön eğitim aşamasında distilasyon tekniklerini kullanarak, BERT'in dil anlama kapasitesinin %90'ını elde tutarak, BERT'ten %60 daha hızlı olmuş ve %40 daha az bellek ihtiyacı gerektirmiştir.

DistilBERT-base Turkish Uncased [30] veya DistilBERTurk modeli BERTurk'ü ön eğitim modeli olarak kullanan ve daha küçük olan 7GB veri seti veri üzerinde eğitilmiş modeldir. Tez çalışmasında bu model üst ve alt kırılımlı VİT veri seti üzerinde ince ayarlar yapılarak kullanılmıştır.

## 3.3 ConvBERT

ConvBERT [36] BERT modelinin dikkat mekanizması üzerine konvolüsyon kullanılarak değiştirilen halidir. ConvBERT ayrıca BERT'in eğitim süresini azaltmaktadır.

ConvBERTurk [13] modeli çok dilli C4 (mC4) [37] derlem üzerinde eğitilmiştir. Eğitim veri seti olan mC4, 31 milyon üzeri sözcükten oluşmakta olup, boyutu 242GB dır. Bu çalışmada ConvBERTurk modeli hiperparametreleri değiştirilerek kullanılmıştır.

## 3.4 XLM-RoBERTa

Diller arası dil modeli [38] (Cross-lingual language model – XLM), diller arası bir hedef olmadan (unsupervised) eğitilen modeldir. RoBERTa [39] büyük İngilizce derlemi üzerinde kendi kendine denetim (self-supervised) yöntemiyle ön eğitimli modeldir.

XLM'in çok dille çalışma özelliği ile RoBERTA'nın ön eğitim yaklaşımını kullanarak XLM-RoBERTa (XLM-R) modeli geliştirilmiştir [26]. Türkçe dahil yüz dil üzerinde ön eğitilmiştir. Modelin özelliklerinden biri ön eğitim aşamasında diller arası bilgi paylaşımıdır. Bu sayede XLM-R ile az kaynaklı diller üzerinde çok dilli BERT (mBERT) modellerine göre daha yüksek başarı elde edilmiştir.

Tez çalışmasında çok dilli XLM-RoBERTa [40] modeli ile üst ve alt kırılımlı VİT veri seti üzerinde ince ayarlar yapılarak kullanılmıştır.

# 4
# DENEYSEL SONUÇLAR

Her model ile gerçekleştirilen testler, öncelikle üst VİT üzerine sonra alt kırılımlı VİT üzerine yapılmıştır. BERT tabanlı üç farklı ve XLM-R modeliyle deneyler gerçekleştirilmiştir. Her model deneyi sonucunda değerlendirmeler her iki veri seti için yapılmıştır. Veri setlerinin %80'i eğitim, %20'si test verisi olarak ayrılmıştır.

## 4.1 Değerlendirme Metrikleri

Değerlendirme metrikleri olarak kesinlik (precision), duyarlılık (recall) ve F1 puanı kullanılmıştır. Model başarısını değerlendirirken, F1 puanı ana başarı göstergesi olarak seçilmiştir.

Kesinlik pozitif olarak tahmin edilen değerlerin gerçekten kaç adedinin pozitif olduğunu göstermektedir. Başka bir deyişle kesinlik, pozitif tahminin doğru olup olmadığına odaklanır (Eşitlik 4.1).

$$Kesinlik = \frac{\text{Doğru Pozitif}}{\text{Doğru Pozitif} + \text{Yanlış Pozitif}} \tag{4.1}$$

Duyarlılık ise pozitif olarak tahmin etmemiz gereken işlemlerin ne kadarını pozitif olarak tahmin ettiğimizi gösteren bir metriktir. Gerçek pozitif tahminlerin sayısının, gerçekte pozitif olan toplam örnek sayısına bölünmesiyle hesaplanır (Eşitlik 4.2).

$$Duyarlılık = \frac{\text{Doğru Pozitif}}{\text{Doğru Pozitif} + \text{Yanlış Negatif}} \tag{4.2}$$

F1 skor, Kesinlik ve Duyarlılık değerlerinin harmonik ortalamasıdır. F1 puanı, eşit dağılmayan veri kümelerinde veya sadece Yanlış Negatif ve Yanlış Pozitif değerlerinin yetersiz kalıp, tüm hata maliyetlerini ölçmek gerektiğinde faydalı görülmüştür (Eşitlik 4.3).

$$F1 = 2 * \frac{\text{Kesinlik} * \text{Duyarlılık}}{\text{Kesinlik} + \text{Duyarlılık}} \tag{4.3}$$

## 4.2 Kullanılan Kütüphaneler

Deneyler, Google Colab platformu üzerinde gerçekleştirilmiştir. Programlama dili olarak Python kullanılmış olup, çalışmada 3.11.5 versiyonu ile çalışılmıştır. Deneyler, HuggingFace tarafından geliştirilen transformers kütüphanesinden alınmış BERTurk tabanlı ve XLM-R modelleriyle gerçekleştirilmiştir. Pytorch kütüphanesinden veri seti gömmesi ve optimizasyon işlemleri için yararlanılmıştır.

## 4.3 BERTurk İle Yapılan Deneyler

BERTurk-128 [30] veya BERTurk modeli BERT tabanlı, 32 GB veri seti üzerinde ön eğitimli modeldir. Çalışmada oluşturulan veri seti üzerinde BERTurk ile transfer eğitim yapılmıştır. Bunun için farklı hiperparametreler ile ince ayarlar yapılmıştır. Eğitim veri setininin tamamını geçme adımı olan epoch sayısı 5 olarak seçilmiştir. Batch sayısı, eğitim sırasında kaç tane veri seti örneğinin paralel şekilde çalıştırıldığı sayısını veren sayıdır. Batch, BERTurk ile olan deneylerde eğitim veri seti için 32 ve test veri seti için 16 olarak seçilmiştir. Maximum Sequence Length veya en büyük dizi uzunluğu, eğitimde her bir girdi dizisine en fazla kaç tane token girebileceğini sınırlandırır. Çalışmada bu sayı BERTurk modeli ön eğitimdeki gibi 128 olarak seçilmiştir.

### 4.3.1 Üst Varlık

Deneyler önce üst varlıklarla etiketlenen veri seti üzerinde yapılmıştır. Başarı sonuçları Tablo 4.1'de gösterilmiştir. BERTurk ile alınan ortalama F1 başarı puanı 90,23 olmuştur. En yüksek başarı 98,76 ile DATE ile alınırken en düşük başarı 87,05 ile ORG için elde edilmiştir.

### 4.3.2 Alt Kırılımlı Varlık

BERTurk ile yapılan alt kırılımlı VİT deneyinin başarı sonuçları Tablo 4.2'de gösterilmiştir. Ortalama 76,85 kesinlik, 82,87 duyarlılık ve 79,53 F1 başarı puanı elde edilmiştir. En yüksek puan DATE için 97,02 olurken, en düşük başarı LOC-diğer için 62,89 ile alınmıştır. LOC kategorisi altında en yüksek başarı 95,12 puan ile LOC-ülke için alınmıştır. PER altında PER-diğer en düşük 62,94, PER-sporcu ise en yüksek 89,63 puanını almıştır. ORG varlığı içinde en düşük 74,31 puanıyla ORG-askeriKuruluş ve ORG-eğitimKurumu, en başarılı 88,87 puanıyla ORG-spor olmuştur.

**Tablo 4.1** BERTurk ile üst varlık tahmin sonuçları

| Varlık | Kesinlik | Duyarlılık | F1 |
|---|---|---|---|
| DATE | 96,14 | 97,96 | 98,76 |
| LOC | 90,39 | 94,75 | 90,11 |
| ORG | 79,56 | 86,02 | 87,05 |
| PER | 82,52 | 89,42 | 87,54 |
| **Toplam** | **87,15** | **92,04** | **90,23** |

**Tablo 4.2** BERTurk ile DATE ve LOC alt kırılımlı varlık tahmin sonuçları

| Varlık | Kesinlik | Duyarlılık | F1 |
|---|---|---|---|
| DATE | 95,57 | 98,52 | 97,02 |
| LOC-diğer | 68,80 | 57,91 | 62,89 |
| LOC-doğaAlanı | 65,32 | 70,81 | 67,99 |
| LOC-il | 92,19 | 90,53 | 91,36 |
| LOC-ilçe | 85,09 | 91,79 | 88,31 |
| LOC-kıta | 55,59 | 83,07 | 66,61 |
| LOC-yerleşim | 86,36 | 89,91 | 88,10 |
| LOC-ülke | 95,13 | 95,11 | 95,12 |
| LOC-şehir | 85,94 | 91,59 | 88,64 |

**Tablo 4.3** BERTurk ile ORG alt kırılımlı varlık tahmin sonuçları

| Varlık | Kesinlik | Duyarlılık | F1 |
|---|---|---|---|
| ORG-askeriKuruluş | 72,92 | 75,76 | 74,31 |
| ORG-diğer | 82,69 | 89,90 | 86,14 |
| ORG-eğitimKurumu | 76,52 | 83,81 | 74,31 |
| ORG-kamusalYapı | 70,49 | 78,57 | 86,14 |
| ORG-kuruluş | 77,91 | 84,08 | 80,00 |
| ORG-spor | 85,04 | 93,07 | 88,87 |

**Tablo 4.4** BERTurk ile PER alt kırılımlı varlık tahmin sonuçları

| Varlık | Kesinlik | Duyarlılık | F1 |
|---|---|---|---|
| PER-akademisyen | 64,50 | 80,16 | 71,49 |
| PER-askeriPersonel | 65,83 | 71,62 | 68,61 |
| PER-diğer | 67,02 | 59,33 | 62,94 |
| PER-kraliyetMensubu | 61,51 | 78,78 | 69,08 |
| PER-politikacı | 73,02 | 79,89 | 76,30 |
| PER-sanatçı | 76,88 | 85,70 | 81,05 |
| PER-sporcu | 86,22 | 93,32 | 89,63 |

Her üst varlık altındaki alt kırılımlı varlıkların tahmin sonucundaki hata matrisi ayrıca oluşturulmuştur. Şekil 4.1'de LOC alt kırılım varlıkları hata matrisi

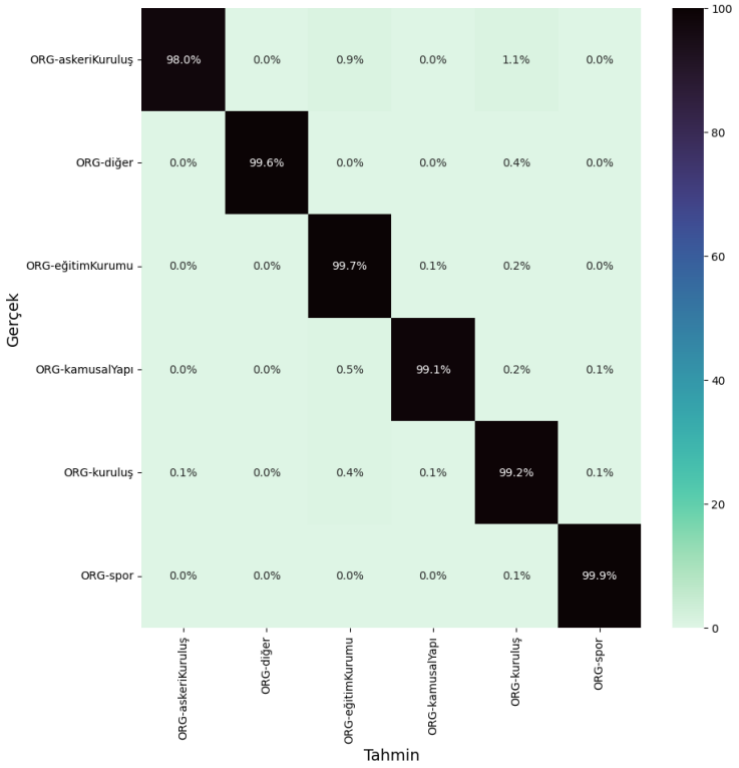
verilmiştir. LOC-diğer varlığı en düşük doğruluğa sahip olup, %20'ye yakın kısmı LOC-yerleşim olarak tahmin edilmiştir. Örneğin, *Badempınarı, Elazığ ilinin Elazığ ilçesine bağlı bir köydür* cümlesinde Badempınarı sözcüğü B-LOC-yerleşim olarak etiketlenmesi gerekirken, B-LOC-diğer olarak etiketlenmiştir. Veri setinde birçok benzer cümlelerde köy kategorisi B-LOC-yerleşim olarak etiketlendiğinden, BERTurk eğitimde bu sözcüğü B-LOC-yerleşim olarak öğrenmiştir.




**Şekil 4.1** BERTurk ile LOC alt kırılımlı VİT hata matrisi

Şekil 4.2'de ORG alt kırılım varlıkları hata matrisinde verilmiştir. ORG alt kırılımlı varlık isimleri arasındaki tahminlerde LOC (Şekil 4.1) ve PER (Şekil 4.3) alt kırılımlı varlık isimleri arasındaki tahminlere göre daha az hatalar görülmüştür.

BERTurk ile yapılan deney sonucu alınan PER alt kırılımlı VİT hata matrisi Şekil 4.3'te gösterilmiştir. LOC alt kırılımlı varlık tahminindeki gibi burada da en düşük oranla doğru tahmin edilen PER-diğer varlığı olmuştur. PER-diğer, hatalı olarak en çok PER-sanatçı (%15,5) ve sonra PER-akademisyen (%5,5) olarak tahmin edilmiştir. Örneğin, *Köncebeyim Nahçıvanski (1841 - ?), Azerbaycanlı şaireydi* cümlesindeki *Köncebeyim Nahçıvanski* varlığı veri setinde PER-diğer olarak etiketlenmiş olup, BERTurk modeli ile PER-sanatçı olarak etiketlenmiştir.




**Şekil 4.2** BERTurk ile ORG alt kırılımlı VİT hata matrisi




**Şekil 4.3** BERTurk ile PER alt kırılımlı VİT hata matrisi

## 4.4 DistilBERT İle Yapılan Deneyler

DistilBERTurk [31] modeli BERTurk'ün daha az parametre ile eğitilmiş versiyonudur. BERTurk'te kullanılan eğitim veri setinin 7 GB'lık küçük hali ile ön eğitim yapılmıştır. Çalışmada DistilBERTurk ile farklı hiperparametler seçilerek deney yapılmıştır. Epoch sayısı 5, eğitim batch sayısı 32 olarak, test batch sayısı ise 16 olarak seçilmiştir.

DistilBERTurk, BERTurk'un daha küçük versiyonu olmasından dolayı, BERTurk'e göre daha düşük başarı puanları alınması beklenmektedir. DistilBERTurk'ün transfer öğrenme süreci diğer BERTurk modellerine göre daha kısa olmuştur [29].

### 4.4.1 Üst Varlık

DistilBERTurk ile üst varlık VİT deneyi sonucundaki alınan başarı sonuçları Tablo 4.3'de gösterilmiştir. Ortalama F1 başarı puanı 87,93 olmuştur. En yüksek başarı 96,25 ile DATE için, en düşük başarı da 76,38 ile ORG için elde edilmiştir.

**Tablo 4.5** DistilBERTurk ile üst varlık tahmin sonuçları

| Varlık | Kesinlik | Duyarlılık | F1 |
|---|---|---|---|
| DATE | 95,04 | 97,48 | 96,25 |
| LOC | 92,56 | 94,05 | 93,33 |
| ORG | 74,87 | 77,95 | 76,38 |
| PER | 79,31 | 84,41 | 81,78 |
| **Toplam** | **86,45** | **88,47** | **87,93** |

### 4.4.2 Alt Kırılımlı Varlık

DistilBERTurk ile yapılan alt kırılımlı VİT deneyleri sonucunda ortalama 73,60 Kesinlik, 74,90 duyarlılık ve 73,66 F1 başarısı elde edilmiştir. Tablo 4.4'de

DistilBERTurk ile DATE ve LOC alt kırılımlı varlık isimleri üzerinde alınan başarı sonuçları gösterilmiştir. DistilBERTurk ile yapılan deney sonucundaki en düşük ve en yüksek başarı puanlı alt kırılımlı varlıklar, BERTurk sonucu gibi olmuştur.

**Tablo 4.6** DistilBERTurk ile DATE ve LOC alt kırılımlı varlık tahmin sonuçları

| **Varlık** | **Kesinlik** | **Duyarlılık** | **F1** |
|---|---|---|---|
| DATE | 95,27 | 97,06 | 96,15 |
| LOC-diğer | 68,95 | 51,37 | 48,55 |
| LOC-doğaAlanı | 62,80 | 39,57 | 48,55 |
| LOC-il | 92,61 | 89,95 | 91,27 |
| LOC-ilçe | 83,16 | 91,65 | 87,16 |
| LOC-kıta | 78,90 | 53,11 | 63,49 |
| LOC-yerleşim | 85,45 | 90,05 | 87,69 |
| LOC-ülke | 92,20 | 95,37 | 93,76 |
| LOC-şehir | 84,00 | 91,92 | 87,78 |

En yüksek başarı 96,15 ile DATE için alınırken en düşük başarı 48,55 ile LOC-diğer ile elde edilmiştir. LOC kategorisi altında en yüksek başarı 93,76 ile LOC-ülke için alınmıştır. PER altında, 55,09 ile PER-diğer en düşük başarıyı göstermiş iken, PER-sporcu 85,23 ile en yüksek başarıyı almıştır. ORG varlığı için de en düşük başarı 63,30 ile ORG-askeriKuruluş ve ORG-eğitimKurumu, en yüksek başarı 84,45 ile ORG-spor'dan alınmıştır. Tablo 4.4'de DistilBERTurk ile yapılan deney sonucu ORG, Tablo 4.5'te ise PER alt kırılımlı varlık isimleri üzerinde alınan başarı sonuçları gösterilmiştir.

**Tablo 4.7** DistilBERTurk ile ORG alt kırılımlı varlık tahmin sonuçları

| Varlık | Kesinlik | Duyarlılık | F1 |
|---|---|---|---|
| ORG-askeriKuruluş | 60,49 | 66,38 | 63,30 |
| ORG-diğer | 59,23 | 80,19 | 68,14 |
| ORG-eğitimKurumu | 63,12 | 75,98 | 68,95 |
| ORG-kamusalYapı | 68,21 | 66,80 | 67,50 |
| ORG-kuruluş | 73,95 | 76,22 | 75,06 |
| ORG-spor | 81,04 | 88,17 | 84,45 |

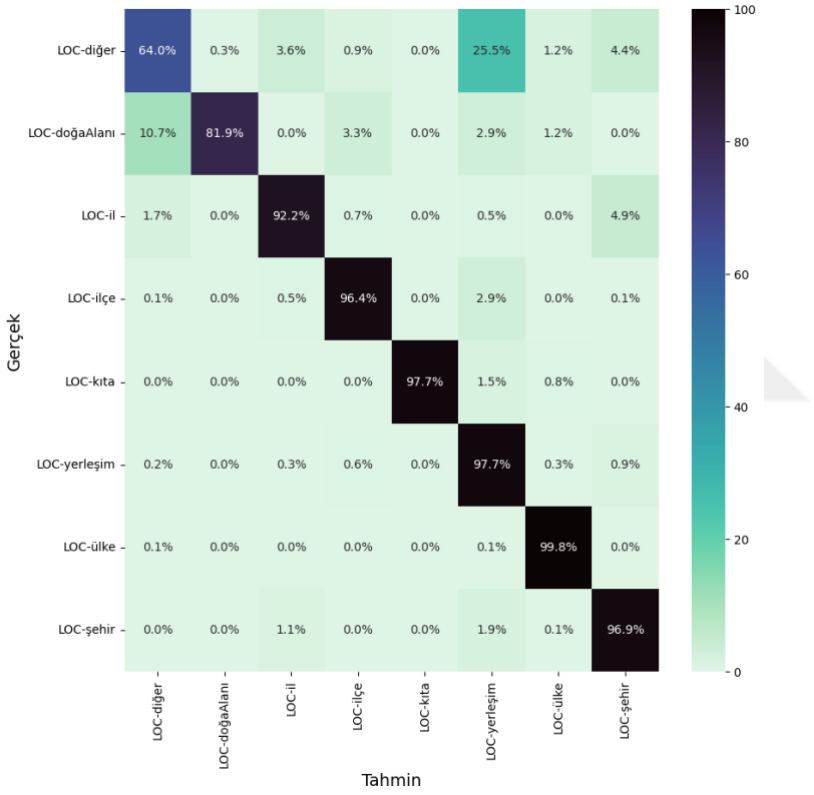
**Tablo 4.8** DistilBERTurk ile PER alt kırılımlı varlık tahmin sonuçları

| Varlık | Kesinlik | Duyarlılık | F1 |
|---|---|---|---|
| PER-akademisyen | 55,78 | 71,03 | 62,49 |
| PER-askeriPersonel | 63,56 | 66,78 | 65,13 |
| PER-diğer | 66,80 | 46,87 | 55,09 |
| PER-kraliyetMensubu | 62,18 | 62,41 | 62,27 |
| PER-politikacı | 66,01 | 78,56 | 71,74 |
| PER-sanatçı | 71,48 | 82,15 | 76,44 |
| PER-sporcu | 84,09 | 86,39 | 85,23 |

Şekil 4.4'de DistilBERT ile yapılan deney sonucu alınan LOC alt kırılım varlıkları hata matrisi verilmiştir. LOC-diğer varlığı en düşük doğruluğa sahip olup, %25'ten üzeri LOC-yerleşim olarak hatalı tahmin edilmiştir.

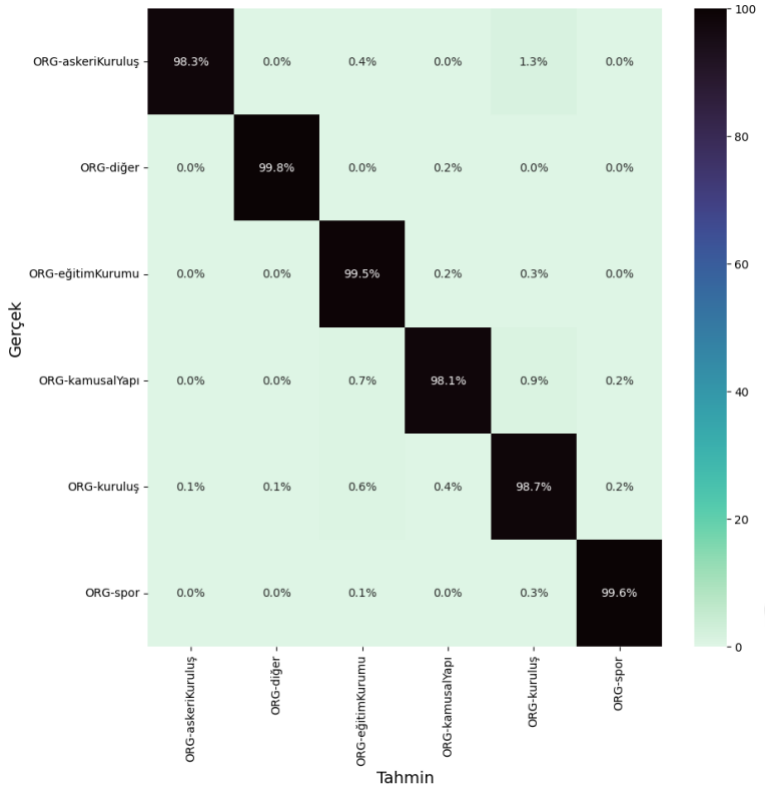
Şekil 4.5'deki ORG alt kırılım varlıkları hata matrisinden görüldüğü gibi, ORG-spor en hatasız varlık olarak tahmin edilmiş olup, ORG-kamusalYapı daha fazla hatalı tahmin edilmiştir.

Şekil 4.6'daki PER alt kırılım varlıkları hata matrisi verilmiştir. PER-diğer varlığının bir kısmı sırasıyla en çok PER-sanatçı (%18,7), PER-politikacı (%11,1) ve PER-akademisyen (%7,4) olarak hatalı tahmin edilmiştir.

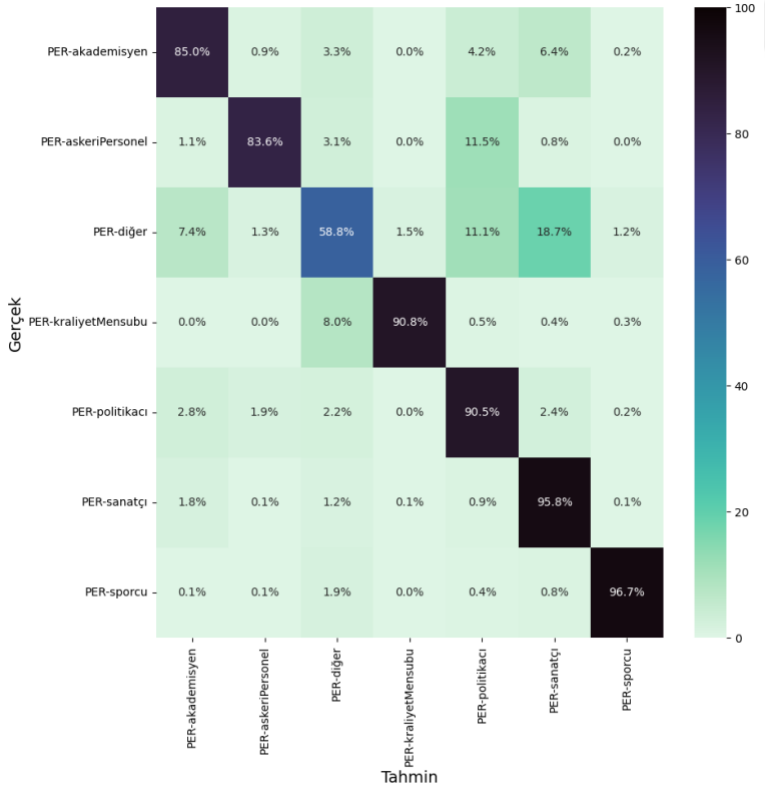



**Şekil 4.4** DistilBERTurk ile LOC alt kırılımlı VİT hata matrisi

PER-askeriPersonel ile PER-polikacı arasındaki hatalı tahminlere de yüksek oranda (%11,5) rastlanmıştır. PER-sporcu ve PER-sanatçı en hatasız tahmin edilen iki PER alt kırılımlı varlıkları olmuştur.




**Şekil 4.5** DistilBERTurk ile ORG alt kırılımlı VİT hata matrisi




**Şekil 4.6** DistilBERTurk ile ORG alt kırılımlı VİT hata matrisi

## 4.5 ConvBERTurk İle Yapılan Deneyler

ConvBERTurk [13] modeli 242 GBlik büyük birçok dilli veri seti üzerinde BERTurk ile ön eğitimli modeldir. Çalışmada ConvBERTurk ile farklı hiperparametler seçilerek deney yapılmıştır. Orijinal modelde maximum length sequence 512 olsa da, veri seti büyüklüğü ve Google Colabdaki kapasiteden dolayı çalışmada bu sayı 128 olarak ayarlanmıştır. Epoch sayısı 5, eğitim batch sayısı 32 olarak, test batch sayısı ise 16 olarak seçilmiştir.

### 4.5.1 Üst Varlık

Deneyler önce üst varlıklarla etiketlenen veri seti üzerinde yapılmıştır. Başarı sonuçları Tablo 4.5'de gösterilmiştir. ConvBERTurk ile alınan ortalama F1 başarı puanı 89,60 olmuştur. DATE için en yüksek 97,12 puanı, ORG için en düşük 81,83 puanı elde edilmiştir.

**Tablo 4.9** ConvBERTurk ile üst varlık tahmin sonuçları

| **Varlık** | **Kesinlik** | **Duyarlılık** | **F1** |
|---|---|---|---|
| DATE | 95,76 | 98,52 | 97,12 |
| LOC | 91,16 | 95,37 | 93,22 |
| ORG | 76,92 | 87,40 | 81,83 |
| PER | 82,49 | 90,32 | 86,23 |
| **Toplam** | **86,58** | **92,90** | **89,60** |

### 4.5.2 Alt Kırılımlı Varlık

ConvBERTurk ile yapılan alt kırılımlı VİT deneyinin başarı sonuçları DATE üst varlığı ve LOC alt kırılımlı varlıklar için Tablo 4.6, ORG alt kırılımlı varlıklar için Tablo 4.7 ve PER alt kırılımlı varlıklar için Tablo 4.8'de gösterilmiştir. Ortalama 75,11 Kesinlik, 83,35 Duyarlılık ve 78,88 F1 başarı puanı elde edilmiştir. DATE

için en yüksek 98,12 puanı, LOC-diğer için en düşük 59,66 puanı elde edilmiştir. LOC kategorisi altında en yüksek 94,17 puanı LOC-ülke için alınmıştır.

ORG varlığı içinde en düşük 71,08 puanıyla ORG-kamusalYapı, en başarılı 89,12 puanıyla ORG-spor olmuştur. ConvBERTurk modeli, ORG-kamusalYapı ile etiketli varlık isimlerinin BERTurk ve DistilBERT modellerine göre daha fazla LOC-diğer olarak hatalı tahmin etmiştir. *Chillon Şatosu, Veytaux sınırları içerisinde yer alan bir şatodur* cümlesindeki *Chillon Şatosu* ORG-kamusalYapı olarak etiketlenmiş olup, ConvBERTurk modeli bunu LOC-diğer olarak tahmin etmiştir.

PER altında PER-diğer en düşük 63,17, PER-sporcu ise en yüksek 88,08 puanını almıştır. ConvBERTurk ile elde edilen PER alt kırılımlı varlıklardaki başarılar, DistilBERT ve BERTurk modeli gibi olmuştur.

**Tablo 4.10** ConvBERTurk ile DATE ve LOC alt kırılımlı varlık tahmin sonuçları

| **Varlık** | **Kesinlik** | **Duyarlılık** | **F1** |
|---|---|---|---|
| DATE | 97,01 | 98,25 | 98,12 |
| LOC-diğer | 66,77 | 53,92 | 59,66 |
| LOC-doğaAlanı | 60,42 | 71,90 | 65,66 |
| LOC-il | 90,53 | 93,33 | 91,91 |
| LOC-ilçe | 85,60 | 93,68 | 89,46 |
| LOC-kıta | 62,37 | 78,40 | 69,47 |
| LOC-yerleşim | 85,74 | 91,81 | 88,67 |
| LOC-ülke | 92,39 | 96,03 | 94,17 |
| LOC-şehir | 81,93 | 93,00 | 87,11 |

**Tablo 4.11** ConvBERTurk ile ORG alt kırılımlı varlık tahmin sonuçları

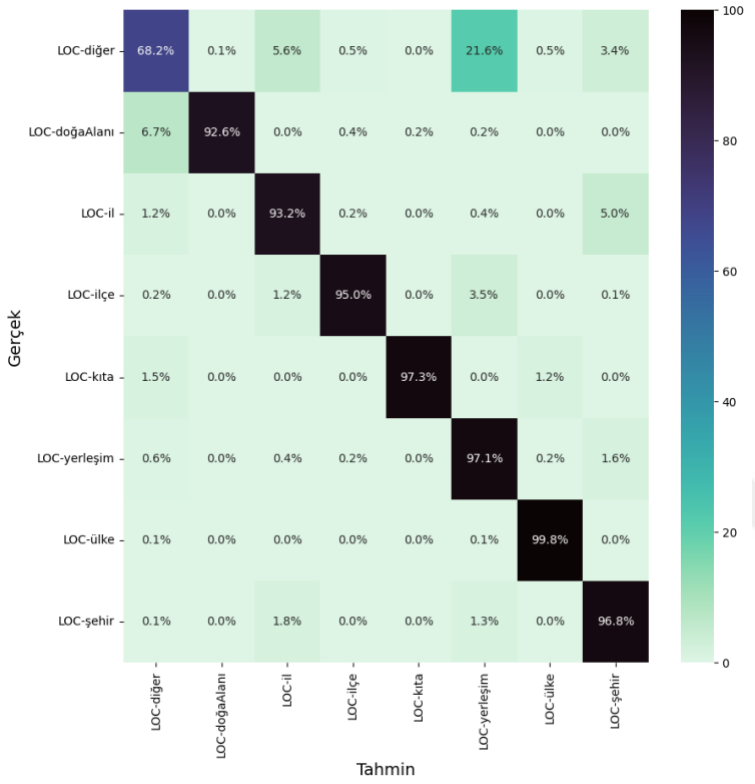
| Varlık | Kesinlik | Duyarlılık | F1 |
|---|---|---|---|
| ORG-askeriKuruluş | 67,13 | 80,81 | 73,33 |
| ORG-diğer | 77,21 | 92,90 | 84,33 |
| ORG-eğitimKurumu | 70,93 | 83,08 | 76,52 |
| ORG-kamusalYapı | 65,62 | 77,53 | 71,08 |
| ORG-kuruluş | 73,07 | 85,64 | 78,85 |
| ORG-spor | 87,05 | 91,29 | 89,12 |

**Tablo 4.12** ConvBERTurk ile PER alt kırılımlı varlık tahmin sonuçları

| Varlık | Kesinlik | Duyarlılık | F1 |
|---|---|---|---|
| PER-akademisyen | 63,82 | 79,41 | 70,77 |
| PER-askeriPersonel | 65,55 | 75,82 | 70,31 |
| PER-diğer | 65,10 | 61,35 | 63,17 |
| PER-kraliyetMensubu | 59,84 | 78,69 | 67,98 |
| PER-politikacı | 72,47 | 80,24 | 76,16 |
| PER-sanatçı | 77,44 | 85,81 | 81,41 |
| PER-sporcu | 85,52 | 90,80 | 88,08 |

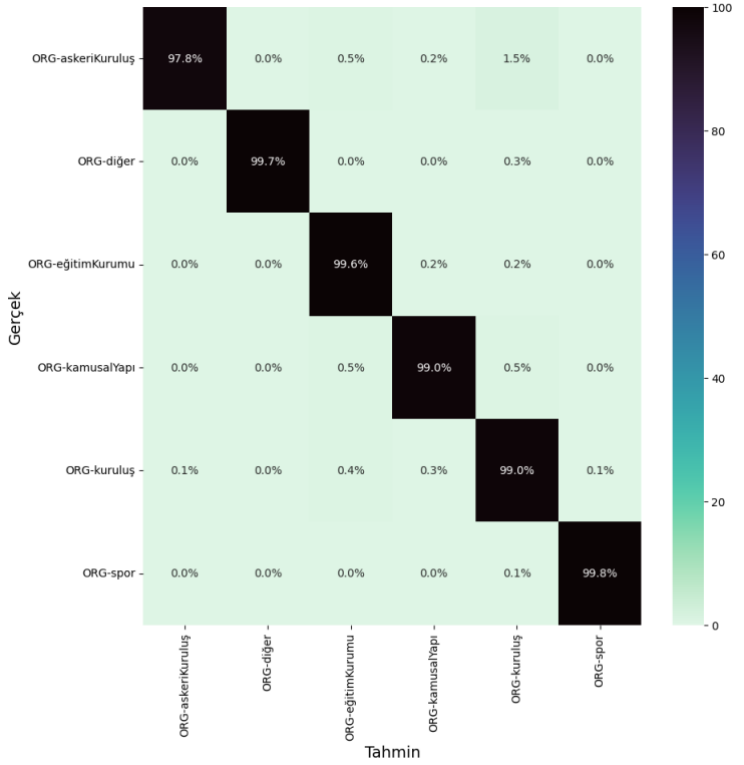
Şekil 4.7'de ConvBERTurk ile yapılan deney sonucu alınan LOC alt kırılım varlıkları hata matrisi verilmiştir. LOC-diğer varlığı en düşük doğruluğa sahip olup, %21,60 LOC-yerleşim, %5,60 ise LOC-il olarak hatalı tahmin edilmiştir.

Şekil 4.8'de ORG alt kırılım varlıkları hata matrisi gösterilmiştir. BERTurk ve DistilBERTurk'te olduğu gibi, ORG alt kırılım varlıkları tahminlerinde hata oranı düşüktür. ORG-spor en hatasız tahmin edilmiş olup, ORG-askeriKuruluş daha fazla hatalı tahmin edilmiştir.

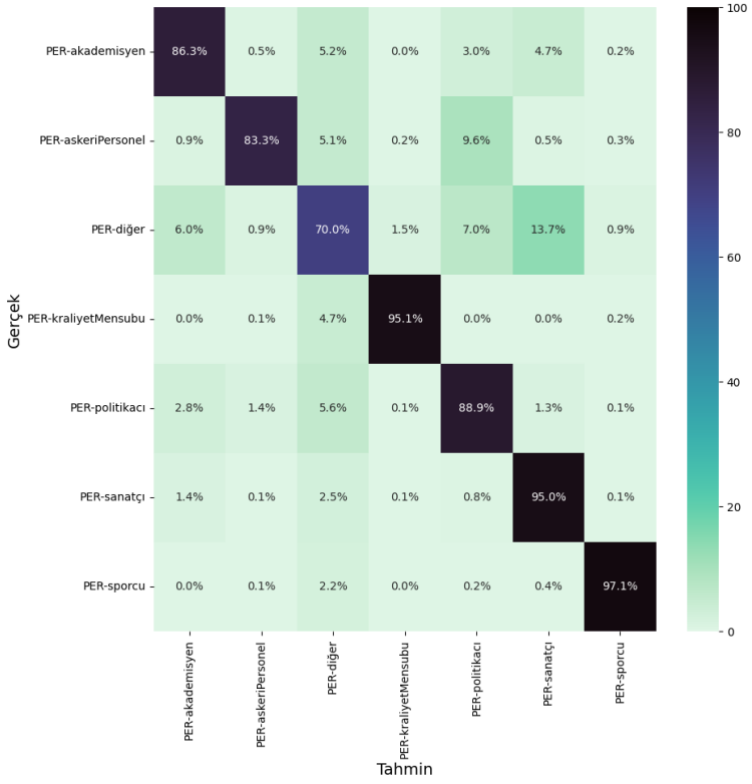



**Şekil 4.7** ConvBERTurk ile LOC alt kırılımlı VİT hata matrisi

Şekil 4.9'daki PER alt kırılım varlıkları hata matrisi verilmiştir. PER-diğer varlığın bir kısmı sırasıyla en çok PER-sanatçı (%13,70), PER-politikacı (%7,00) ve PER-akademisyen (%6,00) olarak hatalı tahmin edilmiştir. PER-askeriPersonel ile PER-polikacı arasındaki hatalı tahminlere de yüksek oranda (%9,60) rastlanmıştır. PER-sporcu ve PER-kraliyetMensubu en hatasız tahmin edilen iki PER alt kırılımlı varlıkları olmuştur.




**Şekil 4.8** ConvBERTurk ile ORG alt kırılımlı VİT hata matrisi




**Şekil 4.9** ConvBERTurk ile PER alt kırılımlı VİT hata matrisi

Alt kırılımlı VİT deneylerinde ConvBERTurk ile alınan ortalama başarı sonucu (78,88), DistilBERT başarı sonucuna (73,66) oldukça göre yüksek olup, BERTurk ile alınan F1 puanına (79,23) daha yakın olmuştur.

## 4.6 XLM-RoBERTa İle Yapılan Deneyler

XLM-RoBERTa veya XLM-R [26] modeli çok dilli model olup, RoBERTA'nın ön eğitim yaklaşımını kullanmıştır. Çalışmada üst ve alt kırılımlı varlıklarla etiketlenmiş veri seti üzerinde XLM-R modelini hiperparametre ayarları değiştirilerek transfer öğrenme yapılmıştır. Epoch sayısı 5, eğitim batch sayısı 32 olarak, test batch sayısı ise 16 olarak ayarlanmıştır. Maximum length sequence 512 olarak ayarlanmıştır.

### 4.6.1 Üst Varlık

Üst varlıklar üzerinde XLM-R ile alınan başarı sonuçları Tablo 4.7'de gösterilmiştir. Alınan ortalama F1 başarı puanı 93,34 olmuştur. DATE için en yüksek 99,22 puanı, PER için en düşük 89,43 puanı elde edilmiştir.

**Tablo 4.13** XLM-R ile üst varlık tahmin sonuçları

| Varlık | Kesinlik | Duyarlılık | F1 |
|---|---|---|---|
| DATE | 97,19 | 98,99 | 99,22 |
| LOC | 92,17 | 94,05 | 93,43 |
| ORG | 82,98 | 84,94 | 90,87 |
| PER | 84,97 | 87,39 | 89,43 |
| **Toplam** | **89,33** | **91,34** | **93,34** |

### 4.6.2 Alt Kırılımlı Varlık

XLM-R ile yapılan alt kırılımlı VİT deneyi sonucunda ortalama 78,00 kesinlik, 82,31 duyarlılık ve 79,82 F1 başarı puanı elde edilmiştir. DATE için en yüksek

99,76 puanı, LOC-diğer için en düşük 59,66 puanı elde edilmiştir (Tablo 4.8). LOC kategorisi altında en yüksek 97,98 puanı LOC-il için alınmıştır.

**Tablo 4.14** XLM-R ile DATE ve LOC alt kırılımlı varlık tahmin sonuçları

| Varlık | Kesinlik | Duyarlılık | F1 |
|---|---|---|---|
| DATE | 98,01 | 99,25 | 99,76 |
| LOC-diğer | 60,75 | 52,78 | 54,82 |
| LOC-doğaAlanı | 58,26 | 59,21 | 58,53 |
| LOC-il | 93,75 | 95,12 | 94,22 |
| LOC-ilçe | 90,35 | 95,53 | 93,16 |
| LOC-kıta | 54,23 | 82,99 | 65,52 |
| LOC-yerleşim | 86,99 | 91,23 | 89,33 |
| LOC-ülke | 98,63 | 96,46 | 97,98 |
| LOC-şehir | 89,58 | 92,13 | 91,54 |

Tablo 4.9'da PER alt kırılımlı varlıkların başarısı gösterilmiştir. PER altında PER-diğer en düşük 61,34, PER-sporcu ise en yüksek 89,65 puanını almıştır. Tablo 4.10'da ORG alt kırılımlı varlıkların başarısı gösterilmiştir. ORG varlığı içinde en düşük 71,27 puanıyla ORG-askeriKuruluş, en yüksek puan ise 90,69 ile ORG-spor varlığında elde edilmiştir.

XLM-R ile çoğu alt kırılımlı varlıklarda BERTurk tabanlı modellere göre daha yüksek sonuçlar elde edilmiştir. Ortalama F1 başarısı tüm modellerle yapılan deneyler arasında en yüksek olmuştur.

**Tablo 4.15** XLM-R ile ORG alt kırılımlı varlık tahmin sonuçları

| Varlık | Kesinlik | Duyarlılık | F1 |
|---|---|---|---|
| ORG-askeriKuruluş | 66,92 | 77,34 | 71,27 |
| ORG-diğer | 86,12 | 86,12 | 86,12 |
| ORG-eğitimKurumu | 85,34 | 88,41 | 86,94 |
| ORG-kamusalYapı | 83,22 | 78,87 | 80,34 |
| ORG-kuruluş | 78,36 | 81,30 | 79,23 |
| ORG-spor | 88,64 | 91,99 | 90,69 |

**Tablo 4.16** XLM-R ile PER alt kırılımlı varlık tahmin sonuçları

| Varlık | Kesinlik | Duyarlılık | F1 |
|---|---|---|---|
| PER-akademisyen | 66,96 | 74,67 | 70,37 |
| PER-askeriPersonel | 59,53 | 69,54 | 64,88 |
| PER-diğer | 71,38 | 53,70 | 61,34 |
| PER-kraliyetMensubu | 77,53 | 87,54 | 82,79 |
| PER-politikacı | 71,14 | 80,26 | 75,53 |
| PER-sanatçı | 75,34 | 84,44 | 79,28 |
| PER-sporcu | 85,99 | 94,22 | 89,65 |

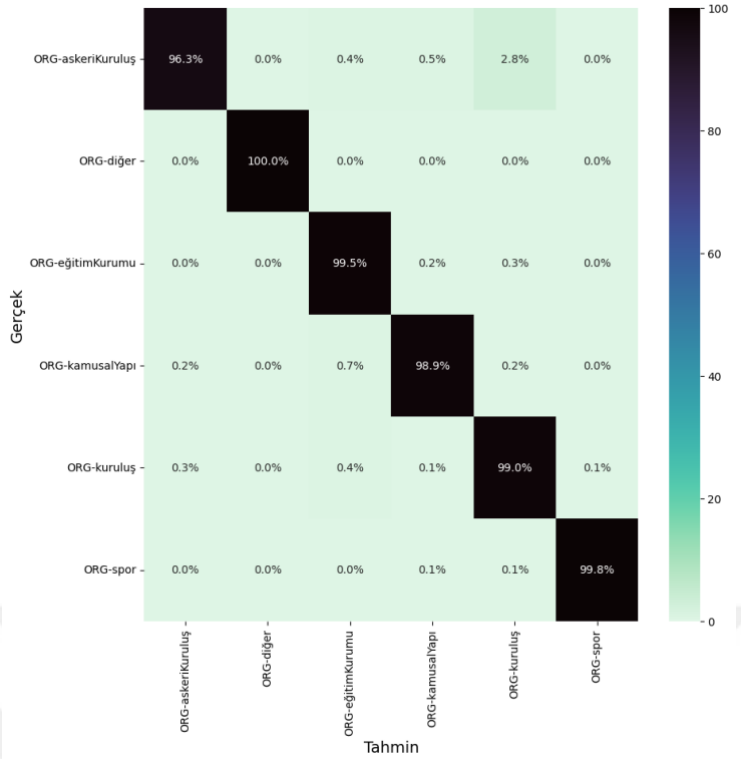
Her üst varlık altındaki alt kırılımlı varlıklar arasındaki ilişkiyi anlamak için hata matrisleri yapılmıştır. Şekil 4.10'deki LOC alt kırılım varlıkları hata matrisinden görüldüğü gibi, LOC-diğer en fazla hatalı tahmin edilmiştir. LOC varlıkları arasından LOC-diğer tahminlerde %55,5 oranında doğru etiketlenmiş olup, %37,2 LOC-yerleşim olarak hatalı etiketlenmiştir. XLM-R ile yapılan deneyde LOC-diğer varlığı için doğruluk oranı, BERT tabanlı modellerde alınan doğruluk oranına göre oldukça düşük çıkmıştır.




**Şekil 4.10** XLM-R ile LOC alt kırılımlı VİT hata matrisi

Şekil 4.11'de görüldüğü gibi ORG varlığı arasında ORG-diğer tamamen hatasız tahmin edilmiş olup, ORG-askeriKuruluş en düşük doğrulukla tahmin edilmiştir. PER varlığı arasında PER-diğer varlığı çok hatalı tahmin edilmiştir (Şekil 4.12). PER-diğer %18,2 oranında PER-sanatçı olarak hatalı tahmin edilmiştir. PER-askeriPersonel PER-polikacı olarak %7,2 oranında hatalı tahmin edilmiştir.

XLM-R modeli hata matrislerinden görüldüğü üzere, XLM-R modeli yakın ilişkili alt kırılımlı varlıklar arasında BERTurk ve ConvBERTurk'e göre daha yüksek hata oranına sahiptir. ORG-askeriKuruluş, PER-diğer ve LOC-diğer kategorisinde alınan hata oranı, BERTurk ve ConvBERTurk modeline göre daha yüksek olmuştur.




**Şekil 4.11** XLM-R ile ORG alt kırılımlı VİT hata matrisi

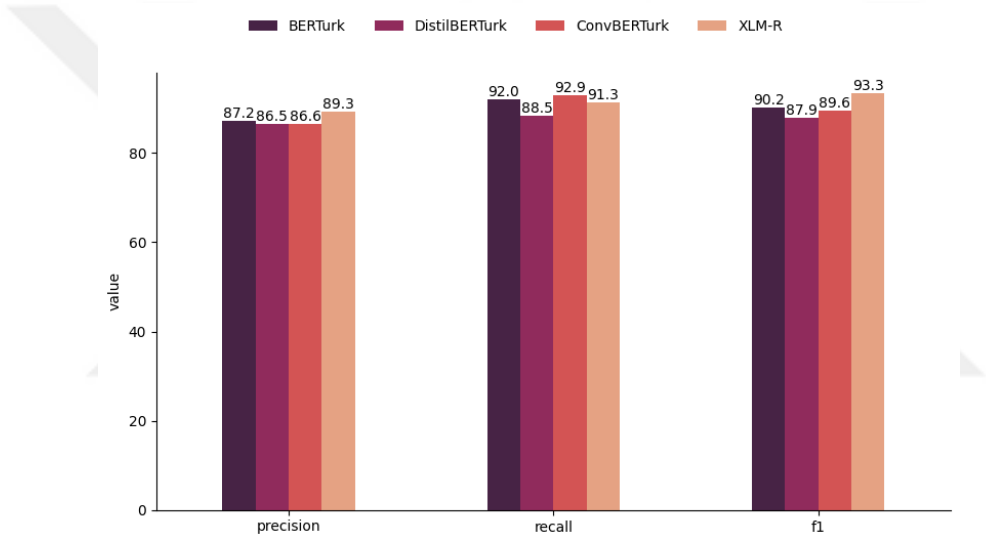



**Şekil 4.12** XLM-R ile PER alt kırılımlı VİT hata matrisi

## 4.7 Sonuçların Özeti

Dört farklı modelle hem alt kırılımlı varlık ismi tanıması, hem de üst VİT deneyleri yapılmıştır. XLM-R modeli hem alt kırılımlı hem de üst varlıklı veri setlerinde en yüksek F1 başarıya ulaşmıştır (Şekil 4.13, Şekil 4.14). Bunu BERTurk-128 ve ConvBERTurk takip etmiştir. En düşük performanslı model DistilBERTurk olmuştur, bu da onun daha küçük veri seti üzerinde eğitilmiş ve daha hafif olmasından kaynaklıdır.

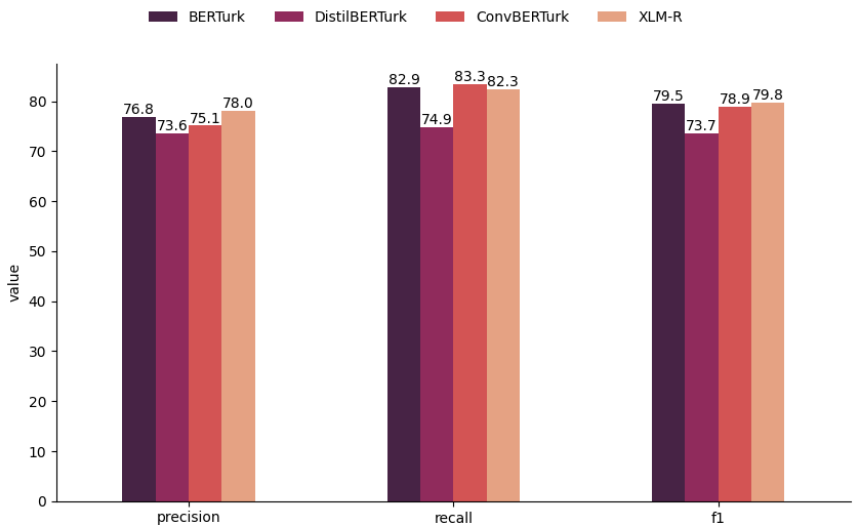
Çalışma veri setine benzer WikiANN [9] veri setinin Türkçe versiyonu üzerinde üst varlık ismi tanıması [8] çalışması sonucunda literatürde en yüksek 93,80'lik ve [12] sonucu 92,23'lük bir F1 başarısı elde edilmiştir.




**Şekil 4.13** Üst varlıkların ortalama başarı sonuçları

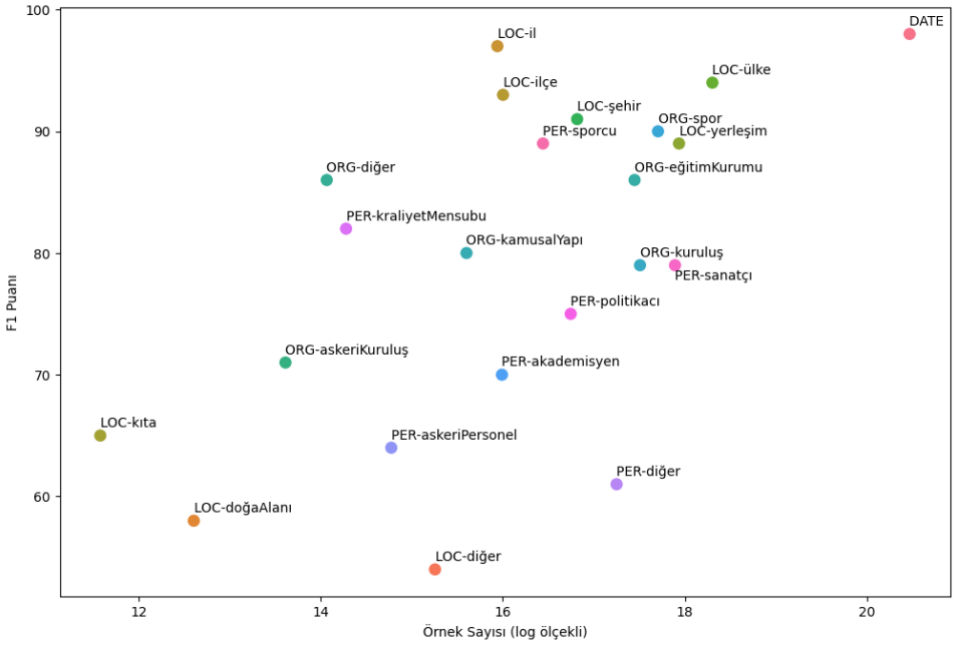
Her bir modelde ilk DATE, sonra LOC-ülke alt kırılımlı varlığı en yüksek başarıya ulaşmıştır. Tüm modellerde en düşük başarıya sahip alt kırılımlı varlık LOC-diğer olmuştur. ORG alt kırılımlı varlıklarından ise birbirine yakın F1 başarısı alındığı görülmüştür. Ayrıca bu varlıklarda, PER ve LOC varlıklarına göre kendi içinde daha az hatalı tahminler yapılmıştır.

Şekil 4.13 ve Şekil 4.14'te modeller arası F1, duyarlılık (recall) ve hassasiyet (precision) karşılaştırması yapılmıştır. XLM-R ile her üç gösterge arası daha dengeli sonuç elde edilmiştir. BERTurk ve ConvBERTurk ile yüksek duyarlılık, ama hassasiyette daha düşük puan alınmıştır. Bu modellerin veri seti üzerindeki eğitimin gerçek pozitif örneklerin çoğunu tanımladığını ancak yanlış pozitiflere eğilimli olduğunu göstermiştir.




**Şekil 4.14** Alt kırılımlı varlıkların ortalama başarı sonuçları

Alınan başarı ile varlık örnek sayısı arasında pozitif bir ilişki kaydedilmiştir (Şekil 4.13). Varlık örnek sayısı ne kadar yüksek ise, o kadar daha yüksek başarı elde edilmiştir. Veri setinde zayıf gözetimle yapılan etiketleme işleminde, LOC-il ve LOC-diğer varlıkları arasındaki hatalı etiketlemeler olmuştur. Bu sebeple, LOC-il ve LOC-diğer varlık isimlerinde, varlık sayısı ve başarı puanı ilişkisinde daha az pozitif bir yön görülmektedir.




**Şekil 4.15** Varlık sayısı ve başarı puanı arasındaki ilişki

Gerçek ve tahmin edilen etiketlemelerin karşılaştırılması yapıldığında dönüştürücü tabanlı modellerin veri üzerinde bağlama daha fazla dikkat ettiği görülmüştür (Şekil 4.16). Vikipedi meslek bilgisine göre etiketlenen varlıklar, bağlama göre daha farklı bir varlık ile de etiketlenebilirdi. PER-askeriPersonel varlığının PER-politikacı olarak, PER-diğer varlığının PER-sanatçı olarak bağlama bağlı şekilde etiketlenmiş olduğu durumları çok görülmüştür.




**Şekil 4.16** Sol: veri seti etiketi, sağ: bağlama bağlı etiket tahmini

Kaynak Vikipedi'de eksik bilgi kutusu ve gürültülü metinler de mevcuttur. Etiketleme işlemi zayıf gözetimli yöntemle yapıldığından eksik etiketlenmiş varlıklar ortaya çıkmıştır. Şekil 4.17'de görüldüğü üzere veri setinde PER-sanatçı olarak etiketlenebilecek varlıklar hatalı veya eksik etiketlenmiştir.




**Şekil 4.17** Sol: Veri setindeki eksik etiketleme, sağ: bağlama bağlı etiket tahmini

Veri setinde bazı varlık isimlerinde doğru ve eksiksiz etiketleme yapılırken, yanlış tahmin edilen durumlar da gözlemlenmiştir. Şekil 4.18'de görüldüğü gibi, kişi ismi ile bağlı organizasyonlar yanlış olarak ORG-kuruluş olarak etiketlenmiştir.




**Şekil 4.18** Sol: Veri setindeki doğru etiketleme, sağ: hatalı etiket tahmini

# 5
# SONUÇ

Bu çalışma sonucunda Türkçe Varlık İsmi Tanıma alanında üst ve alt kırılımlı varlıklarla otomatik olarak etiketlenmiş yeni bir veri seti oluşturulmuştur. Veri Vikipedi metinlerinden alınmıştır. Etiketleme Vikipedi bilgi kutularından yararlanılarak zayıf gözetimli yöntemle yapılmıştır. Etiketler seçilirken Vikipedi bilgi kutuları listesinden DATE, ORG, PER ve LOC varlıkları altında gruplanabilecek kategoriler seçilmiştir. BIO formatında yeni 22 varlık çıkarılmış ve oluşturulan iki farklı veri seti üzerinde farklı dönüştürücü tabanlı modellerle testler yapılmıştır.

Denemeler sonucunda XLM-R modelinin hem üst hem de alt kırılımlı varlıklar üzerinde en yüksek başarıya ulaştığı görülmüştür. DistilBERT ile en düşük performanslar elde edilmiştir. Deneyler sonucunda modellerin bağlama bağlı tahminlere önem verdiği gözlenmiştir.

Gelecekteki çalışmalarda, veri setlerinde üst varlıklar daha fazla kategorilere ayrılarak varlık sayısı arttırılabilir. Modeller ile yapılan deneylerde 128'den daha yüksek olan 512 en fazla dizi uzunluğu (maximum sequence length) ayarı seçilerek daha yüksek başarılar alınabilir. Vikipedi bilgi kutusundaki çoklu meslek bilgisinden faydalanılarak bağlama bağlı etiketleme işlemi yapılabilir.

# KAYNAKÇA


[1] R. Grishmanand, B. Sundheim, "Message Understanding Conference-6: A brief history," in Proceedings of the *16th International Conference on Computational Linguistics*, 1996.

[2] F. Tjong Kim Sang E, "Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition," in Proceedings of the *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*, 2002.

[3] A. Şeker, G. Eryiğit, "Extending a CRF-based Named Entity Recognition Model for Turkish Well Formed Text and User Generated Content," *Semantic Web*, vol 8, pp. 1-18, 2017.

[4] D. Lafferty J, A. McCallum and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in Proceedings of the *Eighteenth international conference on machine learning*, 2001.

[5] G. Çelikkaya, D. Torunoğlu and G. Eryiğit, "Named Entity Recognition on Real Data: A Preliminary Investigation for Turkish," in Proceedings of the *7th International Conference on Application of Information and Communication Technologies*, 2013.

[6] A. Güneş, C. Tantuğ, "Turkish Named Entity Recognition with Deep Learning," in Proceedings of the *26th Signal Processing and Communications Applications Conference*, 2018.

[7] Z. Huang, W. Xu and K. Yu, "Bidirectional LSTM-CRF Models for Sequence Tagging," *arXiv preprint arXiv:1508.019*91v1, 2015.

[8] A. Safaya, E. Kurtulus, A. Göktogan and D. Yuret, "Mukayese: Turkish nlp strikes back," in Proceedings of the *Findings of the Association for Computational Linguistics: ACL 202*2, 2022.

[9] X. Pan, B. Zhang, J. May, J. Nothman, K. Knight and H. Ji, "Cross-lingual Name Tagging and Linking for 282 Languages," in Proceedings of the *5th Annual Meeting of the Association for Computational Linguistics*, 2017.

[10] G. Tür, D. Hakkani-Tür and K. Oflazer, "A statistical information extraction system for Turkish," *Natural Language Engineering*, vol 9, pp. 181-210, 2003.

[11] J. Devlin, M. Chang, K. Lee and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *arXiv preprint arXiv:1810.04805v2*, 2019.

[12] O. Özçelik and Ç. Toraman, “Named entity recognition in Turkish: a comparative study with detailed error analysis,” *Information Processing and Management*, vol. 59, pp. 103065, 2022.

[13] HuggingFace, “Turkish ConvBERT model,” https://huggingface.co/dbmdz/convbert-base-turkish-mc4-uncased, 15.06.2023.

[14] Turkish ELECTRA models, https://github.com/stefan-it/turkish-bert/blob/master/electra/README.md, 04.05.2023.

[15] X. Ling and D. S. Weld, “Fine-grained entity recognition,” Twenty- in Proceedings of the *Sixth AAAI Conference on Artificial Intelligenc*e, 2012.

[16] G. D, N. Lazic, K. Ganchev, J. Kirchner and D. Huynh, “Context-Dependent Fine-Grained Entity Type Tagging,” *arXiv preprint arXiv:1412.1820v2*, 2016.

[17] E. Choi, O. Levy, Y. Choi and L. Zettlemoyer, “Ultra-fine entity typing,” in Proceedings of the *56th Annual Meeting of the Association for Computational Linguistics* (Long Papers), 2018.

[18] C. Lee, H. Dai, X. Song and X. Li, “A chinese corpus for fine-grained entity typing,” in *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 2020.

[19] A. Liqreina H, J. M, K. M, O. El-Shangiti A and M. Abdul-Mageed, “Arabic Fine-Grained Entity Recognition,” *arXiv preprint arXiv:2310.17333v1*, 2023.

[20] M. Jarrar, M. Khalilia and S. Ghanem, “Wojood: Nested Arabic Named Entity Corpus and Recognition using BERT,” *arXiv preprint arXiv:2205.09651v2*, 2022.

[21] D. Dinç, “Financial Named Entity Recognition for Turkish News Texts,” A Thesis Submitted to *The Graduate School of Natural And Applied Sciences Of Middle East Technical University*, 2022.

[22] C. Çetindağ, B. Yazıcıoğlu and A. Koç, “Named-entity recognition in Turkish legal texts,” *Natural Language Engineering*, vol. 29, pp. 615-642, 2023.

[23] Ö. Tanrısever, E. T. Ayan, M. S. Zengin, H. A. Duru and B. Bardak, “Named Entity Recognition for Defense Industry,” in Proceedings of the

30th *Signal Processing and Communications Applications Conference (SIU)*, 2022.

[24] M. Topcu and İ. Durgar El-Kahlout, "TR-SEQ: Named Entity Recognition Dataset for Turkish Search Engine Queries," in Proceedings of the *Recent Advances in Natural Language Processing*, 2021.

[25] Y. P. Kılıç, Named Entity Recognition and Explainability Analysis on Turkish Sports News Texts, A Thesis Submitted to *The Graduate School of Natural And Applied Sciences Of Middle East Technical University*, 2023.

[26] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzman, E. Grave, M. Ott, L. Zettlemoyer and V. Stoyanov, "Unsupervised Cross-lingual Representation Learning at Scale," *arXiv preprint arXiv:1911.02116v2*, 2020.

[27] Vikipedi Bilgi kutuları listesi, https://tr.wikipedia.org/wiki/Vikipedi:Bilgi_kutular%C4%B1_listesi, 07.01.2023

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, "Attention Is All You Need," *arXiv preprint arXiv:1706.03762*, 2017.

[29] S. Schweter, "BERTurk - BERT models for Turkish," https://zenodo.org/records/3770924, 20.05.2023

[30] HuggingFace, Turkish BERT model, https://huggingface.co/dbmdz/bert-base-turkish-uncased, 05.02.2024.

[31] HuggingFace, "Distilled Turkish BERT model," https://huggingface.co/dbmdz/distilbert-base-turkish-cased, 14.06.2023.

[32] K. T. Uçar, "BERT Modeli ile Türkçe Metinlerde Sınıflandırma Yapmak," https://medium.com/@ktoprakucar/bert-modeli-ile-t%C3%BCrk%C3%A7e-metinlerde-s%C4%B1n%C4%B1fland%C4%B1rma-yapmak-260f15a65611, 02.03.2024

[33] Carnegie Mellon University, "Kemal Oflazer," https://sites.google.com/andrew.cmu.edu/kemal-oflazer, 02.05.2023.

[34] Opus corpora, https://opus.nlpl.eu, 02.05.2023.

[35] Open Source Project on Multilingual Resources for Machine Learning (OSCAR), https://oscar-project.org, 02.05.2023.

[36] Z. Jiang, W. Yu, D. Zhou, Y. Chen, J. Feng and S. Yan, “Convbert: Improving BERT with span-based dynamic convolution,” *arXiv preprint arXiv:2008.02496v3*, 2021.

[37] The C4 Multilingual Dataset, https://github.com/allenai/allennlp/discussions/5265, 6.2.2024.

[38] G. Lample, A. Conneau, “Cross-lingual language model pretraining,” *arXiv preprint arXiv:1901.07291v1*, 2019.

[39] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov, “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” *arXiv preprint arXiv:1907.11692*, 2019.

[40] HuggingFace, “XLM-RoBERTa Model,” https://huggingface.co/FacebookAI/xlm-roberta-base, 15.06.2023.

# A
# EK

**Tablo A.1** XLM-R ile yapılan deney sonucunda tokenlerin gerçek ve tahmin etiketleri karşılaştırılması

| **Token** | **Gerçek** | **Tahmin** |
|---|---|---|
| Yamaha | B-ORG-kuruluş | B-ORG-kuruluş |
| , | O | O |
| Torakusu | B-PER-diğer | O |
| Yamaha | I-PER-diğer | B-ORG-kuruluş |
| tarafından | O | O |
| 1887 | B-DATE | B-DATE |
| yılında | O | O |
| kurulmuş | O | O |
| olan | O | O |
| bir | O | O |
| Japon | O | O |
| şirketidir | O | O |
| . | O | O |

# TEZDEN ÜRETİLMİŞ YAYINLAR

**Konferans Bildirileri**

**1**. L.Khudoyberdieva, D.Diri, "Türkçe için Alt Kategorili Varlık İsmi Tanıması", in Proceeding of the *32th Signal Processing and Communications Applications Conference (SIU2024)*, 2024.