



T.C.
EGE ÜNİVERSİTESİ
Fen Bilimleri Enstitüsü



İLAÇ-BİTKİ ETKİLEŞİMLERİ İÇİN BİR VERİ KÜMESİ
OLUŞTURULMASI VE MAKİNE ÖĞRENMESİ İLE
TAHMİNLENMESİ

Doktora Tezi

Erkan YAŞAR

Bilgisayar Mühendisliği Ana Bilim Dalı

İzmir
2022

T.C.
EGE ÜNİVERSİTESİ
Fen Bilimleri Enstitüsü

**İLAÇ-BİTKİ ETKİLEŞİMLERİ İÇİN BİR VERİ
KÜMESİ OLUŞTURULMASI VE MAKİNE
ÖĞRENMESİ İLE TAHMİNLENMESİ**

Erkan YAŞAR

Danışman : Dr. Öğr. Üyesi Özgür GÜMÜŞ

Bilgisayar Mühendisliği Anabilim Dalı
Bilgisayar Mühendisliği Doktora Programı

İzmir
2022

EGE ÜNİVERSİTESİ FEN BİLİMLERİ ENSTİTÜSÜ

ETİK KURALLARA UYGUNLUK BEYANI

EÜ Lisansüstü Eğitim ve Öğretim Yönetmeliğinin ilgili hükümleri uyarınca Yüksek Lisans Tezi / Doktora Tezi olarak sunduğum **“İLAÇ-BİTKİ ETKİLEŞİMLERİ İÇİN BİR VERİ KÜMESİ OLUŞTURULMASI VE MAKİNE ÖĞRENMESİ İLE TAHMİNLENMESİ”** 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.

09 / 12 / 2022

Erkan YAŞAR

ÖZET**İLAÇ-BİTKİ ETKİLEŞİMLERİ İÇİN BİR VERİ KÜMESİ
OLUŞTURULMASI VE MAKİNE ÖĞRENMESİ İLE
TAHMİNLENMESİ**

YAŞAR, Erkan

Doktora Tezi, Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanı: Dr. Öğretim Üyesi. Özgür GÜMÜŞ

Aralık 2022, 109 sayfa

İnsanlık yazılı kaynaklardan önceki ve sonraki binlerce yıldır hastalıklardan korunmak ve hastalıkları tedavi etmek amacıyla şifalı bitkileri kullanmıştır. Son iki yüzyılda ise sanayi ve teknoloji alanında dünyada meydana gelen devrim niteliğindeki gelişmeler sağlık alanında da kendine yer bulmuş ve hastalıkları tedavi etmek amacıyla reçeteli (sentetik) ilaçlar geliştirilmiştir ve tedavi için bitkilerin yerini almışlardır. Bu gelişmelere rağmen bitkisel ilaçların doğal olması kanısı nedeniyle kanser gibi ciddi hastalıkları olan hastalar, tedavinin daha da etkin olabilmesi amacıyla sağlık danışmanlarına haber vermeden bitkisel ilaçları reçeteli ilaçlar ile birlikte almaya devam etmektedirler. Ancak şifa niyetine kullanılan bu bitkisel ilaçlar reçeteli ilaçlarla birlikte eş zamanlı olarak alındıklarında istenmeyen (advers) ciddi etkileşimlere girmekte ve tedavinin başarısız olmasına neden olmaktadır.

Bitki ve ilaçların eş zamanlı alınmasından kaynaklı etkileşimleri bulmak ve gelecekte meydana gelebilecek potansiyel etkileşimleri tahmin edebilmek için veri kümelerine ve makina öğrenmesi tekniklerine ihtiyaç vardır. Ancak bu konu hakkında sınırlı çalışmalar yapılmıştır. Bunun en büyük nedenlerinden biri, bu araştırmaları sonuçlandırabilecek herhangi kapsamlı bir veri kümesinin henüz bulunmaması ve ayrıca standart bir metodolojinin olmamasıdır. Literatürde bitki ilaç etkileşimi ile ilgili yapılan en kapsamlı çalışmalar biyomedikal literatürde yapılandırılmamış düz metin şeklinde araştırma makalelerinde yer almaktadır.

Bu bağlamda bu tez çalışmasında milyonlarca makalenin yer aldığı biyomedikal literatürden ilaç ve bitki varlıklarının birlikte geçtiği cümleleri ve dokümanları otomatik olarak yakalayabilen ve bu varlıklar arasında, ilişki olup olmadığını tespit eden ve ilişkinin tipini ortaya çıkartan bilgi çıkarım (information extraction) sistemlerinin alt dalı olan makina öğrenmesi (SVM, Random Forest,

KNN) ve son teknoloji derin sinir ağı tabanlı (Tok2vec, BERT, RoBERTa, BioBERT) Adlandırılmış Varlık Tanıma (Name Entity Recognition- NER) ve İlişki Çıkarım (Relationship Extraction-RE) modelleri geliştirilmiştir. Ayrıca kural ve makina öğrenmesinin birleştiği zayıf denetimli öğrenme tabanlı bir ilişki çıkarım modelide geliştirilmiştir. Bu modellerin performansları birbiri ile karşılaştırılarak en yüksek performansı veren model temel (baseline) model olarak ele alınmıştır.

Bu modellerin geliştirilebilmesi için eğitim veriseti olarak kullanılabilecek altın standart derlemlere ihtiyaç vardır. Bu nedenle bu tez çalışmasında 3 altın standart derlem geliştirilmiştir. İlk olarak bitki ve ilaç etkileşimleri ile ilgili bilgilerin ve PubMed makale özetlerinin yer aldığı bir veriseti hazırlanmıştır. 680 adet makale özetindeki bitki ve ilaç varlıkları alan uzmanı küratörler tarafından hazırlanan bir protokole göre etiketlenerek varlık etiketli altın standart bir derlem hazırlanmıştır. İkinci olarak bitki ve ilaç varlıkları arasındaki ilişkiler etiketlenerek 100 adet makale özeti ilişki türü (farmakokinetik ve farmakodinamik etkileşimler) seviyesinde etiketlenmiştir ve son olarak 150 adet makale özeti ise ilişki tespit (POSITIVE, NEGATIVE, UNCERTAIN) seviyesinde etiketlenerek altın standart derlemler oluşturulmuştur.

Anahtar sözcükler: Doğal dil işleme, altın standart derlem, bilgi çıkarımı, adlandırılmış varlık tanıma, ilişki çıkarımı, bitki ilaç etkileşimi, makina öğrenmesi ,BERT, BioBERT, RoBERTa, SVM, KNN, Rastgele Orman

ABSTRACT**CREATING A DATASET FOR DRUG-HERB INTERACTIONS AND
PREDICTION WITH MACHINE LEARNING**

YAŞAR, Erkan

PhD in Computer Engineering

Supervisor: Assoc. Prof. Dr. Özgür GÜMÜŞ

December 2022, 109 pages

For thousands of years before and after written sources, humanity has used medicinal herbs to prevent and cure diseases. In the last two centuries, revolutionary developments in the world have found their place in the field of health and prescription drugs have been developed to treat diseases and have replaced herbs for treatment. Despite these developments, patients with serious diseases such as cancer, due to the belief that herbal medicines are natural, In order to make the treatment more effective, they continue to take herbal medicines together with synthetic prescription drugs without informing their health clients. However, when these herbal medicines used for healing purposes are taken simultaneously with prescription drugs, they enter into serious adverse interactions and cause the treatment to fail.

Databases and machine learning techniques are needed to find interactions caused by simultaneous intake of herbs and drugs and to predict potential future interactions. However, limited studies have been done on this subject. One of the biggest reasons for this is that there is not yet any comprehensive database that can conclude these studies, and also that there is no standard methodology. The most comprehensive studies on herb drug interaction in the literature are in unstructured plain text research articles in the biomedical literature.

In this context, in this thesis, machine learning and deep learning-based "named entity recognition" and "relationship extraction" models have been developed that can automatically capture sentences and documents from the biomedical literature where drug and herb entities go together, and reveal whether there is a relationship and the type of relationship between these entities.

In order to develop these models, gold standard corpus that can be used as training datasets are needed. Therefore, 3 gold standard corpus has been developed in this thesis. First, a dataset containing information on herb and drug interactions and PubMed article summaries was prepared. A gold standard corpus was prepared by labeling the herb and drug entities in 680 article abstracts according to a protocol prepared by field expert curators. Secondly, by labeling the relationships between herb and drug entities, 100 article abstracts were labeled at the relationship type (pharmacokinetic and pharmacodynamic interactions) level, and finally, gold standard corpus was created by labeling 150 article abstracts at the relationship detection level (POSITIVE, NEGATIVE, UNCERTAIN).

Keywords: Natural language processing, gold standart corpus, NLP, information extraction, named entity recognition, relation extraction, herb drug interaction, machine learning, BERT, BioBERT, RoBERTa, SVM, KNN, Random Forest

ÖNSÖZ

Bu tez çalışması danışmanım Doktor Öğr. Üyesi Özgür Gümüş, Prof. Dr. Oğuz Dikenelli ve Hollanda Maastricht Üniversitesinden Asistan Prof. Remzi Çelebi ile birlikte yaptığımız ilaç-ilac etkileşimlerinin tahmini çalışmasından sonra ortaya çıkmıştır. Literatürde ilaç-ilac etkileşimi ile ilgili veri kaynakları mevcut iken ilaç-bitki etkileşimleri ile ilgili bir kaynak mevcut değildi. Bu nedenle makina öğrenmesi tabanlı doğal dil işleme teknikleri aracılığıyla bu faydalı bilgilerin çıkarılması için altın standart derlemlere ihtiyaç duyulmuştur. Bunun için Biyokimya bölümünden Prof. Dr. Figen Zihnioğlu ve öğrencisi Araş. Gör. Kerem Tok ile saatler ve günler süren varlık ve ilişki etiketli derlemler hazırlanmıştır. Bu tez çalışması bu nedenle Biyokimya ve Bilgisayar mühendisliğinin ortak emek verdiği disiplinlerarası bir çalışmadır.

İZMİR

09/12/2022

Erkan YAŞAR



İÇİNDEKİLER

	<u>Sayfa</u>
İÇ KAPAK	(ii)
KABUL ONAY SAYFASI	(iii)
ETİK KURALLARA UYGUNLUK BEYANI.....	(iv)
ÖZET	(v)
ABSTRACT	(vii)
ÖNSÖZ	(ix)
İÇİNDEKİLER DİZİNİ.....	(xi)
ŞEKİLLER DİZİNİ	(xvii)
ÇİZELGELER DİZİNİ.....	(xx)
SİMGELER VE KISALTMALAR dİZİNİ.....	(xxiii)
1. GİRİŞ.....	1
2. BİTKİLER ÜZERİNE GENEL BİR BAKIŞ	7
2.1. Geçmişten Günümüze Bitkisel İlaçların Kullanım Tarihi	7
2.2. Bitkisel İlaç Kullanım Nedenleri	9
2.2.1 Erişilebilirlik ve satın alınabilirlik	9
2.2.2 Sağlıkta alternatif bir yaklaşım	9
2.3. İlaç Etkileşim Tipleri	10
2.4 İlaç-Bitki Etkileşimi.....	12

İÇİNDEKİLER (devam)

Sayfa

2.4.1 İlaç-bitki etkileşim mekanizması	12
2.4.1.1 Farmakokinetik ilaç-bitki etkileşimleri	12
2.4.1.2 Farmakodinamik ilaç-bitki etkileşimleri	14
3. DOĞAL DİL İŞLEMEYE GİRİŞ (DDİ)	15
3.1 Doğal Dil İşleme Nedir	16
3.2 Yapay Zeka, Makina Öğrenmesi, Derin Öğrenme ve DDİ İlişkisi	16
3.3 Doğal Dil İşleme Görevleri	17
3.4 Verilerin Yapısı	18
3.4.1 Yapılandırılmış veriler (Structured data)	19
3.4.2 Yarı Yapılandırılmış veriler (Semi Structured data)	20
3.4.3 Yapılandırılmamış veriler (Unstructured data)	20
3.5 Biyomedikal Verilerin Genel Olarak Yapısı	21
4. BİLGİ ÇIKARIMI (INFORMATION EXTRACTION)	23
4.1 Adlandırılmış Varlık Tanıma (Name Entity Recognition-NER)	23
4.1.2 AVT sistemlerine neden ihtiyaç duyulmakta?	24
4.1.3 Biyomedikal alanda AVT (NER) ve karşılaştıkları zorluklar	25
4.2 İlişki Çıkarımı (Relationship Extraction)	27

İÇİNDEKİLER (devam)

Sayfa

4.3 Adlandırılmış Varlık Tanıma ve İlişki Çıkarım Sistemlerinin Makina Öğrenmesi Modelleri İle Tahmini İçin Gerekli Etiketli Veri Kümesi Oluşturma Süreçleri.....	30
5. ÖNCEKİ ÇALIŞMALAR	33
5.1 Adlandırılmış Varlık Tanımlama Sistemleri İçin İlk Yöntemler.....	33
5.1.1 Sözlük tabanlı yöntemler	33
5.1.2 Kural tabanlı yöntemler	35
5.2 AVT (NER) Sistemleri İçin Öğrenme Tabanlı Yöntemler.....	37
5.2.1 Denetimli öğrenme (Supervised learning) tabanlı yöntemler.....	37
5.2.2 Yarı Denetimli (Semi Supervised) öğrenme, Aktif öğrenme ve Zayıf (Weak) öğrenme yaklaşımları.....	39
5.2.3 Denetimsiz (Unsupervised) öğrenme	40
5.2.4 Biyomedikal alanda yapılmış olan varlık etiketli derlemeler.....	41
5.3 İlişki Çıkarımı İçin Yapılan Çalışmalar.....	42
5.3.1. Biyomedikal ilişki sınıflandırmada mevcut derlemeler	43
5.4 Adlandırılmış Varlık Tanıma ve İlişki Çıkarımı İçin Son Teknoloji Yöntemler: Derin Sinir Ağı ve Dönüştürücü (Transformer) Tabanlı Öğrenme Modelleri	44
5.5 Bitki ilaç etkileşimi ile ilgili Yapılan Çalışmalar	50
6. MATERYAL ve YÖNTEMLER.....	52
6.1 Bitki -İlaç Etkileşimleri İçin Veri Seti.....	54

İÇİNDEKİLER (devam)

Sayfa

6.2 BİE İçin Varlık Etiketli Altın Standart Derlem ve AVT Modeli Geliştirme Süreçleri	57
6.2.1 BİE varlıkları için altın standart (gold standart) derlem geliştirme süreci...	58
6.2.1.1 Uzman kişi belirleme	59
6.2.1.2 Rehber hazırlama ve kategori belirleme	60
6.2.1.3 Etiketleme aracını belirleme	64
6.2.1.4 Etiketleme işleminin başlaması.....	65
6.2.1.5 Varlık tanıma için altın standart derlem.....	69
6.2.2 AVT (NER) modeli oluşturma süreci	69
6.2.2.1 Tok2Vec tabanlı AVT modelinin kurulumu	70
6.2.2.2 Transformer (RoBERTa) tabanlı AVT modelinin kurulumu	71
6.2.2.3 Değerlendirme.....	72
6.3 İlişki Etiketli Altın Standart Derlem ve İlişki Çıkarım Modeli Geliştirme Süreçleri	73
6.3.1 İlişki tespiti etiketli altın standart derlem etiketleme süreci	73
6.3.2 İlişki türü etiketli altın standart derlemlerin geliştirilme süreci.....	77
6.4. İlişkilerin Çıkarım Modelinin Geliştirilme Süreci	78
6.4.1 SVM, KNN ve Random Forest ve TF-IDF ile model oluşturma.....	78
6.4.2 BERT ve BioBERT modelleri ile model oluşturma	78

İÇİNDEKİLER (devam)

Sayfa

6.4.3 İlişki Çıkarım modellerinin değerlendirilmesi.....	79
6.5 Zayıf Denetimli Öğrenme (SNORKEL) Modeli İle İlişki Çıkarımı	79
7.DENEYSEL BULGULAR.....	85
7.1 MSKCC Veri Kaynağından Elde Edilen Sonuçlar	85
7.2 BİE için Varlık Etiketli Altın Standart Derlemde Etiketlenen Varlık Sayıları ve Uyum Skoru Sonuçları	86
7.3 AVT Model Sonuçları	87
7.4 BİE için İlişki Etiketli Altın Standart Derlem	88
7.5 BİE için Tespit İlişki Etiketli Model Performansları	89
7.6 BİE için İlişki Türü Etiketli Altın Standart Derlem.....	90
7.7 BİE için Tür İlişkisi etiketli Altın Standart Derlem	90
7.8 Zayıf Denetimli Öğrenme (SNORKEL) Sonuçları	91
8.TARTIŞMA.....	92
8.1 Mskcc Veri seti	92
8.2 Varlık Etiketli BİE Derlemi Üzerine Değerlendirme ve Hata Analizi	93
8.3 AVT Modeli İçin Bir Değerlendirme ve Hata Analizi	94
8.4 İlişki Etiketleme Derlemi.....	95
8.5 İlişki Çıkarım Modeli	97

İÇİNDEKİLER (devam)

	<u>Sayfa</u>
8.6 Zayıf Denetimli Öğrenme (SNORKEL) Modeli ve Hata Analizi	97
9. SONUÇ VE ÖNERİLER	98
KAYNAKLAR DİZİNİ	99
TEŞEKKÜR	108
ÖZGEÇMİŞ	109

ŞEKİLLER DİZİNİ

<u>Şekil</u>	<u>Sayfa</u>
1.1. Bitki dünyası hakkında genel bilgiler	2
2.1. Reçeteli İlaçların etkileşime girdiği bazı maddeler	10
3.1. Sanayi devrimi ve veri arasındaki ilişki.....	15
3.2. Yapay zeka, makina öğrenmesi, derin öğrenme ve doğal dil işleme arasındaki ilişkilere bir bakış	17
3.3. Üst seviye doğal dil işleme uygulamaları.....	18
3.4. Yapılandırılmış, yarı yapılandırılmış ve yapılandırılmamış veriler ve farklarını bir gösterimi.....	19
4.1. İlişki çıkarımı (Relationship extraction) hakkında genel bir örnek	28
4.2. Mekanik Türk'ün rakibiyle oynarken operatörün içeride nasıl oturduğunu düşündüğünü gösteren bir kesiti.	31
5.1. Aktif Öğrenme Döngüsü.....	40
5.2. Genel bir Bert mimarisi	47
5.3. Adlandırılmış varlık etiketleme için ince ayarlanmış (fine tuned) BERT modelinin bir gösterimi.....	48
6.1. Bitki ilaç etkileşimlerinin biyomedikal literatürden çıkarılması için oluşturulan boru hattı (pipeline)	52
6.2. MSKCC verisetinden bitkileri, ilaçları, bitki-ilac etkileşim bilgisini ve PubMed referanslarını ve bu referanslara ait PubMed özetlerinin bir araya getirildiği veriseti. Bu veri setinde yer alan bitki ve ilaç varlıkları bilgi tabanları ile eşleştirilmesi	56

ŞEKİLLER DİZİNİ (devam)

<u>Şekil</u>	<u>Sayfa</u>
6.3. Aktif öğrenme destekli altın standart derlem oluşturma protokolü	59
6.4. Bitki-ilaç etkileşim mekanizmasına göre 4 kategoriden oluşan varlık etiketleme kategorileri ve bazı örnek veriler.....	60
6.5. Prodigy etiketleme aracı çalıştırma komutları	64
6.6. Prodigy etiketleme aracının bir görseli	65
6.7. BİE etkileşimi ile ilgili varlıkların Prodigy etiketleme aracı ile başlatılması	66
6.8. Yaygın olarak kullanılan bir ve birden fazla kelimeden oluşan bitkilerin etiketlenmesi	67
6.9. RoBERTa Dil modeli ince ayarlanarak (fine tune) biyomedikal literatürde yer alan bitki ve ilaç terimlerinin otomatik olarak etiketlenmesi.....	69
6.10. İlişki etiketleme protokolü	74
6.11. Prodigy etiketleme aracı ile ilişki var-yok seviyesinde etiketleme işleminin bir gösterimi	76
6.12. Zayıf denetimli öğrenme çerçevesi SNORKEL modeli ile ilişki çıkarımı .	81
7.1. İlaçlarla en çok etkileşime giren bitki terimleri	85
7.2. Bitkiler ile en çok etkileşime giren ilaç ve sınıfları	86
8.1. BİE derleminde etiketlenen varlıkların denge durumu	93
8.2. POSITIVE, UNCERTAIN ve NEGATIVE ilişkilerin denge durumunun bir gösterimi.....	96

ŞEKİLLER DİZİNİ (devam)

Şekil

Sayfa

8.3. İlişki Türü Etiketli derlemin denge durumu96





ÇİZELGELER DİZİNİ

<u>Çizelge</u>	<u>Sayfa</u>
3.1. Yapılandırılmış, yapılandırılmamış ve yapılandırılmamış medikal kayıtlarla ilgili bir örnek	21
4.1.Bir belge içinde adlandırılmış varlık örnekleri	23
4.2.Biomedikal doğal dil İşleme alanında adlandırılmış varlık örneklerinin bir gösterimi	27
4.3 Kohen kappa katsayısının yorumlanması	32
5.1. Biyomedikal varlıkların hangi sözlüklere ait olduklarını gösteren bir örnek gösterim	34
5.2.Kural tabanlı yöntemlerin avantaj ve dezavantajları	36
5.3. Gen, Protein, Tür, Hastalık, Kimyasallar, Anatomi, ve Hücre varlıklarının etiketlendiği derlemler ve sayıları	41
5.4. Biyomedikal ilişki sınıflandırmada mevcut derlemler	43
6.1. Alan uzmanları tarafından filtrelenen varlıklar	56
6.2. Bilimsel makalelerde sıklıkla geçen bitki ve bitki bileşenlerine ait örneklerin gösterimi	62
6.3. Bir cümlede geçen aday bitki ve ilaçlar arasında meydana gelen etkileşimler	75
6.4. POSITIVE ve ABSTAIN için belirlenen kurallardan bazıları	82
6.5. NEGATIVE ve ABSTAIN için belirlenen kurallardan bazıları.....	83
7.1. BİE altın standart derleminde etiketlenen varlık sayısı	86

ÇİZELGELER DİZİNİ (devam)

<u>Çizelge</u>	<u>Sayfa</u>
7.2. Varlık etiketli altın standart derlem oluşturma sürecinde küratörler arası uyum oranı	87
7.3. Tok2vec modelinin AVT sonuçları.....	88
7.4. RoBERTa modelinin AVT sonuçları.....	88
7.5. BİE altın standart ilişki tespit derlemde etiketlenen varlık sayısı.....	89
7.6. İlişki etiketli model performansları.....	89
7.7. BİE altın standart derleminde etiketlenen varlık sayısı	90
7.8. Tür ilişkisi etiketli model performansı.....	90
7.9. SNORKEL modellerinin performansı	91
8.1. PubMed Sorgusu sonucu çıkan bitki-ilaç etkileşim ile ilgili çalışmaların bir listesinin gösterimi.....	89



SİMGELER VE KISALTMALAR DİZİNİ

<u>Kısaltmalar</u>	<u>Açıklama</u>
MSKCC	Memorial Sloan Kettering Cancer Center
BİE	Bitki İlaç Etkileşimi
M.Ö	Milattan Önce
YZ	Yapay Zeka
MÖ	Makina Öğrenmesi
DÖ	Derin Öğrenme
ADMET	Absorption, Distribution, Metabolism, Elimination, Toxicity –Emilim, Dağılım, Metabolizma, Eliminasyon, Toksisite
NLP	Natural Language Processing
NER	Name Entity Recognition
RE	Relation Extraction
DDİ	Doğal Dil İşleme
AVT	Adlandırılmış Varlık Tanıma
İÇ	İlişki Çıkarma
İS	İlişki Sınıflandırma
BERT	Bidirectional Encoder Representations from Transformers
BioBERT	Biomedical Bidirectional Encoder Representations
RoBERTa	Robustly Optimized BERT

SİMGELER VE KISALTMALAR DİZİNİ

<u>Kısaltmalar</u>	<u>Açıklama</u>
CNN	Convolutional Neural Network-Evrişimli Sinir Ağı
KNN	k Nearest Neighbor-k En yakın komşu
SVM	Support Vector Machine-Destek Vektör Makinası
DSÖ	Dünya Sağlık Örgütü
ABD	Amerika Birleşik Devletleri

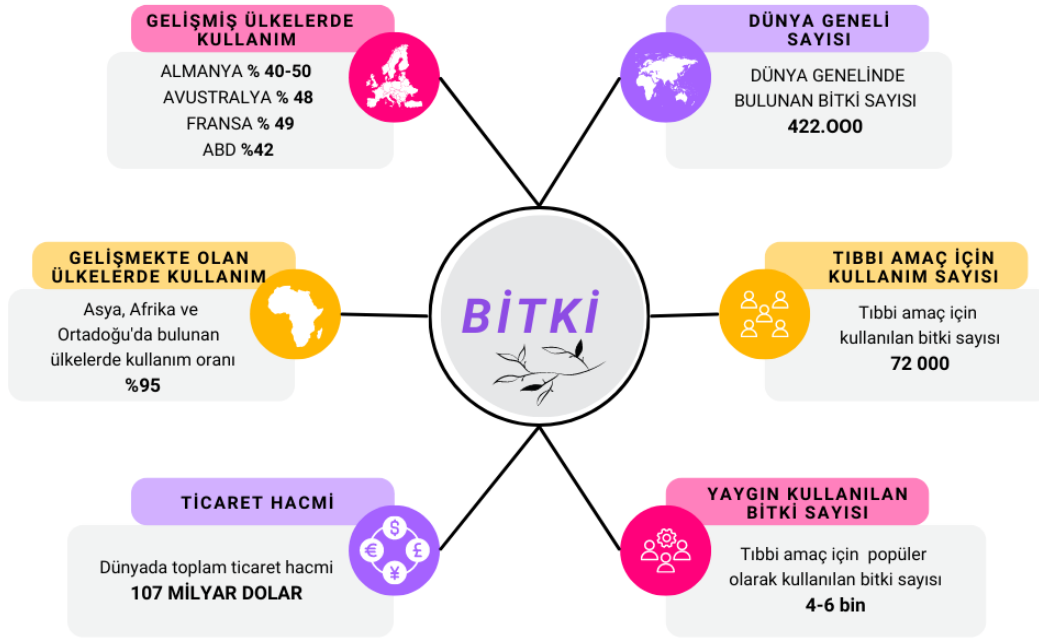


1. GİRİŞ

Hastalıkların tedavisi ve önlenmesi amacıyla bitkisel ilaçların kullanılması binlerce yıldır bilinen en eski tedavi yöntemlerinden birisidir ve halen çeşitli modern tıp sistemlerinin en önemli bileşenleri olarak dünya çapında tedavi edici olarak kullanılmaktadırlar (Health Organization, 2007). Özellikle obezite, hipertansiyon, ağrı sendromları, anksiyete, kilo verme ve sağlıklı olma isteği bitkisel ilaçların sıklıkla kullanımının başlıca nedenleridir (Tachjian ve ark. , 2010). Dünya Sağlık Örgütü (DSÖ) günümüzde dünya nüfusunun yaklaşık %80'ninin (yaklaşık 6 milyar kişinin) bitki ve bitkisel ürünlerden faydalandığını bildirilmektedir. Gelişmiş ülkelerde bu kullanım oranı Almanya'da % 40-50, Avustralya'da % 48, Fransa'da % 49 ve Amerika Birleşik Devletleri'nde (ABD) %42 seviyesinde iken Asya, Afrika ve Orta Doğu'da bulunan gelişmekte olan ülkelerde ise bu oran % 95 gibi oldukça üst seviyelerdedir (Titz, 2004). Bitki ve bitkisel ürün kullanım çeşitliliği ve ticareti de son yıllarda oldukça yaygınlaşmıştır. 2006 yılında yapılan bir çalışmada dünya çapında mevcut 422.000 bitki türünün var olduğu ve bunlardan 72000 tanesinin tıbbi amaç için kullanıldığı belirtilmiştir. Ancak yaygın olarak kullanılan bitki sayısının ise 4-6 bin arasında olduğu ve ticari amaç için bunlardan yaklaşık 3 bin tanesinin kullanıldığı belirtilmiştir (Acibuca and Budak, 2018; Schippmann et al., 2006). 2017 yılında yapılan bir araştırmaya göre de dünya genelinde bitki ve bitkisel ürün ticaretin yıllık yaklaşık 107 milyar dolar gibi oldukça büyük bir rakam olduğu bildirilmiştir (Xu *et al.*, 2019). Bu bilgilere ait detaylı bir özet Şekil 1.1'de görülmektedir.

Bitki ve bitkisel ürünlerin kullanım oranlarının dünya genelinde yüksek olmasının başlıca sebepleri arasında bitkilerin organik kaynaklı olduğu ve hiçbir yan etki ve zararlarının olmadığı kanısı yatmaktadır. Bu nedenle hastalar bitki ve bitkisel ürünleri sıklıkla tedavinin daha da başarılı olabilmesi amacıyla reçeteli ilaçlar ile beraber kullanmaktadırlar. Ancak reçeteli ilaçlar genel olarak Şekil 2.1'de görüldüğü üzere başka ilaçlarla (ilaç-ilaç etkileşimi), yiyeceklerle (ilaç-gıda etkileşimi), kişide bulunan hastalıklarla (ilaç-hastalık) ve ayrıca bitkiler (ilaç-bitki etkileşimi) ile etkileşime girerek tedavinin başarısız olmasına ve hatta yaşamı tehdit edici potansiyel olumsuz etkilere neden olmaktadır (Bent *et al.*, 2005). Örneğin yapılan bir çalışmada özellikle kanser hastalarının yüzde 40'nın (Akhtar, 2018) hiçbir yan etkisi olmadığı kanısı nedeniyle bitki ve bitkisel ürünleri doktorlarına danışmadan reçeteli ilaçlar ile birlikte kullandıkları ve bu nedenle tedavinin başarısız olmasına neden oldukları belirtilmiştir. Bazı yaygın kullanılan bitkilerden;

- Bilişsel işlevi iyileştirdiği için sıkça kullanılan **Ginkgo biloba**'nın antikoagülanlar ve antiplatelet ilaçlarla etkileşime girdiği ve spontan bir kanamaya neden olduğu
- Fiziksel ve zihinsel etkileri nedeniyle sıkça kullanılan **Ginseng**'in, **varfarin** ilacına verilen yanıtın azalmasına neden olduğu
- Sarı kantaronun (St. John's wort) ise irinotekan, dosetaksel ve imatinib ilaçları ile etkileşime girerek plazma seviyelerini düşürdüğü bildirilmiştir (Asimwe et al., 2021).



Şekil 1.1. Bitki dünyası hakkında genel bilgiler

Bitkisel ürün pazarında dünya genelinde oldukça popüler olduğu göz önüne alındığında bitki ve ilaç etkileşimlerinden kaynaklı riskleri belirlemek hem bilim dünyasının hem de ilaç üreten ticari firmaların dikkatini çekmiştir. Reçeteli ilaçların özellikle ilaç-ilaç (Lim ve ark. 2018), ilaç-gıda (Kim ve ark. 2022), ilaç-hedef (Aldahdooh ve ark. 2022) ve ilaç-hastalık (Lim ve ark. Shengjergji ve ark. 2021) gibi etkileşimlerinin tahmini üzerine birçok çalışma yapılmıştır. Ancak ilaç-bitki etkileşimi üzerine ise sınırlı sayıda çalışma yapılmıştır. Bunun en büyük nedenlerinden biri, bu araştırmaları sonuçlandırabilecek herhangi bir kapsamlı veri tabanının henüz bulunmaması ve ayrıca standart bir metodolojinin olmamasıdır. Çünkü insan sağlığı ile ilgili kapsamlı bir veri tabanını oluşturma süreci;

- zaman alıcı
- maliyetli

➤ uzman bilgisi gerektiren

zahmetli bir süreçtir. İlaç-bitki etkileşimlerinin varlığı ile ilgili en kapsamlı mevcut çalışmalar genel olarak milyonlarca makalenin bulunduğu PubMed gibi literatür kaynaklarında yapılandırılmamış düz metin şeklinde bulunmakta ve kısmen de MSKCC (Memorial Sloan Cancer Center, 2022) ve DrugBank (Drugbank, 2022) gibi veri kaynaklarını referans almaktadır. Ancak DrugBank veri tabanında bulunan etkileşim bilgileri daha çok tavsiye niteliğinde (örn. warfarin ile beraber alkol kullanmayınız) referanssız bilgilerdir. Bu nedenle kapsamlı bir bitki-ilaç etkileşim veri tabanının varlığı ilerde oluşabilecek olumsuz etkileşimlerin tahmini ve bu konularda yapılması düşünülen araştırma problemlerine büyük bir yarar sunacaktır. Bunu sağlamak için gün geçtikçe büyüyen biyomedikal literatürden bitki ve ilaç varlıklarını (terimlerini) otomatik olarak tanımlayacak ve bu varlıklar arasında farmakokinetik ve farmakodinamik ilişki türlerini çıkaracak bir metodoloji gereklidir. Biyomedikal ilişki çıkarımı alanında veri tabanı bazlı, ontoloji bazlı ve metin madenciliği bazlı yöntemler sıklıkla kullanılmaktadır. Ancak karmaşık biyomedikal metinlerden anlamlı ilişkilerin çıkarılması hususunda bu yöntemler yetersiz kalmaktadır (Trinh, Pham and Le, 2018).

Bu bağlamda bu tez çalışması milyonlarca makalenin yer aldığı biyomedikal literatürden ilaç ve bitki varlıklarının birlikte geçtiği cümleleri ve dokümanları otomatik olarak yakalayabilen ve bu varlıklar arasında, ilişki olup olmadığını ve ilişkinin tipini ortaya çıkartan makina öğrenmesi ve derin öğrenme tabanlı adlandırılmış varlık tanıma (Name Entity Recognition -NER) ve ilişki çıkarım (Relationship Extraction-RE) sistemleri (boru hatı, pipeline) tasarlamaya odaklanmıştır. Ancak makina ve derin öğrenme tabanlı NER ve ilişki çıkarımı modelleri etiketli veri setleri gerektiren maliyetli sistemlerdir ve literatür incelendiğinde makina ve derin öğrenme yöntemleri için varlık tanıma ve ilişki etiketleme düzeyinde etiketlenmiş kapsamlı bir bitki-ilaç etkileşim derlemi bulunmamaktadır. Bu nedenle bu tez çalışmasının genel amacı literatürün kapsamlı ve büyük eksikliklerden biri olan bitki-ilaç etkileşim veri setinin tasarımı için alan uzmanı küratörlerle çalışarak biyomedikal literatürden elde edilen yeterli derecede ilgili kaynağı varlık düzeyinde ve ilişki düzeyinde etiketleyerek (anote etme işlemi) makina öğrenmesi ve derin öğrenme yöntemleri için gerekli olan altın standart derlemler oluşturmaktır.

Çalışmanın literatüre temel katkıları 3 bölümde anlatılmıştır ve her bölümde kendi içinde bir kaç aşamadan oluşmuştur. Buna göre;

1- Onaylanmış bitki-ilaç etkileşim verilerinin ve makale özetlerinin çeşitli kaynaklardan toplanarak bir araya getirildiği bir veri seti literatüre kazandırılmıştır.

- Bitki ve ilaç etkileşimi üzerine yapılan biyomedikal çalışmaları bir araya getirmek için MSKCC (2022) ve PubMed (2022) veri kaynakları belirlenmiştir. MSKCC veri kaynağı genel olarak PubMed veri kaynağında bulunan makalelerden manuel olarak çıkarılan toplam 279 adet bitkinin, bitkisel ürünün ve vitaminin ilaçlar ile olan etkileşim bilgisinin referansları ile birlikte yer aldığı yarı-yapılandırılmış bir kaynaktır.
- MSKCC veri kaynağında bulunan yarı-yapılandırılmış bitki-ilaç etkileşim (BİE) verileri web kazıma araçları ve metin ayrıştırma (text parsing) teknikleri kullanılarak tablo formatında yapılandırılmış bir veri seti haline getirilmiştir. Bu değerli bilgilerin tablo haline getirilmesi, bitki-ilaç etkileşimi ile ilgili yapılacak olan çeşitli analizlerin yapılmasını kolaylaştırmakta ve literatüre fayda sağlayacaktır.
- MSKCC kaynağının referans verdiği BİE ile ilgili PubMed makale özetleri bir api aracılığıyla çıkartılarak tabloya ayrı bir sütun olarak eklenmiştir. Böylece BİE'leri için zenginleştirilmiş bir veri seti oluşturulmuştur.
- Tabloda yer alan bitki ve ilaç varlıklarından bitkiler FooDB (2022) ve ilaçlar DrugBank (2022) bilgi tabanları ile eşleştirerek varlıkların bilgi tabanı seviyesinde analizi yapılmıştır. Varlık bağlama (entity linking) olarak da adlandırılan bu yöntem genel olarak varlıklar arasındaki belirsizlik (disambiguation) durumunu kaldırma amacıyla kullanılmaktadır. Bu çalışmada bu analiz ile bilgi tabanlarının bitki ve ilaç varlıklarının ham veri ile eşleşmesi test edilmiştir ve çıkan sonuç bitki ve ilaç varlıklarını otomatik yakalayabilecek bir adlandırılmış varlık tanıma (NER) modeli ihtiyacını doğrulamıştır.
- Son olarak MSKCC uzmanları tarafından manuel olarak çıkarılmış potansiyel etkileşime sahip bitki varlığı, ilaç varlığı, özetlenmiş BİE verileri, ilgili bilgilerin referansı, bu referanslardan elde edilen PubMed makale özetleri ve MSKCC'den elde edilen potansiyel bitki ve ilaç varlıkların eşleştiği bilgi tabanının ID'lerinin bulunduğu yapılandırılmış bir

BİE veri setinin literatüre kazandırılması sağlanmıştır. Tablonun genel bir tasarımı aşağıdaki gibidir.

[[bitki varlığı], [ilaç varlığı], [MSKCC uzmanları tarafından PubMed'ten alınan
özet bitki-ilaç etkileşim bilgileri],
[PubMed makale özeti],[FoodDB ID,Drugbank ID]]

2- Makina öğrenmesi algoritmaları için gerekli olan eğitim verilerini sağlayacak etiketli altın standart derlemler oluşturularak bu derlemler literatüre kazandırılmıştır.

- MSKCC veri kaynağında referans verilen BİE ile ilgili PubMed makale özetleri ve ayrıca PubMed sorgusu yapılarak elde edilen makale özetleri içinden bitki ve ilaç terimleri (varlıkları) alan uzmanları seviyesinde hazırlanan bir protokole göre etiketlenerek varlık etiketli altın standart bir derlem literatüre kazandırılmıştır.
- Bitki ve ilaç varlıkları arasındaki ilişkiler Emilim, Dağılım, Metabolizma, Eliminasyon ve Toksisite olarak adlandırılan ADMET ve Potansiyel, Sinerjistik ve Katkısal olarak kategorize edilerek ilişki seviyesinde ilişki var yok derlemleri alan uzmanları tarafından etiketlenerek ilişki seviyesinde altın standart derlemler literatüre kazandırılmıştır.

3- Adlandırılmış varlık tanıma ve ilişki çıkarımı sistemleri için makina öğrenmesi tabanlı modeller geliştirilmiştir.

- Varlık etiketli altın standart derlem tabanlı sistemle biyomedikal literatürde bulunan ilaç ve bitki varlıklarını otomatik yakalayabilen aktif öğrenme destekli makina ve derin öğrenme bazlı bir model geliştirilmiştir.
- Varlıklar arasındaki ilişki etiketli derlem tabanlı sistemle biyomedikal literatürde bulunan ilaç ve bitki varlıkları arasındaki ilişkileri otomatik yakalayabilen makina ve derin öğrenme bazlı bir model geliştirilmiştir.
- Bitki ve ilaç arasındaki ilişkileri yakalayabilmek için iki varlık arasında ilişki belirten anahtar kelimelerin ve kuralların çıkarılarak makina öğrenmesi ve kural tabanlı yöntemlerin birleştirilmesi ile oluşan bir sistemin geliştirilmesi sağlanmıştır.

Bu tez çalışmasının akış şemasına değinilirse tezin ikinci bölümünde bitkilerden ve bitki-ilaç etkileşim mekanizmalarından bahsedilmiştir. Üçüncü bölümde Doğal dil işleme (DDİ) nedir ve biyomedikal verilerin yapısı üzerinde durulmuştur. Dördüncü bölümde tezin ana konusu olan adlandırılmış varlık tanıma (AVT, NER) ve ilişki çıkarım (RE) yöntemlerinden bahsedilmiştir. Dördüncü bölümde önceki çalışmalar, beşinci bölümde materyal metod, altıncı bölümde sonuç, yedinci bölümde tartışma, sekizinci bölümde sonuç ve dokuzuncu bölümde ilgili referanslar yer almaktadır.



2. BİTKİLER ÜZERİNE GENEL BİR BAKIŞ

2.1. Geçmişten Günümüze Bitkisel İlaçların Kullanım Tarihi

Çok eski tarihlerden beri insanlık, ağrıları gidermek ve hastalıkları iyileştirmek için şifalı bitkilerden faydalanmıştır. Hastalıkları tedavi etmek için şifalı bitkilerin kullanımına dair en eski kanıtlar Kuzey Irak'ta yapılan çeşitli arkeolojik mezar kazılardan elde edilen 60.000 yıl öncesine dair kayıtlara dayanmaktadır. Bu mezarda aralarında Efedra bitkisinin de bulunduğu en az 7 adet şifalı bitki mezarda yatan kişi ile çevrili bir halde bulunmuştur (Giannenas *et al.*, 2020). Bir diğer tarih öncesi kayıt ise kutsal bir kitap olarak düşünülen Avesta'nın (M.Ö 1000-500) kayıtlarında M.Ö 6500'lere dayanan İran Aryan uygarlığı sırasında çeşitli bitkilerin şifalı ot olarak kullanıldığı ve “**haoma**” olarak adlandırılan Efedra bitkisinin kutsal sayıldığıdır (Jamshidi-Kia, Lorigooini and Amini-Khoei, 2018). Ancak yazılı anlamda tıbbi bitkilerin kullanımını doğrulayan kanıt ise yakın zamanda Mezopotamya'da bulunan 5000 yıllık bir Sümer kil levhasının keşfi olmuştur. Bu levhada toplam 250 adet bitkinin tedavi amacıyla kullandığı saptanmıştır (Sewell and Rafieian-Kopaei, 2014). Bitkilerin şifalı ot olarak kullanılmaları ile ilgili ilk kapsamlı çalışma ise M.Ö 2700'lü yıllarda Çin imparatoru Shen Nong Ben Jing'in yazdığı tahmin edilen “Büyük Bitki” olarak çevrilen “Pen Ts'ao” adlı Çin kitabıdır. Bu kitapta sarı centiyan, ginseng, tarçın, jimson otu ve efedra dahil olmak üzere çeşitli hastalıkların tedavisinde kullanılan 365 bitkisel ilaç tarif edilmiştir. Bitki biliminin buradan aşamalı olarak ilk önce Hindistan, Mezopotamya, Pers ve Mısır'a yayıldığı daha sonra oralardan da Doğu Akdeniz'de Yunanlara ve son olarak Yunanlılardan tüm Avrupaya yayıldığı düşünülmektedir (Jones, 1996).

Eski Mısırdaki hastalıkların tedavisi için tıp ve sihir bir arada kullanılmıştır. Mısır tıbbı hakkındaki bilgiler papirüsler aracılığıyla öğrenilmektedir. Papirüsler otsu bir bitkiden oluşmuştur ve üzerlerine harf, yazı ve işaretlerin kolay bir şekilde yazılması sayesinde Mısırlılar kültür ve uygarlık yaşamında büyük bir gelişme göstermişlerdir. Çeşitli keşifler neticesinde bulunan papirüslerden Kahun papirüsü (M.Ö 2000) ve Smith (M.Ö 1600) papirüsü Mısırlıların hastalık tedavisinde çeşitli cerrahi ve sihir yöntemlerini kullandıklarına değinmiştir (Ceran, 2008). Ancak bitkisel ilaçların kullanımına dair ders kitabı seviyesinde yazılmış olan en etkili papirüs ise 1862 yılında Thebes'te bir mezarın içerisinde bulunan M.Ö 1553-1550 yılları arasında yazıldığı tahmin edilen Ebers papirüsüdür (Hayirlidağ, 2021). Buna göre Mısırlılar, *Mentha x piperita* L. (nane), *Sinapis nigra* L. (siyah hardal), *Cassia*

acutifolia L. (sinameki), Papaver somniferum L. (haşhaş), Scilla maritima L. (ada soğanı), Datura stramonium L. (tatula) gibi tıbbi bitkileri hastalık tedavilerinde kullandıkları görülmüştür (Kaya, 2010).

Bitkisel ilaç kullanımı Yunan ve Roma imparatorluklarında ise hem terapötik hem de nefret dolu (Sokratesin baldıran zehiriyle idam edilmesi) amaçlarla kullanılmıştır. Yunanistan'ın Kos adasında yaşayan Hipokrat (MÖ 466-377) hastalığı bir büyü ve mistik bir olay olarak görmemiş bunu doğal bir fenomen olarak görmüştür ve bu mistik olayı bitkiler ve çeşitli ilaçlar kullanarak ortadan kaldırmıştır. Yaklaşık 400 çeşit bitki bazlı ilaç kullanarak hastalığın doğal seyrini, tedavisini ve sonuçlarını incelemiştir. Bu tıbbi ve bitkisel bilgiler İskenderiye Tıp Okulunun kurulmasına yol açmıştır. Yazılan bir diğer ilk şifalı bitkiler arasında MÖ 2. yüzyılda Nikander tarafından yazılan “zehirler ve panzehirleri” eseridir. Pontus Kralı Eupator Mithridates (MÖ 132-63) tarafından bu eser güncellenmiş ve belgelenmiştir. Eupatorium purpureum (çakıl kökü veya Jo Pye) otu ile adı bazı şifalı bitkilerde hatırlanır. İlk gerçek tıbbi botanikçi ise MS birinci yüzyılda Roma İmparatorluğu zamanında yaşayan Tarsus’lu Dioskorides’tir ve onun ünlü kitabı *Materia Medica*’sı (Müdavi İlaçlar), sonraki 1500 yıl boyunca bitkiler için model olarak kalmıştır (Jones, 1996).

İslam uygarlığı döneminde ise dönemin meşhur astronom, filozof ve yazarı olan meşhur İbn-i Sina (Avicenna) (980-1037) “Şifa” ve “El-Kanun fi’l-Tıb” (Tıbbın Kanunu) adında iki eser üretmiştir. Bu eserlerde 900’den fazla tıbbi bitki, hayvansal ve inorganik kökenli ilaç yer almaktadır. Son olarak bitkisel ilaç literatürü Amerika kıtasının keşfi ile daha da zenginleşmiştir. Bu kıtada keşfedilen bitkilere örnek olarak *Erytroxylon coca* Lam. (koka), *Cinchona succubra* Pav. (kinin ağacı), *Theobroma cacao* L. (kakao ağacı), *Hydrastis canadensis* L. (hidrastis), *Polygala senega* L. (senega) verilebilir (Kaya, 2010).

Ondokuz ve yirminci yüzyılı ise kimya ve biyokimya bilimlerindeki gelişmeler neticesinde ilaç sanayisinin arttığı yıllar olmuştur. Antibiyotik ve birçok hastalığa yanıt veren laboratuvar kaynaklı ilaçlar keşfedilmiş ve oldukça popüler olmuşlardır. Buna rağmen bitki kökenli ilaçlarda ilaç sanayisinin her zaman önemli bir bel kemiğini oluşturmuştur. Yapılan bir çalışmada mevcut ilaçların 1/4’ünün bitkisel kökenli olduğu bildirilmiştir (Şarışen and Çalışkan, 2005). Çok bilinen bitki kaynaklı ilaçlardan bazılarını aşağıdaki gibidir (Giannenas *et al.*, 2020);

➤ aspirin (söğüt kabuğundan)

- digoksin (yüksük otundan)
- kinin (kınakına kabuğundan)
- morfin (afyon haşhaşından)

Günümüzde ise bitkisel ürün pazarı, 1960'lardan itibaren oldukça büyümeye başlamış ve bu hız son yıllarda da devam etmektedir. Özellikle reçeteli ilaçlarda meydana gelebilen ciddi yan etkilerden ve ekonomik sorunlardan kaynaklı düşünceler ve ayrıca doğanın yan etkisi olmadığı kanısı bitkisel tedaviyi yeniden popüler bir duruma getirmiştir. 1974'te, DSÖ geliştirmekte olan ülkeleri reçeteli ilaçlarla karşılanmayan ihtiyaçlarını karşılamak için geleneksel (traditional) bitkisel ilaçlarla desteklemeye teşvik etmiştir (Winslow ve Kroll, 1998). Bitkisel ürün satışları 2012 yılında yalnız ABD'de yaklaşık 5,6 milyar dolara ulaşmış ve giderek artmaya devam etmektedir (Lindstrom ve ark., 2013).

2.2. Bitkisel İlaç Kullanım Nedenleri

2.2.1 Erişilebilirlik ve satın alınabilirlik

Bitkisel ilaçların kullanım nedenlerine dair yapılan bir araştırmada özellikle Hindistan gibi aşırı nüfuslu ülkelere kırsal nüfusun modern tıba erişiminin zor olduğu ve bu nedenle temel sağlık ihtiyaçları için bitkisel ilaçların kullanımının oldukça fazla olduğu bildirilmiştir. Bir başka neden ise bitkisel ilaçların reçeteli ilaçlara göre daha uygun fiyatlı olmasından kaynaklanmaktadır. Örneğin Afrika ülkelerinden Zambiya, Tanzanya ve Uganda gibi ülkelere kişi başına düşen bitkisel ilaç doktorlarının nüfusa oranı 1:200–1:400 oranında iken gelişmiş batılı ülkelere ise bu oran 1:20.000 civarında veya daha azdır (Z. Msomi and B.C. Simelane, 2019).

2.2.2 Sağlıkta alternatif bir yaklaşım

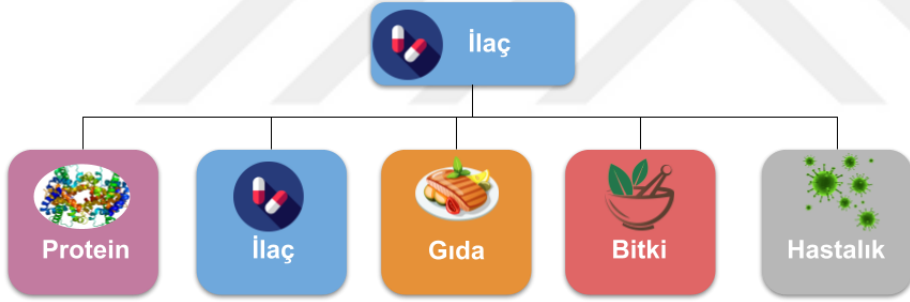
Gelişmiş ülkelere bitkilerin kullanım nedenleri araştırıldığında bitkilerin geleneksel biyosentetik ilaçlardan daha sağlıklı olduğu algısı yatmaktadır. Örneğin Amerika Birleşik Devletleri'nde yapılan bir araştırmada ciddi bir sağlık sorunu olan Amerikalıların yaklaşık yüzde 25'i bitki ve bitkisel ürünlerden faydalanmakta ve bu hastaların yalnızca yüzde 70'i doktorlarını bu tür bir kullanım hakkında haberdar ettiği gözlemlenmiştir (Cupp, 1999).

Yapılan bir araştırmada bitkisel ilaçlara olan ilginin yakın zamanda yeniden canlanması;

- bitkisel ürünlerin üretilen ürünlerden daha üstün olduğuna dair yanlış inanç
- geleneksel ilaçlardan elde edilen sonuçlardan duyulan memnuniyetsizlik ve geleneksel tedavilerin ve ilaçların etkisiz olduğu kanıtlanmış bazı hastalıkların tedavisinde bitkisel ilaçların etkili olabileceğine dair inanç
- bilim ve teknolojinin gelişmesiyle bitkisel ilaçların kalitesi, etkinliği ve güvenliği hakkındaki gelişmeler
- hastaların, doktorlarının sorunu gerektiği gibi tanımlamadığına dair inançları; bu nedenle bitkisel ilaçların başka bir seçenek olduğu hissi gibi çeşitli faktörlere bağlanmıştır (Ekor, 2014).

2.3. İlaç Etkileşim Tipleri

Etkileşimin temel konsepti: İki maddenin (bitkiler, ilaçlar, besinler, gıdalar) aynı anda veya kısa bir süre içinde beraber tüketilmesi sonucu her iki maddenin tek başına alındığında oluşabilecek etkilerinin toplamından daha farklı fizyolojik etkilerle sonuçlanması olayı etkileşim olarak tanımlanmaktadır (Rehman *et al.*, 2015).



Şekil 2.1. Reçeteli İlaçların etkileşime girdiği bazı maddeler

Bir başka deyişle, etkileşim bir maddenin başka maddenin aktivitesini artırdığı, azalttığı ya da hiçbirinin kendi başına üretemediği yeni bir etki durumudur (Bushra, Aslam and Khan, 2011). Oluşan bu etkileşimlerin bir kısmı hastaya faydalı olabilecek olumlu sinerjik etkileşimler iken bir kısmı ise ilaçların farmakokinetik (vücudun ilaca yaptığı etki) veya farmakodinamik (ilacın vücuda yaptığı etki) mekanizmasını etkilemekte ve insan sağlığı açısından ölüme dahi sebep olabilecek ciddi riskler oluşturmaktadır. Şekil 2.1’de en popüler ilaç etkileşimlerinden bazıları görülmektedir. Bunlardan bitki ilaç etkileşimleri (bkz bölüm 2.4) ayrı başlık altında açıklanırken diğer etkileşimler aşağıdaki gibidir;

- **İlaç-ilaç Etkileşimi:** İlaç-ilaç etkileşimleri bir ilacın birlikte alınan başka bir ilaç tarafından etkinliğinin değişmesi olayıdır. Bu etkiler istenen veya istenmeyen etkiler olabilir. İlaç-ilaç etkileşimlerinin nedenleri genel olarak farmakokinetik ve farmakodinamik olmak üzere iki gruba ayrılır. Farmakokinetik etkileşimler vücudun ilaca yaptığı etkilerden oluşmakta ve emilim, dağılım, metabolizma ve boşaltım (ADME) olarak sınıflandırılmıştır (Takeda *et al.*, 2017). Farmakodinamik etkileşimler ise ilacın vücuda yaptığı etkilerden oluşmakta ve genel olarak sinerjik, katkısız (additive) ve antagonistik olarak sınıflandırılmışlardır (Niu, Straubinger and Mager, 2019).
- **İlaç-hastalık etkileşimi:** İlaç-hastalık etkileşimi, bir hastalığı tedavi etmek için kullanılan bir farmakoterapinin başka bir hastalığın kötüleşmesine neden olduğu durumlardır. Bu durumlarda ya hasta ilacı kullanmayı bırakmak da ya da ilaç dozajları ayarlanarak tedavi devam etmektedir (van Tongeren *et al.*, 2020).
- **İlaç-hedef etkileşimi:** Küçük ilaç molekülleri genellikle birden fazla protein hedefine veya istenmeyen hedef dışı hedeflere bağlanır. İlaç karışıklığı (drug promiscuity) olarak adlandırılan bu olay genellikle istenmeyen veya açıklanamayan ilaç reaksiyonlarına yol açarak yan etkilere neden olmakta ya da ilacın yeniden konumlandırması (ilacın başka bir hastalık için kullanılması) gibi fırsatlara yol açmaktadır. Bu nedenle, potansiyel ilaç-hedef etkileşimlerini belirlemek farmakoloji için önemli bir konu olduğundan son zamanlarda geniş bir ilgi görmektedir (Wang *et al.*, 2013).
- **İlaç-gıda etkileşimi:** Klinik olarak ilgili gıda-ilaç etkileşimlerinin çoğu, ilacın biyoyararlanımında meydana gelen gıda kaynaklı değişikliklerden kaynaklanır. Bu nedenle yaygın olarak kullanılan bazı gıda, meyve ve alkol, tedavinin başarısız olmasına neden olmaktadır. Bazı gıdaların ilaçlar üzerindeki başlıca yan etkileri, yağlı, yüksek proteinli ve lifli diyetler tarafından emilimdeki değişiklikleri içerir (Bushra, Aslam and Khan, 2011).

2.4 İlaç-Bitki Etkileşimi

Bitkilerin ilaçlarla eş zamanlı kullanımı ile bir madde diğer maddenin etkilerini taklit edebilir, artırabilir veya azaltabilir, bu da klinik olarak önemli bitki-ilaç etkileşimlerine neden olabilir (Asiimwe et al., 2021). Olumsuz sonuçları en aza indirmek ve bu etkileşimleri bulabilmek ve bu etkileşimlerin potansiyel faydalarından yararlanmak için sistematik bir yaklaşım gereklidir ve bu mekanizmaları bilmek gerekir.

2.4.1 İlaç-bitki etkileşim mekanizması

Bitki ve bitkisel ürünlerin ilaçlarla olan etkileşiminin altında yatan mekanizma, ilaç-ilaç etkileşiminde olduğu gibi farmakokinetik ve farmakodinamik mekanizmalar ile ilişkili olduğu düşünülmektedir ancak bileşikler söz konusu olduğunda daha karmaşık bir mekanizmaya sahiptir. Başlıca olası potansiyel İBE'leri özetlersek (Ong and Pan, 2014):

- Gastrointestinal (Sindirim sistemi) emiliminde değişiklik
- Plazma proteinlerine bağlanma
- Taşıyıcı proteinlere bağlanma
- Reseptör düzeyinde farmakodinamik etkileşimler
- Metabolizmanın inhibisyonu
- Metabolizmanın indüksiyonu
- Aktif renal atılım için rekabet

2.4.1.1 Farmakokinetik ilaç-bitki etkileşimleri

Farmakokinetik etkileşimler vücudun ilaç ve bitkiye olan etkilerinden oluşmaktadır ve bu etkiler ilaç-ilaç etkileşimine benzer bir şekilde emilim, dağılım, metabolizma ve boşaltım ve toksisite (ADMET) olarak sınıflandırılmıştır. Genel olarak en çok rapor edilen BİE'ler farmakokinetik etkileşimlerdir ancak bu etkileşimleri tahmin etmek de farmakodinamik etkileşimlere göre çok daha zordur. Farmakokinetik etkileşimlerin altında yatan mekanizma karaciğer CYP (sitokrom P450) enzim ailesi, ilaç taşıyıcıları (drug transporter), akış proteinleri (efflux protein) ve P-glikoproteinlerin (P-gp) indüksiyonu ve inhibisyonudur. İlaçların çoğu CYP3A izoformu ile etkileşime girer (Oga *et al.*, 2016). Bitkiler bu yapıları inhibe ve indükleyerek bunların substratı olan ilaçların emilim, dağılım, metabolizma ve atılımlarını değiştirmektedirler. Farmakolojik etkiler şunlardır;

- **Emilim (Absorption) Etkileşimi:** Bitki ve ilaç etkileşimlerinin sindirim (gastrointestinal) sistemi seviyesinde olan etkileşim tipidir. Sindirim sistemi ortamını etkileyen bir bitki, eş zamanlı kullanılan bir ilacın emilim düzeyini değiştirir ve bir bitki-ilaç etkileşimine neden olur (Borse, Singh and Nivsarkar, 2019). Örneğin kabızlık tedavisi için kullanılan, Plantago (karnıyarık otu) bitkisi Warfarin, Karbamazepin ve lithium ilaçlarının emilimini azaltır ve bir ilaç bitki etkileşimi olur (Haddadian, Haddadian and Zahmatkash, 2014).
- **Dağılım (Distribution) Etkileşimi:** Bu etkileşimler yüksek plazma protein bağlama (higher plasma protein-binding) özelliğine sahip ilaçlarda, düşük dağılım hacmine (volume of distribution) sahip ilaçlarda ve dar terapötik pencereye (narrow therapeutic window) sahip ilaçlarda ortaya çıkabilir. Bu ilaca warfarin örnek verilebilir.

Özellikle warfarin ile etkileşime giren bitkisel ilaçlar, ilaç etkinliğini artırarak uzun süreli kanamaya veya ilaç etkinliğini azaltarak kan pıhtılaşma riskine neden olurlar ve ölümcül olabilecek riskler oluşmaktadır. Örneğin St John Wort bitkisi warfarin ilacının antikoagülan etkisini azaltırken Allium sativum (sarımsak) ise kanama riskinin artırır (Hodges and Kam, 2002).

- **Metabolizma (Metabolism) Etkileşimi:** Metabolizma, canlı organizmalar tarafından ksenobiyotiklerin biyokimyasal olarak modifikasyonudur. Metabolizma hızı, bir ilacın farmakolojik etkisinin süresini ve yoğunluğunu belirler. Bitkisel bileşenler, indüksiyon ve/veya inhibisyon yoluyla metabolize edici enzimleri değiştirebilir. Bununla birlikte, enzim(ler)in indüksiyonu, ilaç plazma düzeylerinin azalmasına (artan ilaç metabolizması yoluyla) ve ardından ilaç etkilerinin azalmasına yol açabilir.

Örneğin “*Sophora bitkisinin indinavir ile eş zamanlı kullanımı, plazma indinavir konsantrasyonlarında doza bağlı bir düşüşe neden olmuştur*”.

- **Eliminasyon Etkileşimi:** İlacın vücuttan atılımı ile ilgili etkileşim tipidir. İlaçlar idrar, dışkı, ter, gözyaşı, vb. kaynaklardan dışarıya atılırlar. Bu atılım süreci P-gp gibi hücre taşıyıcı protein/enzimler, OATP gibi yapılardır. Bu

yapılar bitki ve ilacın birlikte kullanılmasıyla etkilenebilir ve bitki-ilac etkileşimine neden olunur.

2.4.1.2 Farmakodinamik ilaç-bitki etkileşimleri

Farmakodinamik etkileşimler temel olarak enzim ve reseptör düzeyinde meydana gelen etkileşimlerden oluşur. Bu tür etkileşimler, antagonistik, katkı maddesi veya sinerjistik etkilerle sonuçlanabilir. Bu etkileşimlerden bazıları olumlu etkiler iken bazıları ciddi olumsuz etkilerdir.

1-Sinerjik ve Katkısal (Additive) Etkileşimler: Bir bitki, bir ilaç ile aynı etkiyi üretebilir ve birlikte eş zamanlı olarak alındıklarında ise bu etkiyi daha da artırabilirler. Bu nedenle bazı sakinleştirici bitkiler ve ilaçlar aynı amaç için eş zamanlı alındığında ilacın etkilerini daha da artırabilir. Bu etkiler katkısal ve sinerjik olarak değerlendirilebilir. Katkısal etki bir ilaç ve bir bitkinin toplamı kadar etkinin oluşması olayıdır.

$$\text{Katkısal (additive) etki: } 1+1=2 \quad (2.1)$$

Sinerjistik etki ise bir ilaç ve bitkinin toplamından daha büyük bir etkinin oluşması olayıdır.

$$\text{Sinerjik etki: } 1+1=3 \quad (2.2)$$

Örneğin agrimony bitkisi anti-diyabetik ilaçlarla birlikte alındığında ilave kan şekeri düşürücü (additive blood-glucose-lowering) etkiler üretilir. Varfarinin antikoagülan etkisi ginkgo, sarımsak ve zencefil bitkileri ile artar (Swanston-Flatt *et al.*, 1990) .

2-Antagonistik etkileşimler: Bir bitki, birlikte uygulanan bir ilacın istenilen etkisine aykırı bir etki üretebilir ve böylece ilacın etkisini azaltabilir.

$$\text{Antagonist etki: } 1-1=0 \quad (2.3)$$

Örneğin efedra veya kafein içeren otlar (kola fıstığı, guarana, mate, yeşil çay), ek kardiyovasküler etkiler üretmek için birçok bitkisel zayıflama ürünüde sıklıkla birlikte kullanılır ancak bu kullanım antihipertansif ilaçların etkilerini antagoneze edebilir (Scott and Elmer, 2002).

3. DOĞAL DİL İŞLEMeye GİRİŞ (DDİ)

M.Ö 1500’lü yıllarda yaşamış olan bir çiftçinin hayatı 2000 yıl sonra yani M.S 1500 yıllarda yaşamış olan bir çiftçinin hayatından muhtemelen çok farklı değildir. Çünkü insanlık tarihi 1800’lü yıllara kadar oldukça yavaş ilerlemiştir. Ancak 1800’lerde İngiltere’de sanayi devrimiyle başlayan keşifler, sonraki yıllarda yeni bulgulara yol açarak Homo sapiens’in dünyaya ilk ayak basmasından bu yana görülmemiş bir ölçekte buluşlara ve icatlara giden yolu belirlemiştir. Bu iki yüzyılın en önemli buluşlarından biri olan buharlı makinelerin icadı, sonrasında petrol ile çalışan yakıtlı motorların icadı ve neticesinde trenlerin ve uçakların icadı ile aylar süren kıtalar arası seyahatler bir günden daha kısa zamana inmiştir (Sullivan,1989). 1900’lerin başında antibiyotiğin keşfi ile binlerce yıldır tedavi edilemeyen veba gibi hastalıklar sona ermiş, elektriğin keşfi ile radyo, televizyon ve bilgisayarlar gibi teknolojik devrimler, radikal bir değişime ve günümüzde giderek artan bir hızla devam eden muazzam bir büyümeye yol açmıştır. Şekil 3.1 bu devrimsel yolculuğun tarihsel bir görselini göstermektedir.



Şekil 3.1. Sanayi devrimi ve veri arasındaki ilişki

Son yıllarda ise bilişim teknolojilerinde meydana gelen yüksek teknolojik gelişmeler paralelinde büyük miktarda bir verinin artmasına ve bu verileri işleme kapasitesinin artmasına olanak sağlamıştır. Özellikle dünyanın her yerindeki araştırmacılar tarafından muazzam miktarda biyomedikal veri üretilmektedir. Ancak biyomedikal verilerin büyük bir kısmı, yapılandırılmamış düz metin şeklinde biyomedikal literatürde (örneğin ACM, Springer, ScienceDirect, PubMed ve IEEE Xplore) bulunmaktadır (Zhang et al., 2021). Veri miktarındaki bu muazzam büyüme hızı nedeniyle yapılandırılmamış düz metin verilerinden anlamlı

bilgiyi çıkarabilmek oldukça zahmetli ve zor bir süreçtir. Bu verilerden potansiyel faydaların elde edilebilmesi için teknolojik gelişmelere ihtiyaç duyulmaktadır. Doğal dil işleme (NLP) yöntemleri ve araçları, araştırma makaleleri ve klinik notlar gibi yapılandırılmamış düz metinlerden biyomedikal kavramların çıkarılmasını kolaylaştırmakta ve bu metinlerden anlamlı ilişkiler çıkarmaktadır (Peng et al., 2020). Bu bölümde doğal dil işleme ve biyomedikal verilerin yapısını oluşturan kavramlar ele alınmıştır.

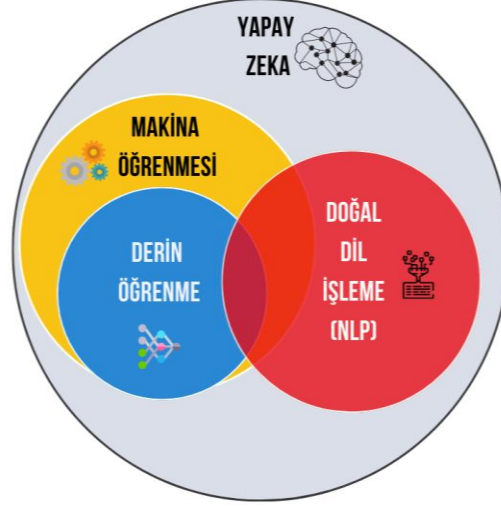
3.1 Doğal Dil İşleme Nedir

İnsan dili ya da bir başka deyişle doğal dil, karakter, kelime, cümle vb. gibi kurucu bileşenlerin karmaşık kombinasyonlarını içeren yapısal bir iletişim sistemidir (Vajjala *et al.*, 2020). Doğal dil işleme (DDİ, NLP) ise genel olarak insan dilini (doğal dil) bilgisayarların ve makinaların anlayabilmesi, yorumlayabilmesi ve analiz edebilmesi için otomatik veya yarı otomatik olarak işlenmesini sağlayan yapay zekanın bir alt dalıdır. Biyomedikal alanda ise yapılandırılmamış metinden anlamlı ve gizli bilgiyi çıkarma ve analiz etme işidir.

3.2 Yapay Zeka, Makina Öğrenmesi, Derin Öğrenme ve DDİ ilişkisi

Yapay zeka (YZ) , makina öğrenmesi (MÖ), derin öğrenme (DÖ) ve doğal dil işleme (DDİ) kavramları birbiri ile sıklıkla karıştırılmaktadır. MÖ, DÖ ve DDİ alanları birbiri ile bir miktar örtüşsede birbirlerinden farklı alanlardır. Bu kavramlar arasındaki ilişki Şekil 3.2'de gösterilmektedir. Bu kavramlara kısaca açıklanırsa;

- **Yapay zeka (YZ)**, insan zekası gerektiren görevleri yerine getirebilecek sistemler oluşturmayı amaçlayan bir bilgisayar bilimi dalıdır. Temelleri 1950'lerde Dartmouth Koleji'nde düzenlenen bir çalışmaya dayanan ilk YZ çalışmaları mantık, buluşsal ve kural tabanlı sistemlerden oluşmaktaydı.
- **Makine öğrenimi (ML)**, el yapımı kurallar gerektirmeden birçok görevi büyük miktarda örneğe dayalı olarak öğrenen algoritmaların geliştirilmesi ile ilgilenen bir yapay zeka dalıdır.
- **Derin öğrenme (DÖ)**, yapay sinir ağı mimarilerine dayanan makine öğreniminin bir alt dalını ifade eder. Derin öğrenmedeki son gelişmeler sayesinde, NLP uygulamaları performansta benzeri görülmemiş bir artış elde etmiştir (Lauriola, Lavelli and Aiolfi, 2022).



Şekil 3.2. Yapay zeka, makina öğrenmesi, derin öğrenme ve doğal dil işleme arasındaki ilişki

- **Doğal Dil İşleme (DDİ)** ise yapay zekânın ilk yıllarında olduğu gibi insan dilini kural tabanlı ve buluşsal yöntemlere dayanarak yorumlayan ancak son yıllarda birçok görevi sıklıkla makina öğrenmesi ve derin öğrenme yöntemleri kullanarak gerçekleştiren yapay zekânın bir alt dalıdır.

3.3 Doğal Dil işleme Görevleri

Doğal dil işleme görevleri alt seviyede temel araştırmalar ve üst seviyede uygulama araştırmaları olmak üzere iki ana alt dala ayrılmaktadır (Lauriola, Lavelli and Aioli, 2022). Temel araştırmalar doğal dilin biçimi ile ilgilenen araştırma alanıdır. Bu alanlardan bazıları;

- **Kelime bilimi analizi (morphological analysis):** Kelime yapısı ve aralarındaki sistematik ilişki ile ilgilenir.
- **Sözdizimsel analiz(syntactic analysis):** Cümle yapısı ile ilgilenir.
- **Anlamsal analiz (semantic analysis):** Kelime, deyim ve cümlelerin gerçek anlamları ile ilgilenir.
- **Yorumsal (pragmatics):** Dilin farklı kullanımlarını inceler.

Uygulama araştırmaları ise Şekil 3.3’de görüldüğü üzere daha üst seviye araştırmalardan oluşmaktadır (Lauriola, Lavelli and Aioli, 2022). Bu alanlar şunlardır;

- **Bilgi Çıkarımı:** Metinlerden ilgili bilgilerin otomatik olarak çıkarılmasını sağlar örneğin, metin içinden Kişi, Kuruluş, Hastalık terimlerini bulma ve bu terimler arası ilişkileri çıkarmak.
- **Makina Çeviri:** Metinlerin farklı dillere çevrilmesini sağlar. Örneğin İngilizceden Türkçeye veya Rusçaya çeviri yapmak.
- **Soru Cevaplama:** Soruların otomatik olarak yanıtlanmasını sağlar. Örneğin banka veya firma chatbot'larına soru sormak ve gelen cevaba göre işlem yapmak.
- **Metin Sınıflandırma:** Belgelerin sınıflandırılması ve kümelenmesi gibi uygulamalı konularla ilgilenir.



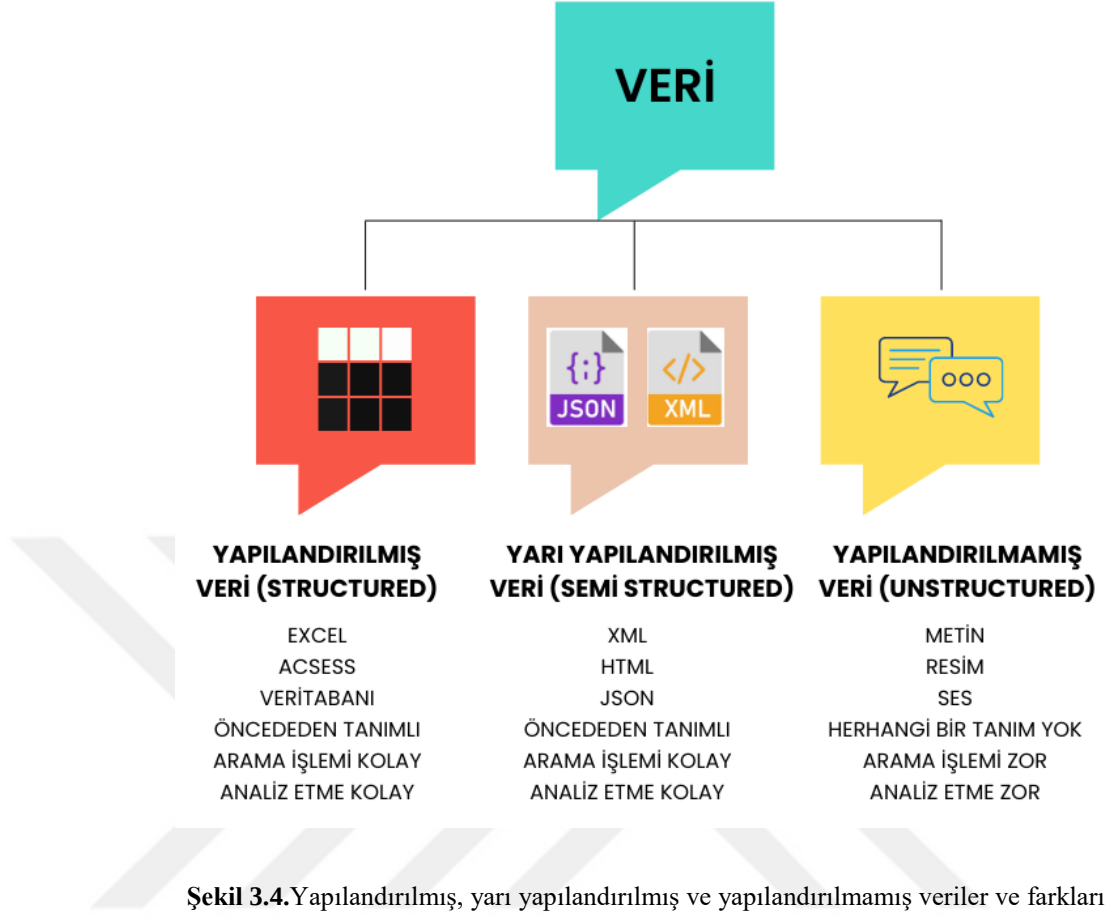
Şekil 3.3. Üst seviye doğal dil işleme uygulamaları

3.4 Verilerin Yapısı

Günlük hayatta veriler birçok farklı biçimde bulunabilir. Bu veriler genel olarak 3 kategori altında sıralanmıştır bunlar Şekil 3.4'te de görüldüğü üzere;

- yapılandırılmış
- yarı yapılandırılmış
- yapılandırılmamış

olmak üzere üç bölüme ayırabiliriz. Bu veri yapılarının detaylı bir anlatımı bölüm 3.4.1, bölüm 3.4.2 ve bölüm 3.4.2 'de verilmiştir.



3.4.1 Yapılandırılmış veriler (Structured data)

Yapılandırılmış veriler belirli bir biçime (formata) uygun şekilde bulunan verilerdir. Genel olarak satır ve sütunlardan oluşan veri tabanı biçimindedirler. Hem bilgisayar hem de insanlar için okunması kolay ve anlaşılabilirlerdir (Monica and Kumar, 2013). Özellikle makina öğrenmesi (sinir ağları, karar ağaçları) ve istatistiksel algoritmaları bir arada kullanan veri madenciliği yöntemlerinin analizi için sıklıkla kullanılan veri tipleridir (Gharehchopogh and Khalifelu, 2011). Çünkü veri madenciliği araçlarının metin veya web içinde anlamlı örüntüler (paterenler) bulabilmesi için bu verilerin yapısal olması gerekmektedir. Bu veri tipine;

- Veri tabanları (Access vb.)
- Excel tabloları

örnek olarak verilebilir.

3.4.2 Yarı Yapılandırılmış veriler (Semi Structured data)

Yarı yapılandırılmış veriler, ilişkisel veya tablo biçimindeki veri modeli yapısına uymayan ancak anlamsal öğeleri ayırmak için etiket veya başka işaretler içeren yapılandırılmış verilerdir (Buneman, 1997). Bilgi alışverişi için tam metin belgeleri ve veri tabanlarının artık tek veri biçimi olmadığı düşüncesi ile karşımıza çıkan yarı yapılandırılmış veriler son zamanlarda giderek daha yaygınlaşmaktadır (Monica and Kumar, 2013). Yarı yapılandırılmış veri modeli, birbiriyle ilişkili ancak farklı özelliklere sahip çeşitli kaynaklardan gelen bilgilerin (XML ve Doc vb.) tek bir bütün halinde birbirine uymasına izin verirler. (Hänig, Schierle and Trabold, 2010). Bu veri tipine;

- XML
- RDF
- HTML
- JSON
- JSONL

örnek olarak verilebilir.

3.4.3 Yapılandırılmamış veriler (Unstructured data)

Yapılandırılmamış veriler önceden tanımlanmış belirli bir yapısı olmayan verilerdir. Tipik olarak bir veri tabanının parçası olmayan görüntü, metin, e-posta ve diğer veri türlerini içerir (Feldman and Sanger, 2007). Ancak genel olarak metin ağırlıklıdır. Metni genellikle yapılandırılmamış veri olarak adlandırılır. Bununla birlikte, gerçekte, metnin birçok yapısı vardır - ancak çoğu açık değildir, bu nedenle metin içindeki bilgileri aramayı veya analiz etmeyi zorlaştırır (Grishman, 2015). Bu tez çalışmasının konusu olan bitki ve ilaç etkileşimleri gibi bilgilerin çoğu ise biyomedikal literatürde bulunan araştırma makalelerinde mevcuttur.

Yapılandırılmamış veriler, yapılandırılmış verilerle aynı şekilde sıralanamaz, aranamaz, görselleştirilemez veya analiz edilemezler bu nedenle bilgi paylaşmak ve değer sağlamak için yeni araçlar ve süreçler gereklidir (Hanig, Schierle ve Trabold, 2010).

Özellikle biyomedikal alanda PubMed veri kaynağı her yıl milyonlarca makalenin yayınlandığı bir sisteme dönüşmüştür. Bu milyonlarca

yapılandırılmamış makale içinden anlamlı bir verinin çıkarılabilmesi ve işlenebilmesi için çeşitli dil işleme teknikleri gereklidir. Bunun için doğal dil işleme teknikleri kullanılmaktadır.

3.5 Biyomedikal Verilerin Genel Olarak Yapısı

Biyomedikal metinlerin belge yapısı biraz karmaşık bir olaydır. Biyomedikal metinler doğal işlemenin ilgili olduğu iki alan vardır. Bu alanlar aşağıda verilmiştir;

- bilimsel yayınlar
- klinik belgeler

Bilimsel yayınlar makale özeti ve makale gövdesinden oluşmaktadır. Yapısal olarak makale gövdeleri giriş, araç ve yöntemler, sonuç ve tartışma gibi birçok bölümden oluşmakta ve bu bölümlerde birçok noktalama işareti, matematiksel denklem, şekil ve tablolar bulunmaktadır. Bu parantez ve noktalama işaretleri doğal dil işleme uygulamaları için bir takım problemler oluşturmaktadır ve işlenmesi zordur (Cohen, Christiansen and Hunter, 2011). Makale özetleri ise makale gövdelerine göre daha az bilgi içermekte olmalarına rağmen daha az noktalama işareti ve daha az cümle içermekte ve böylece daha anlaşılır yapıdadırlar. Bu nedenle bilimsel makaleler üzerine yapılan araştırmaların çoğunluğu makale gövdeleri yerine makale özeti üzerine çalışmalarından meydana gelmektedir (Cohen and Demner-Fushman, 2014).

Çizelge 3.1. Yapılandırılmış ve yapılandırılmamış medikal kayıtlarla ilgili bir örnek

Veri Biçimi	Yapılandırılmış (Structured)	Yapılandırılmamış (Unstructured)
Zaman	dd/mm/yyyy 24 Eylül 2022	'Bugün erken saatlerde Bay X sırt ağrısı yaşadı'
Semptom	N242300 Sinirsel Ağrı	...gece boyunca sağ kolunda meydana gelen ağrı uykusunu böldü'
Teşhis	C109923 Tip2 diyabet	Hastanın gizli diyabeti vardır
Reçete	0104234 (BNF codeine)	Px codeine 60 mg 7 gün
Test	43F1.00 Romatoid faktör pozitif	Romatoid faktör 42 IU/mL olarak pozitif sonuç verdi.

Klinik belgeler ise sađlık hizmetleri içindeki ana iletişim biçimini temsil eder ve aslında yapısal bir biçimdedirler ancak sayısız farklı biçimlerde bulunmaktadır. Örneđin Çizelge 3.1 incelendiğinde medikal kayıtlar her ne kadar yapılandırılmış görünselerde aslında yapılandırılmamış birer belgedirler. Bu yapılandırılmamış verilerin düzenlenmesi ve yapılandırılmış bir biçime getirilmesi, makina öğrenmesi algoritmalarının bu veriler içinden anlamlı bilgiler çıkarılabilmesi ve analiz edilebilmesi için önemlidir.



4. BİLGİ ÇIKARIMI (INFORMATION EXTRACTION)

Bilgi çıkarma yapılandırılmamış metinsel verilerden varlık, ilişki, iddia, konu ve buna benzer faydalı bilgileri yapılandırılmış bir şekilde çıkarma işlemidir. Bilgi çıkarma sistemleri doğal dil işleme, dil modelleme ve yapılandırılmış çıkarım (structured extraction) tekniklerine dayanmaktadır (Adnan and Akbar, 2019). Biyomedikal alanda ise bilgi çıkarımının (relation extraction) amacı, literatürde bulunan yapılandırılmamış biyomedikal metinden ilaç, bitki, gen, hastalık, protein vb. varlık adlarını ve protein-protein etkileşimi (PPI), ilaç-ilâç etkileşimi (DDI'ler), kimyasal-protein etkileşimi (CPI) ve bu tez çalışmasının konusu olan ilaç-bitki etkileşimi gibi değerli biyomedikal ilişkileri gelişmiş doğal dil işleme (NLP) teknikleri aracılığıyla yüksek doğruluk ve verimlilikle otomatik olarak yakalamak ve çıkarmaktır.

Bilgi Çıkarımının en önemli 3 görevi şunlardır:

- Adlandırılmış Varlık Tanıma (Name Entity Recognition)
- Varlık Bağlama (Entity Linking)
- İlişki Çıkarımı (Relation Extraction)

4.1 Adlandırılmış Varlık Tanıma (Name Entity Recognition-NER)

Adlandırılmış Varlık Tanıma (AVT, Name Entity Recognition-NER), bir metinde bulunan varlıkları tanımlamayı ve daha sonra bu varlıkları Çizelge 4.1'de yer alan kuruluş, konum, zaman kategorisine veya alana özgü olarak hastalık, ilaç ve gıda kategorisine göre sınıflandırmayı amaçlamaktadır.

Çizelge 4.1. Bir belge içinde adlandırılmış varlık örnekleri

Tip	Etiket	Örnekler
Kişiler (Person)	PER	Steve jobs, Aziz Sancar, Cengiz Aytmatov
Kuruluşlar (Organisation)	ORG	Apple, Amazon, Space X
Konumlar (Location)	LOC	Kırgızistan, İzmir, Erciyes, Berçelan
Zaman (Date)	DATE	15 Mart 1959, 1/04/1965, 26.05.1982

Bir AVT modelinin çalışma prensibi şu şekildedir;

Verilen bir $X=\{x_1,x_2,x_3\}$ giriş cümlesinde x_i , i-inci kelime/token ve N ise cümle uzunluğunu temsil etsin. Bu durumda; AVT sisteminin amacı X'teki her kelimeyi/tokeni sınıflandırmak ve onlara karşılık gelen bir $y \in Y$ etiketine atamaktır; burada Y önceden tanımlanmış etiket türlerinin (örneğin, İLAÇ, HASTALIK ve PROTEİN) bir listesidir.

Varlıkların ilk olarak tanımlanması (adlandırılmış varlık tespiti) ve daha sonra bunların sınıflandırılması (anlamsal sınıflandırma) AVT'nin alt görevleridir. Örneğin aşağıdaki cümleyi ele alalım;

“Apple CEO’su Tim Cook yeni ürün tanıtımını Cupertino’da Flint Merkezinde gerçekleştirdi (Song *et al.*, 2018).”

yukarıdaki cümlede bir AVT sistemi cümleyi şu şekilde kategorize edecektir.

“Apple [Organization] CEO’su Tim Cook [Person] yeni ürün tanıtımını Cupertino’da [Location] Flint Merkezinde [Organization] gerçekleştirdi (Song *et al.*, 2018).”

Bu cümlede Apple bir kuruluş, Tim Cook bir kişi, Cupertino bir lokasyon ve Flint Merkezi’de bir kuruluş olarak kategorize edilmiştir.

Adlandırılmış Varlık terimi ilk olarak 1996 yılında altıncı MUC (Message Understanding Conference) konferansında “varlıkların benzersiz tanımlayıcılarına (unique identifiers of entities)” atıfta bulunmak için ortaya çıkmıştır (Chinchor and Robinson, 1997). Özel adların otomatik olarak çıkarılması, makine çevirisi, bilgi alma, soru yanıtlama ve özetleme gibi birçok problem için birçok faydalı çözüm sunmaktadır. Tüm metni sınıflandırmak yüksek depolama gerektiren ve maliyetli bir süreçtir adlandırılmış varlıklar tanıma sistemleri ile bir belgeyi kolayca tarayabilir ve bu belge içindeki ilgili ana öğeleri etkili bir şekilde sınıflandırabilir (Mansouri, Affendey and Mamat, 2008).

4.1.2 AVT sistemlerine neden ihtiyaç duyulmakta?

Haber Siteleri: Haber ve yayınevleri günlük olarak büyük miktarda çevrim içi içerik üretir. Bu içeriklerden en iyi şekilde yararlanmak için bunları doğru bir şekilde kategorize etmek çok önemlidir. Adlandırılmış Varlık Tanıma, tüm haber makalelerini otomatik olarak tarayabilir ve bunlarda tartışılan başlıca kişilerin,

kuruluşların ve yerlerin hangileri olduğunu ortaya çıkarabilir. Her makale için ilgili etiketleri bilmek, makalelerin tanımlanmış hiyerarşilerde otomatik olarak kategorilere ayrılmasına yardımcı olur ve daha iyi içerik keşfi sağlar (Ruokolainen *et al.*, 2020). Bunun dışında bir haberin yalan haber mi yoksa orijinal bir haber olup olmadığı kanısına karar vermek amacıyla AVT sistemleri kullanılabilir (Al-Ash and Wibowo, 2018).

Müşteri hizmetleri: Soru cevaplama sistemleri, sorulara somut cevaplar sağlayan sistemlerdir ve AVT teknikleri bu sorulara verilecek cevapların seçimini kolaylaştırmak için sıklıkla kullanılırlar. Örneğin, bir işletmenin dünya genelinde birkaç şubesi olduğunu ve bir müşterinin şubelerinden biri hakkında çevrim içi şikayetde bulunduğunu varsayalım. Şikâyetin ilgili şubeye iletilmesi yöneticiler için zaman alıcı ve sıkıcı bir süreç olabilir. AVT sistemi, şikâyetin içinde belirtilen yere, ürüne ve kişiye göre sınıflandırılmasına yardımcı olabilir. Ardından herhangi bir insan müdahalesine gerek kalmadan şikâyet otomatik olarak ilgili şubeye ve ekibe iletilebilir.

Bilimsel yayınlar: Bir çevrim içi dergi veya yayın sitesinde milyonlarca araştırma makalesi ve bilimsel makale bulunur. Yalnızca tek bir konu hakkında yüzlerce hatta binlerce makale olabilir. Tüm bu verileri iyi yapılandırılmış bir şekilde organize etmek zor olabilir. Belirli bir bilgiyi aramak için gözden geçirme veya bir takım sorgular muhtemelen yetersiz kalabilir. Belgeleri, sahip olduğu ilgili varlıklar bazında ayırmak, konuyla ilgili çok sayıda bilgiyi gözden geçirme zahmetinden kurtarabilir. Örneğin, Makine Öğrenimi üzerine PubMed de bir sorgu yapıldığında 98.000 adet makale çıkmaktadır. Eğer bu makaleler üzerindeki varlık etiketleri çıkartılırsa, örneğin “ilaç-ilaç etkileşimleri için evrimsel sinir ağlarının kullanımı” üzerine tartışılan makaleler hızlı bir şekilde bulunabilir.

4.1.3 Biyomedikal alanda AVT (NER) ve karşılaştıkları zorluklar

Varlıklar, bir kişi, konum ve kuruluşlar gibi genel bir tanım olabileceği gibi herhangi bir alana özgüde olabilirler. Biyomedikal alanda AVT (NER) sistemlerinin amacı milyonlarca yapılandırılmamış biyomedikal literatürden “PROTEİN, DNA, RNA, ENZİM, GEN, HÜCRE TİPİ, KİMYASALLAR, HASTALIK, GIDA ve İLAÇ” gibi biyomedikal varlıkları önce tanımlamak ve daha sonra sınıflandırmaktır. Çizelge 4.2.’de biomedikal DDİ alanında yapılan adlandırmış varlık örneklerinin bir gösterimi verilmiştir. Biyomedikal alanda AVT

sistemleri kiři ve konum gibi genel adlandırmalara göre çıkarılması zor görevlerdir. Bu zorlukların başlıca nedenlerini aşağıda sıralanmıştır ;

1. Mevcut araştırma sayılarının gittikçe artması ve bu nedenle yeni teknik terimlerinde sayısının artması. Dolayısıyla tüm yeni kelimeler için bir sözlük oluşturulması zor bir görevdir (Song *et al.*, 2018).
2. Aynı kelimeler veya ifadeler, bağlamları açısından farklı adlandırılmış varlıklar olarak sınıflandırılabilir (Song *et al.*, 2018).
3. Biyomedikal varlık uzunlukları oldukça uzun olabilir ve aralarında sıklıkla kısa çizgi gibi kontrol karakterleri içerebilir (örneğin, “**12-o-tetradekanoilforbol 13-asetat**”) (Song *et al.*, 2018).
4. Bir varlık farklı araştırma makalelerinde farklı adlandırılmış olabilir. Örneğin e.g., “**Lymphocytic Leukemia**” ifadesi bir başka makalede “**Lymphoblastic Leukemia**” olarak ifade edilmiştir (Perera, Dehmer and Emmert-Streib, 2020).
5. Dördüncüsü, kısaltma ifadeleri biyomedikal alanda sıklıkla kullanılır ve anlam belirsizliği yaşanır. Örneğin, “**TCF**” ifadesi “**T cell factor** (T hücre faktörü)” veya “**Tissue Culture Fluid** (Doku Kültürü Sıvısı)” anlamına gelebilir (Wang *et al.*, 2008).
6. Bazı adlandırmalar iç içe geçmiş şekildedir. Örneğin “**91 and 84 kDa proteins**” ifadesi “**91 kDa protein**”ve “**84 kDa protein**” olarak da ifade edilebilir (Perera, Dehmer and Emmert-Streib, 2020).
7. Son olarak, biyomedikal terimlerde normal terimler veya fonksiyonel terimler sıklıkla birleştirilir, bu nedenle biyomedikal terimler çok uzun olabilir. Örneğin, “**HTLV-I-infected**” ve “**HTLV-I-transformed**”, kelimeleri “**I**”, “**infected**” ve “**transformed**” gibi normal terimlerin birleşmesiyle oluşmuştur. Bu tarz cümleleri adlandırmak zordur.

Çizelge 4.2. Biomedikal Doğal Dil İşleme alanında adlandırılmış varlık örnekleri (Cohen and Demner-Fushman, 2014).

Biyomedikal Varlık (Sınıf)	Örnekler
Hücre Tipi (Cell Type)	Natural killer cell, NK cell
Kimyasal (Chemicals)	Citric acid, 1,2-diidepentane
Hastalık (Disorder)	Amyotrophic lateral sclorosis
İlaç (Drug)	Cyclosporin A, warfarin
Gıda (Food)	empty passion fruit juice, Colby-Jack cheese
Gene/Protein	HSP60, protein kinase C
Mutasyonlar (Mutations)	C10T, Ala64→Gly
Maligniteler (Malignancies)	Carcinoma, breast neoplasm

4.2 İlişki Çıkarımı (Relationship Extraction)

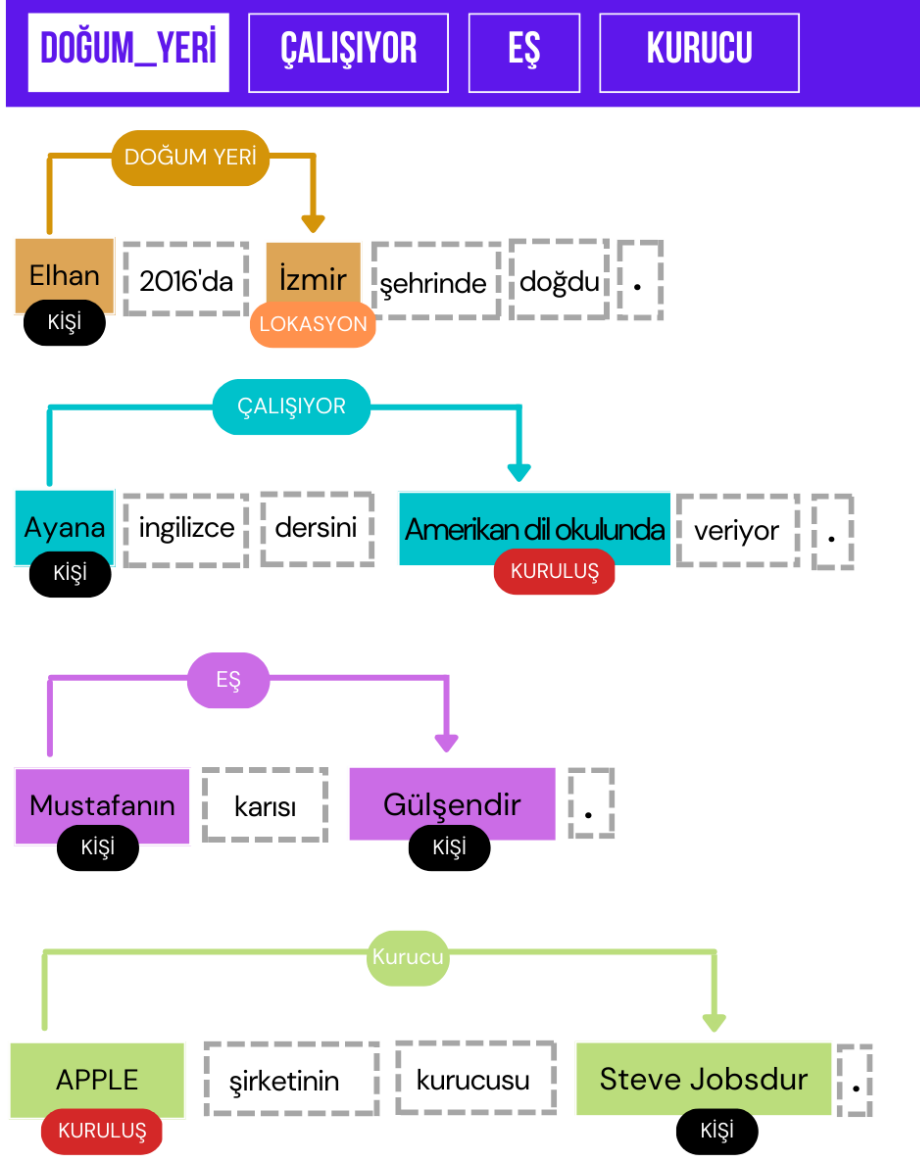
İlişki çıkarımının (relationship extraction) amacı bir cümle veya bir belge içinde geçen kişi, kuruluş ve lokasyon veya biyomedikal gibi özel bir alan içindeki cümle ve paragraflardan protein, ilaç ve hastalık gibi aday varlık çiftleri (candidate entity pair) arasında anlamsal bir ilişkisinin olup olmadığını belirlemektir. Örneğin Şekil 4.1’de yer alan cümleler incelenirse ;

“Elhan 2016 yılında İzmir şehrinde doğdu.” cümlesinde (Elhan,İzmir)**DOĞUM_YERİ** ilişkisi vardır. Bu cümlede bir kişi olan **Elhan** varlığı ve bir lokasyon olan **İzmir** varlığı arasında **Doğum yeri** ilişkisi bulunmaktadır.

“Ayana ingilizce dersini Amerikan Dil okulunda veriyor.” cümlesinde ise (Ayana, Amerikan Dil Okulu)**ÇALIŞIYOR** ilişkisi vardır. Bu cümlede bir kişi olan **Ayana** varlığı ve bir kurum olan **Amerikan Dil Okulu** varlığı arasında **Çalışıyor** ilişkisi bulunmaktadır.

“Mustafa’nın karısı Gülşendir.” cümlesinde ise (Mustafa, Gülşen)**EŞ** ilişkisi vardır. Bu cümlede bir kişi olan Mustafa varlığı ve bir başka kişi olan **Gülşen** varlığı arasında **eş** ilişkisi bulunmaktadır.

“APPLE şirketinin kurucusu Steve Jobs’tur.” cümlesinde ise (APPLE, Steve Jobs)**KURUCU** ilişkisi vardır. Bu cümlede ise bir şirket olan **APPLE** varlığı ve **Steve Jobs** varlığı arasında **kurucu** ilişkisi bulunmaktadır.



Şekil 4.1. İlişki çıkarımı (Relationship extraction) hakkında genel bir örnek

Bir cümle içinde geçen iki varlık arasındaki ilişki şu şekilde formülize edilebilir.

$$(v_1, v_2, s) \rightarrow r \quad (4.1)$$

Burada “ v_1 ” ve “ v_2 ” simgeleri “ s ” cümlesinde bulunan varlıkları ifade etmekte ve “ r ” ise olası bir ilişkinin türünü ifade etmektedir. Bir cümledeki her

varlık çifti arasında bir ilişki olmayabilir bu nedenle (v_1, v_2, s) ilişkisi aday ilişki olarak tanımlanır. Olası r değerleri kümesi ile en az üç görev olabilir.

1-Yalnızca Tespit: Her bir ilişki adayı için gerçek bir ilişkinin olup olmadığını tahmin etme görevidir. Bu görev ikili bir sınıflandırma gibi düşünülebilir. $S=w_1, w_2, \dots, v_1, \dots, w_j, \dots, v_2, \dots, w_{n+1}$ cümlesi verildiğinde “ v_1 ” ve “ v_2 ” varlıkları ifade ederse bu durumda $f(.)$ fonksiyonu şu şekilde formülize edilir.

$$fr(T(s)) = \begin{cases} +1, & \text{eğer } v_1 \text{ ve } v_2, "r" \text{ etiketi ile ilişkili ise} \\ -1, & \text{aksi durum} \end{cases} \quad (4.2)$$

Bu formülde bulunan $T(s)$, “s” cümlesinden çıkarılan öznitelikleri ifade eder. Bir r etiketi, +1 (iki varlık birbiriyle ilişkilidir) veya -1 (iki varlık birbiriyle ilişkili değil) değerini alır. Örneğin aşağıdaki cümle incelenirse;

"Ginkgo biloba'nın aspirin ile kullanımına bağlı olası kanama riskleri vardır"

Bu cümlede **Ginkgo biloba** bitkisi ve **aspirin** ilacı arasında bir ilişki olduğu görülmektedir. Burada $v_1 = \text{“Bitki”}$ ve $v_2 = \text{“İlaç”}$ varlıklarını ifade eder.

$$(\text{Ginkgo biloba}_{(\text{Bitki})}, \text{aspirin}_{(\text{İlaç})})_{\text{POSITIVE}} = (v_1, v_2)_{\text{POSITIVE}} \quad (4.3)$$

Bu durumda bu cümlede ilişki çıkarımının amacı v_1 ve v_2 varlıkları arasında POZİTİF bir ilişkinin olup olmadığını tanımlamak olabilir.

2-Sadece sınıflandırma: Her bir ilişki adayı için N adet sınıflandırmanın olduğu ilişki tipidir. Sınıflandırıcının performansını yalnızca N ilişki türleri üzerinde dikkate alınır.

3- Tespit ve sınıflandırma birleşimi: Her bir ilişki adayı için N ilişki kategori sayısını gösterebilir. Bu kategoriye ek olarak iki varlık arasında herhangi bir ilişkinin olmadığı -1 (negatif) ilişki kategorisi eklenirse bu durumda bu sınıflandırma N+1 adet sınıflandırma olarak gerçekleştirilir.

4.3 Adlandırılmış Varlık Tanıma ve İlişki Çıkarım Sistemlerinin Makina Öğrenmesi Modelleri İle Tahmini İçin Gerekli Etiketli Veri Kümesi Oluşturma Süreçleri

Denetimli öğrenme yöntemleri için eğitim verisi oluşturmak birçok farklı bileşenin bir araya gelmesiyle oluşur. Oluşturulacak olan etiketli veriler için gerekli olan yapılar şunlardır (Campos, Matos and Oliveira, 2012).

1-Derlem (Korpus): Bir derlem genel olarak herhangi bir alana özgü verilerin bir arada olduğu ve bu alana özgü terim veya ilişkilerin etiketli veya etiketlenmemiş olduğu veri kümeleridir. Örneğin proteinler ile ilgili bir AVT sistemi yapılacaksa ilk olarak biyomedikal literatürden proteinler ile ilgili metin belgeleri bir araya getirilir bu etiketlenmemiş bir derlemdir. Eğer bu derlemde bulunan protein varlıkları ve aralarındaki ilişkiler alan uzmanı veya kitle kaynak yöntemleri kullanılarak etiketlenmiş ise o zaman bu etiketli ya da altın standart derlem olarak adlandırılır.

2-Alan Uzmanı Küratör: Bir derlem hazırlamak için alan uzmanı kişilere ihtiyaç duyulmaktadır. Örneğin biyomedikal alanda uzman küratörler bilimsel metinlerden ilaç, hastalık, protein gibi varlıkları ve bu varlıklar aralarındaki ilişkileri çıkartan doktor, profesör, araştırmacı veya laborant gibi kişilerden oluşmaktadır.

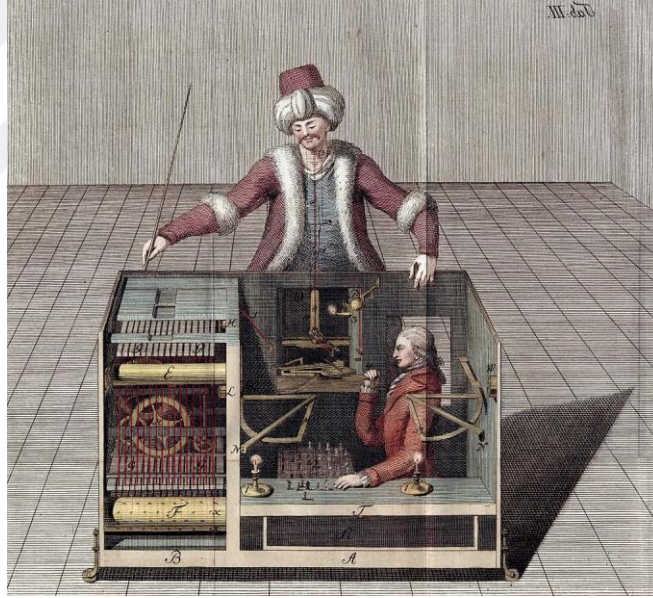
3-Kitle kaynak kullanımı (crowdsourcing): Kitle kaynak kullanımı (Crowdsourcing), özellikle İnternet üzerinden uzman ve uzman olmayan kişilerden hizmet, bilgi ve fikir alma süreci olarak tanımlanır. Kitle kaynaklı olabilecek görevler genel olarak insan uzmanlığı gerektiren bilimsel sorunları içermektedir. Kitle kaynaklı problemler genel olarak “mikro görevler” veya “mega görevler” olarak sınıflandırılır (Khare *et al.*, 2016).

Mikro görevler, genellikle zamandan tasarruf etmek ve toplu bir çözüm oluşturmak için çok sayıda insanın paralel olarak katkıda bulunduğu küçük görevlerdir; Bu görevler:

- Kelime anlam belirsizliğini giderme görevleri
- Bilimsel metinlerden hastalık, ilaçlar, genler vb. gibi varlıkları adlandırılmış varlık tanıma (AVT) sistemleri için etiketleme görevleridir.

Mega görevler ise daha büyük karmaşık problemlerin çözümüne dayanmaktadır. Örneğin protein katlama oyunu olan foldfit (Miller, Vandome and McBrewster, 2010) buna örnek verilebilir.

Kitle kaynak kullanımı için en popüler platform istenilen göreve göre belirli bir ücret karşılığı hizmet veren Amazon Mechanical Turk (MTurk) (Mechler, 2017) platformudur. İsmi 1769 yılında yapılmış olan (bkz. Şekil 4.2) bir masa ve bir Osmanlı figüründen oluşan satranç otomatından alır. Mekanik Türk'ün içine kimsenin göremeyeceği bir şekilde gizlice saklanmış oldukça yetenekli usta bir satranç uzmanı durmaktadır. Bu uzman karşısındakileri hayrete düşürecek derecede zeki ve etkileyici bir satranç oynayarak Napoleon Bonaparte ve Benjamin Franklin gibi devlet adamlarının da aralarında bulunduğu birçok rakibi yenerek Avrupa ve Amerika'da oldukça ün yapmıştır (Barthelmeh, 2011); (Fiske, 1859). Bu aletten esinlenen Amazon MTurk'de arkasındaki insan zekâsını gizleyerek benzer bir akıllı sistem oluşturmaya çalışan bir platformdur.



Şekil 4.2. Mekanik Türk'ün rakibiyle oynarken operatörün içeride nasıl oturduğunu düşündüğünü gösteren bir kesiti. (Wikipedia contributors, 2022)

4-Etiketleme Aracı: Başarılı bir doğal dil işleme modeli elde etmek için metin içerisindeki varlıkların ve ilişkilerin iyi belirlenmesi gerekmekte ve çok iyi başarılı bir etiketleme aracına ihtiyaç duyulmaktadır. Bunun için çeşitli görevler için etiketleme araçları mevcuttur. Etiketleme işlemi için Brat (2022), Docano (2022), DataTurks (2022) ve son olarak aktif öğrenme desteği vererek etiketleme işleminin süresini azaltan Prodigy (Prodi.gy annotation tool, 2022) araçları kullanılabilir.

4-Küratörler arası uyum (Inter Annotator Agreement) : Herhangi bir etiketleme görevi ile ilgili en büyük problem tutarlılıktır. Tutarlılık iki veya daha fazla adayın etiketleme görevinde etiketledikleri varlık veya ilişkilerin uyuşma oranı ile ilgilenir. Uzmanların etiketlediği birbirinden ne kadar farklı olduğunu ölçmek veri kalitesi bakımından gerekli bir olaydır. Bu bağlamda küratörler (alan uzmanı, kitlek kaynak kullanımı) arasındaki uyuşmazlığı hesaplamak için kategorik verilerde yaygın olarak Cohen Kappa istatistik ölçüsü kullanılmaktadır. Cohen kappa aşağıdaki gibi tanımlanmaktadır:

$$\kappa = (p_o - p_e) / (1 - p_e) \quad (4.4)$$

Burada “ p_o ”, örneğe atanan etiketin gözlemlenen uyuşma oranını ifade ederken “ p_e ” ise her iki küratörün rastgele etiketler atadığında beklenen uyuşmadır. Bu denklemde “ p_e ”, sınıf etiketleri için küratör başına deneysel ön olasılıklar kullanılarak tahmin edilir.

Çizelge 4.3. Cohen kappa katsayısının yorumlanması

Cohen's Kappa istatistiği(κ)	Uyuşma oranı
<0.0	Zayıf uyuşma
0.00-0.20	Önemsiz uyuşma
0.20-0.40	Orta derece uyuşma
0.41-0.60	Normal derece uyuşma
0.61-0.80	Önemli derece uyuşma
0.81-1.00	Mükemmel uyuşma

Cohen Kappa değerini hesaplamak için Kohen kappa Python Scikit-learn (2022) kütüphanesi kullanılmıştır. Çizelge 4.3'e göre Landis ve Koch (1977), 0.81'den büyük bir κ değerinin neredeyse mükemmel bir uyuşmaya işaret ettiğini belirtmektedir.

5-Altın Standart Derlem: Alan uzmanı veya kitle kaynak küratörleri tarafından belirli rehberlere ve belirli kurallara göre herhangi bir alana özgü kritik öneme sahip terimlerin ve aralarındaki ilişkilerin manuel olarak etiketlendiği verilerdir (Campos, Matos and Oliveira, 2012). Altın standart, belirli bir durumda mevcut olan en iyi ölçüttür. Altın standart bir çalışma, kapsamlı bir şekilde test edilmiş ve bu alanda güvenilir bir yönteme atıfta da bulunabilirler (Xintong *et al.*, 2014).

5. ÖNCEKİ ÇALIŞMALAR

Bu bölümde adlandırılmış varlık tanıma ve ilişki çıkarımı için yapılan çalışmalar 3 aşama olarak ele alınmıştır. İlk aşama adlandırılmış varlık tanıma üzerine yapılan çalışmalar, ikinci aşama ilişki çıkarımı üzerine yapılan çalışmalar ve son aşama ise bitki-ilaç etkileşimi üzerine yapılan çalışmalardan oluşmaktadır.

5.1 Adlandırılmış Varlık Tanımlama Sistemleri İçin İlk Yöntemler

Adlandırılmış varlık tanıma için ilk yaklaşımlar sözlük tabanlı ve kural tabanlı yaklaşımlardır (Goyal, Gupta and Kumar, 2018). Bu yaklaşımlar detaylı olarak aşağıdaki bölümlerde ele alınmıştır.

5.1.1 Sözlük tabanlı yöntemler

Bu yaklaşım çok sayıda örnek içeren önceden tanımlanmış bir sözlük veya terminoloji kaynağına bağlıdır ve verilen bir metin ifadesini bu daha önce tanımlanmış sözlükte bulunan kelimelerle eşleştirmeye çalışır. Bir başka deyişle bir sözlük ile metnin bir kısmı arasında birebir eşleşme olup olmaması ile ilgilenir (Quimbaya *et al.*, 2016). Örneğin aşağıdaki yapı incelenirse;

$$G = \{ \text{'diabet'}, \text{'hipertansiyon'}, \text{'COPD'}, \text{'grip'} \}$$

DocX = "Suzan 67 yaşında bir diyabet hastası ve CODP ve hipertansiyon şikayetleri vardır. Suzan hanım 20 gün önce hastaneye gribal enfeksiyondan kaynaklı zatüre nedeniyle..."

Burada "**G**" hastalık teşhis terminolojisine ait bir sözlük olsun ve "**DocX**" ise bir derlemde bulunan bir belge olduğunu varsayalım. Bu cümlede sözlük tabanlı bir adlandırılmış varlık sistemi çalıştırılırsa DocX cümlesindeki varlıklar şu şekilde etiketlenir:

DocXAnnot = "Suzan 67 yaşında bir **diyabet** hastası ve **CODP** ve **hipertansiyon** şikayetleri vardır. Suzan hanım 20 gün önce hastaneye gribal enfeksiyondan kaynaklı zatüre nedeniyle..."

Yukarıdaki cümle incelendiğinde **G** hastalık teşhis terminoloji sözlüğü **DocX** begesinde yer alan diyabet, CODP ve hipertansiyon kelimeleri ile eşleşerek bu kelimeleri etiketlemiştir. Ancak "**grip**" hastalığı "**gribal enfeksiyon**" şeklinde

yazıldığından “G” sözlüğü bu kelimeyi algılayamamış ve bu nedenle herhangi bir eşleşme olmamıştır (Quimbaya *et al.*, 2016).

Bunun dışında biyomedikal literatürde bulunan varlıklar için bir sözlük tabanlı AVT modeli oluşturulacak olunursa dört adet sözlük tabanlı özelliğe odaklanılır. Bu özellikler aşağıda detaylı bir şekilde açıklanmıştır (Alshaikhdeeb and Ahmad, 2016):

Değiştirici (Modifier): Genel olarak biyomedikal varlıklar tarafından takip edilen kelime grubundan oluşur. Bir başka deyişle, biyomedikal varlıklardan önce meydana gelen anahtar kelimeleri içerir.

Önemsiz (Trivial): Bu özellik, biyomedikal varlıklar için kullanılabilecek alternatif adlar veya eşanlamlıları ile ilişkilidir. Örneğin kimyasal bileşiklerde "etanol" yerine "ethyl" gibi önemsiz adlandırmaların kullanılması.

Aile (Family): Biyomedikal varlıkların ailesidir. Örneğin “Etanol” varlığının ailesi “Alkol’dür”. Bu özellik, biyomedikal varlıklar arasındaki 'kısmi' ilişkileri gösterir.

Kısaltma (Abbreviation): Biyomedikal varlıkların kısaltmalarını içerir. Örneğin Kalsiyum için "cl" bir kısaltmadır.

Sum (moleküler formül): Bazı biyomedikal varlıklar iki biyomedikal varlığın birleşiminden oluşur. Örneğin “carbonhydrat” varlığı "carbon" ve "hydrat" varlıklarının birleşmesiyle meydana gelir.

Çizelge 5.1. Biyomedikal varlıkların hangi sözlüklere ait olduklarını gösteren bir örnek

Token	Değiştirici	Önemsiz	Aile	Kısaltma	Sum
iron (iii)	0	1	0	0	0
complexes	1	0	0	0	0
acids	0	0	1	0	0
NPS	0	0	0	1	0
CH3CN	0	0	0	0	1

Çizelge 5.1 incelenirse “iron (iii) (demir)” varlığı “oxide (oksit)” için önemsiz bir varlık olarak kabul edilir. Bu nedenle **Önemsiz** sütununa 1 atanmıştır. Öte yandan ikinci örnekte yer alan “complexes (kompleksler)” genellikle biyomedikal varlıklardan sonra “iron(iii) complexes (demir (iii) kompleksleri)” olarak meydana geldiği için **Değiştirici** sütununa 1 atanmıştır. Üçüncü örnek olan “acids (asitler)”, asitler ailesinin bir alt kimyasal örneğini içerdiği için **Aile** sütununa 1 atanmıştır. Dördüncü örnek yer alan “NPS” ise, 'New Psychoactive Substances'ın” ilacının kısaltması olduğu için **Kısaltma** sütununa 1 atanmıştır. Son olarak Methene (CH₃) ve Siyanür (CN) varlıklarının kombinasyonu olan Asetonitrilinin kimyasal bileşiği 'CH₃CN' olarak ifade etmiştir ve bu nedenle **Sum** sütununa 1 atanmıştır.

Bu yöntemin sınırlamaları şunlardır (Song, Yu and Han, 2015);

- AVT modelinin başarısı sözlük ile doğru orantılıdır. Sözlük ne kadar başarılı ise model ancak o kadar başarılıdır.
- Biyomedikal literatür çalışmalarının sayısı gittikçe artmakta ve buna paralel olarak da yeni terimlerde ortaya çıkmaktadır ve sözlüklerin yeni terimleri yakalaması için sürekli güncellenmeleri gerekmektedir.
- Bir kısaltma başka bir varlığın kısaltması olabilir. Bu sorun ancak kısaltmaların sözlükten atılmasıyla çözülebilir.
- Sözlük tabanlı yaklaşımlar tarafından kullanılan tam eşleştirme teknikleri, farklı şekilde adlandırılmış terimleri farklı terimler olarak ele alır. Örneğin "DC2-dopamin reseptörü" bir gen adıdır ve bu ifade "dopamin DC2 reseptörü" gibi farklı yazım varyantına sahiptir. Eğer bu iki adlandırma sözlükte yoksa ya bulunamayacak terimleri farklı terimler olarak ele alır.

5.1.2 Kural tabanlı yöntemler

Bu sistemler dil uzmanlarından elde edilen el yapımı kurallara bağlıdır. Örüntü tabanlı ve bağlam tabanlı kurallar olarak ikiye ayrılırlar . Ayrıca kalıplara dayalı kuralların yanı sıra buluşsal yöntemleri (örneğin, fiilden önce bir edat gelemmez) ve bilgi tabanlarını da kullanabilir (Goyal, Gupta and Kumar, 2018). Dil bilimini kullandıkları için verimli oldukları kabul edilirler (Shaalan, 2010). Bunun dışında yeterli doğruluğu elde etmek için alana özgü özellikler kullanırlar.

Kural tabanlı yöntemlere örnek olarak şöyle bir cümlemiz olsun;

“Prof. Dr. Mustafa ve Doc Dr. Gülşen yazdıkları doğal dil işlemenin bir alt dalı olan bilgi çıkarımı üzerine bir ders verdi.” ve şöyle bir kural yazarsak:

Başında herhangi bir unvan bulunan bir isimden sonra gelen isimler özel isimdir (Budi and Bressan, 2003);

<Prof. Dr., Capitalized_word(X)> => person_named(X)

<Doc. Dr., Capitalized_word(X)> => person_named(X)

Bu durumda yukarıdaki aşağıdaki şekilde etiketlenir.

“**Prof. Dr. Mustafa** [PERSON] ve **Doc. Dr. Gülşen** [PERSON] yazdıkları doğal dil işlemenin bir alt dalı olan bilgi çıkarımı üzerine bir ders verdi”.

Bu cümlede yazılan kural neticesinde cümlede geçen Prof. Dr. ve Doc Dr. kelimelerinden sonraki kelimeler kişi olan PERSON etiketi ile etiketlenmiş olduğu görülmektedir. Kural tabanlı yaklaşımlara biyomedikal alanda bir örnek verilirse;

Örneğin, "Metanol veya N-metilformamid gibi polar organik çözücüler lipazları etkisiz hale getirir" cümlesini incelersek.

Bu cümlede geçen kimyasal terimleri (büyük harfler, sayılar ve özel semboller gibi yüzey özelliklere sahip olanlar) yakalamak için bir takım kimyasal kök formları tanımlanabilir. Örneğin içinde met ve metil geçen kelimeler kimyasal varlıklardır denildiğinde bu durumda yukarıdaki cümlede geçen;

Metanol ve **N-metilformamid** varlıkları "metil" ve "met" kimyasal kök formlarını içerdikleri için kimyasal temel terimler olarak tanımlanır (Narayanaswamy, Ravikumar and Vijay-Shanker, 2002).

Çizelge 5.2. Kural tabanlı yöntemlerin avantaj ve dezavantajları (Chiticariu, Li and Reiss, 2013;Goyal, Gupta and Kumar, 2018)

Avantaj	Dezavantaj
Bildirimsel (Declarative)	Sezgisel
Buluşsaldır (Heuristic)	Sıkıcı
Alan bilgisini birleştirmek kolay	Alana özgüdür ve taşınabilir değil
Hata nedenini öğrenmek ve düzeltmek kolay	Maliyetlidir

Bu yöntemin sınırlamaları Çizelge 5.2’de verildiği gibi şunlardır (Chiticariu et al., 2013; Goyal, Gupta and Kumar, 2018);

- Bu sistemler hem zaman alıcı hem de oldukça maliyetlidirler.
- Alana özgü olmaları ve bu nedenle taşınabilir olmamalarıdır.
- Bu sistemlerin gelişimi için programlama becerisi ve dil bilgisi açısından alan uzmanı gereklidir.
- Bunun yanı sıra, kural tabanlı sistemler etki alanları arasında aktarılamaz.

5.2 AVT (NER) Sistemleri İçin Öğrenme Tabanlı Yöntemler

Makine öğrenimi, veriler hakkında en fazla verimli kararı veren karmaşık kalıpları veya dizi etiketleme algoritmalarını otomatik olarak öğrenme bilimi olarak ifade edilir. Bu yöntem öğrenme temelli yaklaşımlar olarak da ifade edilebilir. Bu yaklaşımlar denetimli öğrenme, yarı denetimli öğrenme (aktif öğrenme ve zayıf öğrenme dahil) ve denetimsiz öğrenme olarak kategorize edilebilir. Bu yaklaşımlar aşağıda farklı bölümler olarak detaylı bir biçimde açıklanmıştır.

5.2.1 Denetimli öğrenme (Supervised learning) tabanlı yöntemler

Denetimli öğrenme temelli yaklaşımlar, alan uzmanları (domain expert) veya kitle kaynak (crowdsourcing) platformları tarafından etiketlenmiş eğitim verilerini kullanarak bir belgeye açıklama (etiketleme) ekleyen yaklaşımlardır. Geleneksel denetimli öğrenme teknikleri, x girdilerinden y çıktılarına eşleşen bir işlev sağlar:

$$y = f(x) \quad (5.1)$$

Denetimli öğrenme tabanlı yöntemler eğitim (train) ve tahmini etiketleme (annotation) olmak üzere iki adımdan meydana gelirler (Campos, Matos and Oliveira, 2012).

Eğitim verileri, denetimli öğrenme algoritmaları için ön koşuldur. Makina öğrenmesi (ML) modeli, alan uzmanları veya kitle kaynak platformu aracılığıyla etiketlenmiş altın standart veri kümesini bir makina öğrenme algoritması kullanarak eğitir. Bu adım, modelin karmaşıklığına ve mevcut hesaplama kaynaklarına bağlı olarak biraz zaman alabilir. Daha sonra eğitilen bu model kullanılarak herhangi bir veriye tahmini olarak varlık adları sağlayarak otomatik olarak etiketleme yapar.

Denetimli öğrenme için AVT sistemlerinde Koşullu Rastgele Alanlar (CRF'ler) (Lafferty, McCallum and Pereira, 2001), Gizli Markov Modelleri (HMM'ler) (Baum and Petrie, 1966) ve Maksimum Entropi Markov Modelleri (MEMM'ler) (McCallum, Freitag and Pereira, 2000) olmak üzere farklı denetimli modelleri kullanılmıştır. CRF'ler, diğer yöntemlere göre çeşitli avantajlar sundukları için son yıllarda aktif olarak kullanılmaktadır. En popüler denetimli öğrenme algoritmaları şunlardır;

SVM (Destek Vektör Makinesi): Sınıflandırma ve regresyon problemleri için kullanılan doğrusal bir modeldir. Doğrusal ve doğrusal olmayan sorunları çözebilir ve birçok pratik sorun için iyi çalışır. SVM algoritması, verileri sınıflara ayıran bir çizgi veya hiper düzlem oluşturur. SVM'lerin yaptığı, iki sınıfın verileri arasında bir ayırma çizgisi (veya hiper düzlem) bulmaktır. Verileri girdi olarak alır ve mümkünse bu sınıfları ayıran bir satır çıktısı veren bir algoritmadır (Cherkassky, 1997; Islam et al., 2017).

KNN (K-En Yakın Komşu): En basit tembel makine öğrenme algoritmalarından biridir. Algoritmanın amacı, belgeleri makine öğrenimi tarafından oluşturulan bir örnek grubun önceden tanımlanmış sınıflarından birine sınıflandırmaktır. Algoritma, sınıflandırma yapmak için eğitim verilerinin kullanılmasını gerektirmez, eğitim verileri test aşamasında kullanılabilir. KNN, karşılıklı Öklid mesafesi hakkında örnek gruplarından en benzer belgeleri bulmaya dayanır (Trstenjak et al., 2014).

Rastgele Orman (Random Forest): 2001 yılında L. Breiman tarafından önerilen rastgele orman algoritması, genel amaçlı bir sınıflandırma ve regresyon yöntemi olarak son derece başarılı olmuştur. Birkaç rastgele karar ağacını birleştiren ve tahminlerini ortalama alarak toplayan yaklaşım, değişken sayısının gözlem sayısından çok daha fazla olduğu ortamlarda mükemmel performans göstermiştir (Biau and Scornet, 2016).

5.2.2 Yarı Denetimli (Semi Supervised) öğrenme, Aktif öğrenme ve Zayıf (Weak) öğrenme yaklaşımları

Yarı denetimli öğrenme, özel bir öğrenme şeklidir. Geleneksel sınıflandırıcılar, önemli miktarda açıklamalı eğitim verisi gerektirir. Özellikle biyomedikal alanda eğitim verilerini etiketleme görevi uzman kişi çabası gerektirdiğinden pahalı, zor ve zaman alıcı bir iştir. Bu nedenle yarı denetimli öğrenme genellikle etiketlenmiş verinin az, etiketlenmemiş verinin ise fazla olduğu durumlarda kullanılan bir yöntemdir. Sorunu hem etiketli hem de etiketsiz verileri kullanarak ele alır (Goyal, Gupta and Kumar, 2018). Etiketlenmiş veriler bir denklem ile ifade edilirse (Perera, Dehmer and Emmert-Streib, 2020);

$$X(x_1, x_2, \dots, x_n) \rightarrow L(l_1, l_2, \dots, l_n) \quad (5.2)$$

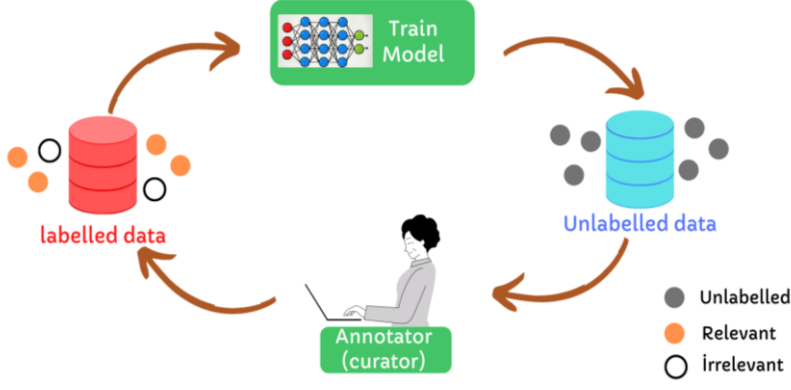
burada “X” veri kümesini, “L” ise etiket kümesini ifade eder , Buradaki amaç görev ;

$$Y(y_1, y_2, \dots, y_m) \rightarrow L(l_1, l_2, \dots, l_m) \quad (5.3)$$

ifadesini doğru şekilde eşleyen bir model geliştirmektir. Burada $m > n$ ve Y, etiketlerle eşlenmesi gereken etiketlenmemiş veri kümesidir.

Bu yöntemler, etiketlenmemiş verileri etiketlemek için "tohum (seed)" adı verilen az sayıda eğitim örneği kullanır. Elde edilen sonuçlar daha çok etiket sağlanması amacıyla tekrar eğitilmek için kullanılır. Bu süreç, öğrenme kararlarını rafine hale getirmek için birkaç kez devam eder. En popüler yöntem, birçok araştırmacı tarafından kullanılan ve popülerlik kazanan “bootstrapping (önyükleme)”dir (Bhagavatula, Gsk and Varma, 2012). Bunun dışında bir diğer yöntem olan aktif öğrenmede aslında bir çeşit yarı denetimli öğrenme sayılabilir.

Aktif Öğrenme: Bu teknik veri etiketleme sürecini en aza indiren ve model performansını en üst düzeye çıkaran yarı denetimli bir öğrenme tekniğinin farklı bir biçimi olup, veri etiketlemek/üretmek için makine öğrenmesi ve insan etkileşimini birleştiren bir tekniktir. Aktif öğrenme algoritması, hangi örneklerin etiketleneceğini ve eğitim setine ekleneceğini seçebilir. Yani, modelin yeniden eğitilmesi üzerinde en büyük etkiye sahip olacak öğeleri açıklama için seçerek yapılması gereken etiketleme sayısını azaltır.



Şekil 5.1. Aktif Öğrenme Döngüsü

Şekil 5.1'de gösterildiği gibi, genellikle bir sınıflandırıcı küçük bir etiketli örnek kümesiyle başlar, etiketlenmemiş bir depodan birkaç bilgilendirici örnek seçer ve bir kahinden (uzman kişi, ya da kitle kaynak) etiketleri sorgular. Bu süreç, yinelenmeli olarak etiketlenmiş verileri eklemek ve bir modeli yeniden eğitmektir. Buradaki genel amaç, bir sınıflandırıcıya yedek veri sağlamak ve yüksek performanslı doğru bir model eğitmektir.

Zayıf Denetimli (Weak Supervised) Öğrenme: Zayıf denetim, çoğunlukla düşük kalitede veya zayıf olan çeşitli veri kaynaklarından gelen bilgileri harmanlayan bir öğrenme tekniğidir (Zhou, Z. H., 2018). Atif öğrenmede ile farkı şu şekildedir;

- Aktif öğrenmede veriler insan tarafından etiketlenirken zayıf öğrenmede uzman olmayan kişilerden ve daha birçok farklı yerden gelen daha ucuz ve daha düşük kalitede etiketli veriler olabilir.
- Aktif öğrenmede etiketlerin mutlaka doğru olduğu varsayılırken zayıf öğrenmede etiketler mutlaka çok doğru veya eksiksiz değildir.
- Atif öğrenmede veriler yalnızca bir kaynaktan etiketlenirken zayıf öğrenmede birden fazla veri kaynağı kullanılmış olabilir.

5.2.3 Denetimsiz (Unsupervised) öğrenme

Denetimsiz öğrenme, sınıflandırma ve etiket bilgileri kullanamayan bir algoritmadır. Karar vermek için yalnızca etiketlenmemiş verileri kullanan sistemlerdir. Bu yöntemin amacı, verilerin yapısal ve dağıtım (distributional) özelliklerini dikkate alan bir öğrenme modeli oluşturmaktır (Bose *et al.*, 2021).

Bu yöntemlere kümeleme (clustering) ve birliktelik kurallarına dayalı (association rules-based) yaklaşımlar örnek olarak verilebilir. Kümeleme tabanlı yaklaşım, bağlam benzerliğini kullanarak etiketlenmemiş verilerden adlandırılmış varlıkları çıkarmak için dağıtım istatistiklerini kullanır. Birliktelik kurallarına dayalı teknik ise büyük veritabanlarındaki öğeler arasındaki ilişkileri bulmakla ilgilidir.

5.2.4 Biyomedikal alanda yapılmış olan varlık etiketli derlemeler

Biyomedikal literatürde makina öğrenmesi algoritmaları için eğitim verisi olarak kullanılacak birçok altın standart derlem mevcuttur. Bu derlemeler bir tablo halinde Çizelge 5.3’de gösterilmiştir. Çizelge incelendiğinde gen çeşitleri üzerine 3 adet altın standart derlemin olduğu ve bu derlemlerde en az 20 bin adet cümlelerin etiketlendiği görülmektedir. Bunun dışında hastalıklar ile ilgili en az 3 binin üzerinde cümlelerin etiketlendiği altın standart derlemlerin olduğu ve vücut anatomisine ait en az 4 binin üzerinde cümlelerin etiketlendiği derlemlerin olduğu görülmektedir. Tür ve kimyasal varlıklar içinde birden fazla derlemin olduğu görülmektedir.

Çizelge 5.3. Gen, Protein, Tür, Hastalık, Kimyasallar, Anatomi ve Hücre varlıklarının etiketlendiği derlemler ve sayıları

Varlıklar	Derlem	Metin Tipi	Etiketlenen (Cümle)	Referans
Gene ve Protein	GENETAG	Cümle	20.000	(Tanabe <i>et al.</i> , 2005)
	JNLPBA	Özet	22.402	(Zhu <i>et al.</i> , 2018)
	PennBioIE	Özet	≈22.877*	(Mandel, 2006)
Türler	OrganismTagger	Tam metin	9.863	(Naderi <i>et al.</i> , 2011)
	Corpus Linnaeus	Tam metin	19.491	(Naderi <i>et al.</i> , 2011)
	Corpus SCAI Disease	Özet	≈3.640*	(Wissler <i>et al.</i> , 2014)
Hastalıklar	EBI Disease	Cümle	600	(Hahn <i>et al.</i> , 2012)
	BioText	Özet	3.655	(Prabavathy and Sumathi, 2014)
	SCAI IUPAC	Cümle	20.300	(Klinger <i>et al.</i> , 2008)
Kimyasallar	SCAI General	Cümle	914	(Sammartino, Krallinger and Valencia, 2016)
Anatomi Çeşitli	AnEM	Cümle	4.700	(Ohta <i>et al.</i> , 2012)
	CellFinder	Tam metin	2.100	(Stachelscheid <i>et al.</i> , 2014)

5.3 İlişki Çıkarımı İçin Yapılan Çalışmalar

Biyomedikal ilişki sınıflandırma iki kategoriye ayrılmaktadır.

- Örüntü tabanlı yaklaşımlar (pattern-based approaches)
- Makine öğrenmesi temelli yaklaşımlar (machine learning-based approaches)

Örüntü tabanlı yaklaşımlar bu alandaki geleneksel yöntemlere girmektedir. Biyomedikal varlıklar arasındaki anlamsal ilişkiyi tanımlamak için belirli model türlerini ve eşleştirme kurallarını kullanır. Örüntü tabanlı yaklaşımlar genel olarak etiketli veriye ihtiyaç duymazlar ancak örüntü formları (pattern form) veya elle kodlanmış kuralları (hand-coded rules) işlemek ve tasarlamak için bir biyomedikal alan uzmanına ihtiyaç duyarlar. Örüntü tabanlı yaklaşımların asıl amacı örüntülerin veya kuralların kalitesi ve sayıdır. Ancak önceden tanımlanmış (predifined) örüntüler veya kurallar genellikle serbest metindeki ifade varyasyonlarına uyum sağlayamadığından, örüntü tabanlı yaklaşımlarda düşük duyarlılık oranı problemi mevcuttur.

Bir diğer yöntem denetimsiz öğrenme yaklaşımıdır. Denetimsiz öğrenmede etiketli eğitim verileri gerekli değildir. Bu bakımdan iyi görünmesine rağmen sistemin performansı büyük ölçüde kısıtlamaların ve buluşsal yöntemlerin ne kadar iyi yapılandırıldığına bağlı olmasına dayanmaktadır. Son zamanlardaki ilişki çıkarma yaklaşımlarının çoğu, denetimli makine öğrenimi teknikleri kullanılarak etiketlenmiş derlemeler üzerinden eğitilmiştir ve böylece oldukça yüksek performanslar elde edilmiştir. Bu yöntemlerin tek sınırlaması, etiketli verileri toplamanın ve açıklama eklemenin kapsamlı ve sıkıcı bir görev olmasıdır. Bunların dışında son zamanlarda denetimli yöntemlerin bir türevi olan zayıf denetimli ve uzaktan denetimli yöntemlerde ilişki çıkarımı için sıklıkla kullanılmaktadır.

Bir alternatif olarak makina öğrenmesi tabanlı yaklaşımlar biyomedikal ilişki sınıflandırmasını denetimli bir öğrenme sınıflandırma görevi (supervised learning classification task) olarak ele alır. Makina öğrenmesi tabanlı yaklaşımlar 3 alt grupta kategorize edilebilir. Bunlar;

- Öznitelik tabanlı (feature-based) yaklaşımlar
- Çekirdek tabanlı (kernel-based) yaklaşımlar ve
- Son sinir ağı tabanlı (recent neural network-based) yaklaşımlardır.

Öznitelik tabanlı yaklaşımlar için çeşitli sözcüksel (lexical) ve sözdizimsel (syntactic) öznitelikler uzmanlar tarafından tanımlanır ya da çıkarılır ancak bu çok beceriye dayalı bir iştir. Sözdizimsel ayrıştırma ağacı (syntactic parse tree) veya bağımlılık grafi (dependency graph), biyomedikal ilişki sınıflandırması için değerli bilgiler taşır. İyi tasarlanmış bir çekirdek işlevine (well-designed kernel function) dayanan çekirdek tabanlı (kernel-based) yaklaşımlar, sözdizimsel ağaçları (syntactic trees) veya bağımlılık graflarını (dependency graphs) yüksek boyutlu (high-dimension) bir öznitelik alanına (feature space) eşleyebilir (map). Örneğin biyomedikal ilişki sınıflandırması için graf çekirdekleri (graph kernels alt diziler çekirdekleri (subsequences kernels) ve sığ dilbilimsel çekirdekler (shallow linguistic kernels) gibi yaklaşımlar önerilmiştir. Çekirdek tabanlı yaklaşımlar, öncelikle yapı verileri (structure data) arasındaki benzerliği (similarity) etkili bir şekilde hesaplayan çekirdek işlevlerine (kernel functions) dayanır. Bununla birlikte, uygun çekirdek işlevlerini (suitable kernel functions) tasarlamak, genellikle öznitelikleri seçmek (choosing) veya oluşturmaktan çok daha zordur.

5.3.1. Biyomedikal ilişki sınıflandırmada mevcut derlemler

Biyomedikal alanda ilişki çıkarımı üzerine birçok derlem mevcuttur. Biyomedikal ilişki sınıflandırması için mevcut ana derlemler Çizelge 5.4'de listelenmiştir. Özellikle protein-protein etkileşimleri (PPI) üzerine birçok derlem hazırlanıldığı görülmektedir.

Çizelge 5.4. Biyomedikal ilişki sınıflandırmada mevcut derlemler

Derlem	İlişki	Açıklama	URL
IEPA	PPI	486 cümle, 335 olumlu örnek ve 482 olumsuz örnek dahil 817 ikili PPE ilişkisi	http://mars.cs.utu.fi/PPICorpora
HPRD50	PPI	145 cümle, 163 olumlu örnek ve 270 olumsuz örnek dahil 433 ikili PPE ilişkisi	http://www.hprd.org
LLL	PPI	77 cümle, 164 olumlu örnek ve 166 olumsuz örnek olmak üzere 300 ikili PPE ilişkisi	http://genome.jouy.inra.fr/texte/LLLchallenge

DDI 2011	DDI	DrugBank'tan 579 belge, 3158 pozitif örnek ve 27.695 negatif örnek dahil 30.853 ikili DDI	http://labda.inf.uc3m.es/DIEExtraction2011/dataset.html
DDI 2013	DDI	DrugBank'tan 792 belge, 233 Medline özeti, 18,502 farmakolojik madde ve hem farmakokinetik hem de farmakodinamik etkileşimler dahil 5028 çok sınıflı DDI.	http://labda.inf.uc3m.es/doku.php?id=en:labda_ddicorpus
CDR	CID	1500 PubMed makalesi, 4409 açıklamalı kimyasal, 5818 açıklamalı hastalık ve 3116 CID	http://www.biocreative.org

5.4 Adlandırılmış Varlık Tanıma ve İlişki Çıkarımı İçin Son Teknoloji Yöntemler: Derin Sinir Ağı ve Dönüştürücü (Transformer) Tabanlı Öğrenme Modelleri

Makine öğreniminin yeni bir dalı olan derin sinir ağları (deep neural networks) , son zamanlarda doğal dil işleme ve bilgisayarla görme (computer vision) gibi yapay zekâ alanlarında oldukça başarılı bir potansiyel göstermiştir. Sinir ağı tabanlı yaklaşımlar, veri odaklı bir yöntemdir ve uzmanların kalıpları, öznitelikleri ve çekirdek işlevlerini dikkatlice oluşturmaya gerek kalmadan etiketli eğitim verilerinden gizli öznitelik (latent feature) temsilini doğrudan öğrenebilir. Sinir ağı temelli yaklaşımlar, biyomedikal ilişki sınıflandırmasında baskın bir yöntem haline gelmiş ve giderek artan bir ilgi görmüştür.

Sinir ağı tabanlı yaklaşımların çoğu denetimli (supervised) yöntemlerdir ve sinir modellerini eğitmek için etiketli verilere (labeled data) ihtiyaç duyarlar. Bu nedenle, biyomedikal ilişki sınıflandırmasında sinir ağı tabanlı yöntemler kullanılacaksa bu yöntemlerde etiketli derlemeler önemli bir rol oynar. Kamuya açık (publicly available) derlemeler, performans değerlendirmesi ve karşılaştırma yapabilmek için önemlidir.

Derin sinir ağı ile ilgili yaygın kullanılan yöntemler şunlardır.

- İleri beslemeli sinir ağı (feedforward neural network, FFNN) (Furrer *et al.*, 2019)
- Tekrarlayan sinir ağları (recurrent neural networks, RNN) (Zhang *et al.*, 2018)

- Evrişimsel sinir ağları (convolution neural networks, CNN) (Zhu *et al.*, 2018)
- Son bir kaç yılda ise transformer tabanlı dil modeline dayalı yaklaşımlar özellikleri temsil etmek için bağlamsal bilgiler hesaba katılarak kullandığından, birçok NLP görevinde CNN ve RNN teknolojilerinden daha gelişmiş performans göstermiştir. Bu yöntemlerde girdiler her örnek için iyi tanımlandığından, aynı zamanda denetimli bir öğrenme modelidir. Özellikle Google'ın 2018 yılında çıkardığı bir dil modeli olan BERT (Bidirectional Encoder Representations from Transformers) (Devlin *et al.*, 2018), doğal dil işlemede devrim niteliğinde bir gelişme olmuştur.

Bu devrimsel teknolojinin çıkmasına yol açan kilit süreçlere tarihsel olarak incelenecek olunursa;

TF-IDF: Her bir belgedeki her terimin (veya kelimenin) ağırlığının belirlenmesini sağlayan sayısal bir istatistik yöntemidir. Yöntem, belge toplamada terimlerin (veya kelimelerin) önemini değerlendiren ağırlığı belirler. Belgelerde geçen metin sayısı ile orantılı olarak metnin önemi artmaktadır (Zhang *et al.* 2011).

2013 yılında: Bu devrimsel teknolojinin ilk yolu “Vektör Uzayında Kelime Gösterimlerinin Verimli Tahmini” adında Word2Vec makalesi düşünülebilir Word2Vec (Mikolov *et al.*, 2013). Bu makalede kelimelerin anlamsal anlamını ve benzerliğini daha doğru bir şekilde belirlemek için kelimeler vektör uzayında ifade edilmeye çalışmıştır. Birbirine benzer kelimeler uzayda birbirine daha yakın durmaktadır ve aralarındaki benzerlik esas alınarak bir tahmin yürütülmektedir.

2015 yılında "Sequence to Sequence Learning with Neural Networks " makalesinde seq2seq yaklaşımı geliştirilmiştir (Sutskever, Vinyals and Le, 2015). Seq2seq model, Tekrarlayan Sinir Ağları (RNN'ler) adı verilen bir teknik kullanmıştır. Bu teknik Kodlayıcı (Encoder)- Çözücü (Decoder) mimarisi üzerine kuruludur. Bu mimarinin ortaya çıkışıyla beraber doğal dil işleme görevleri performanslarında başka bir devrim yaşandı denebilir.

- **Kodlayıcı (Encoder):** Bu mimari ilk olarak giriş metnini işler, önemli kısımları arar ve cümledeki diğer kelimelerle alaka düzeyine göre her kelime için bir gömme oluşturur.

- **Kod Çözücü (Decoder):** Burası ise, bir gömme olan kodlayıcının çıktısını alır ve ardından bu gömmeyi tekrar bir metin çıktısına, yani girdi metninin çevrilmiş versiyonuna dönüştürür.

2017 yılında yayınlanan “Attention is All You Need (İhtiyacınız Olan Herşey Dikkat)” (Vaswani et all, 2017) makalesi doğal dil işleme üzerine “Dönüştürücü (Transformer)” olarak adlandırılan dikkat (attention) mekanizmasını tanıtmıştır. RNN mimarisi her ne kadar bir devrim isede uzun girdilerde sorunlar yaşamaktaydı. Özellikle uzun cümlelerin başka dillere çevrildiği problemlerde bu cümledeki ilk kelimelerin önemini ister istemez azaltabiliyordu. Dikkat (Attention) adı verilen ve Encoder’larda bütün bilginin sabit uzunluktaki bir vektörle ifade edilmesi ile bu hatırlama problemi kısmen de olsa ortadan kaldırılmıştır.

2018 yılında Google, belirli makine öğrenimi görevlerinde ince ayar (fine tune) yapmak için kullanılabilecek önceden tanımlı (pre-defined) bir dil modeli olan BERT dil modelini geliştirmiştir (Devlin *et al.*, 2018). BERT model mimarisi, çok katmanlı çift yönlü bir dönüştürücüye (transformatör) dayanmaktadır. Önceki dil modelleri, iki tek yönlü dil modelinin (yani soldan sağa ve sağdan sola) bir kombinasyonu ile sınırlı iken BERT, bir dizideki rastgele maskelenmiş kelimeleri tahmin eden maskeli bir dil modeli kullanmış ve bu nedenle çift yönlü temsilleri öğrenmek için kullanılmıştır. Bu yapısı nedeniyle BERT, AVT, ilişki çıkarımı, metin sınıflandırma gibi birçok doğal dil işleme görevinde oldukça başarılı sonuçlar üretmiştir. Bu da onu son yılların en popüler tekniklerinden biri haline getirmiştir.

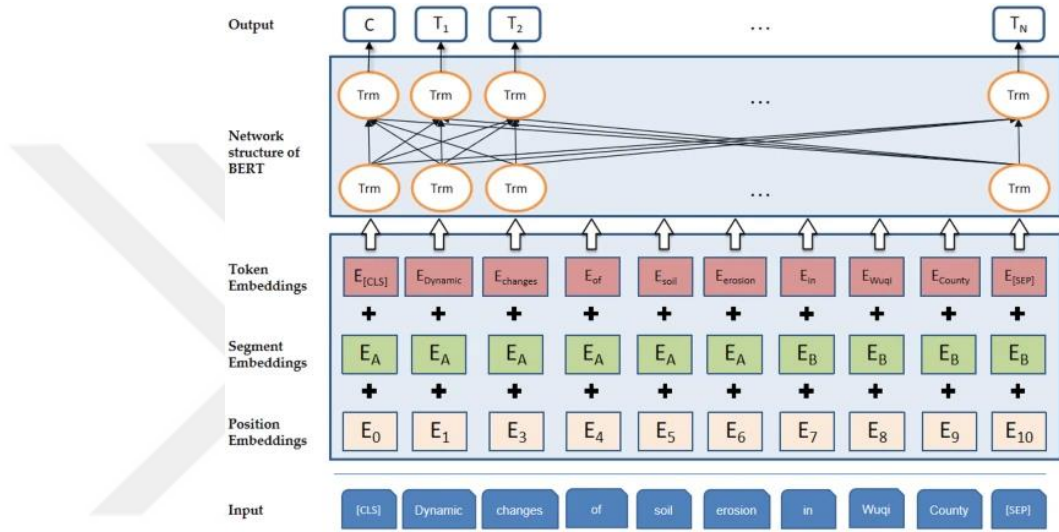
BERT’in önemi: Vektör gömme tabanlı yöntemlerden Word2vec veya türevi GloVe gibi bağlamdan bağımsız modeller (Context-free models), kelime dağarcığındaki her kelime için tek bir kelime gömme temsili oluşturur. Örneğin aşağıdaki cümleyi incelenirse;

“Bugün **yüz** sayfa kitap okuduktan sonra elimi **yüzümü** yıkadım ve dışarıda koyunların derilerini **yüzen** bir adam gördüm ve hemen oradan uzaklaşarak **yüzmeye** gittim.”

Bu kelime Word2vec ile temsil edildiğinde sayı olan yüz, çehre olan yüz, deri soyma eylemi olan yüz ve spor aktivitesi olan yüz aynı temsile sahip olacaktır. Bunun yerine bağlamsal modeller, cümledeki diğer kelimelere dayanan her kelimenin bir temsilini oluşturur. Bu nedenle BERT sayı olan yüz, çehre olan yüz, deri soyma eylemi olan yüz ve spor aktivitesi olan yüz için ayrı ayrı temsiller sağlar.

BERT, bağlamsal bir model olarak bu ilişkileri çift yönlü bir şekilde yakalar. Word2vec kelime bazlı öğrenme ve Elmo karakter bazlı temsil yöntemini kullanırken BERT, girdiyi alt sözcükler olarak temsil eder ve alt sözcükler için gömmeleri öğrenir. Örneğin BERT “playing” kelimesini “play” ve “##ing” olarak böler. Bu yöntem kelime ve karakter tabanlı yöntemlere göre daha başarılı sonuçlar vermiştir (Kim and Lee, 2020).

BERT, ingilizce Wikipedia ve BooksCorpus veri kaynakları kullanılarak eğitilen bir dil modelidir.



Şekil 5.2. Genel bir Bert mimarisi (Devlin et al., 2018)

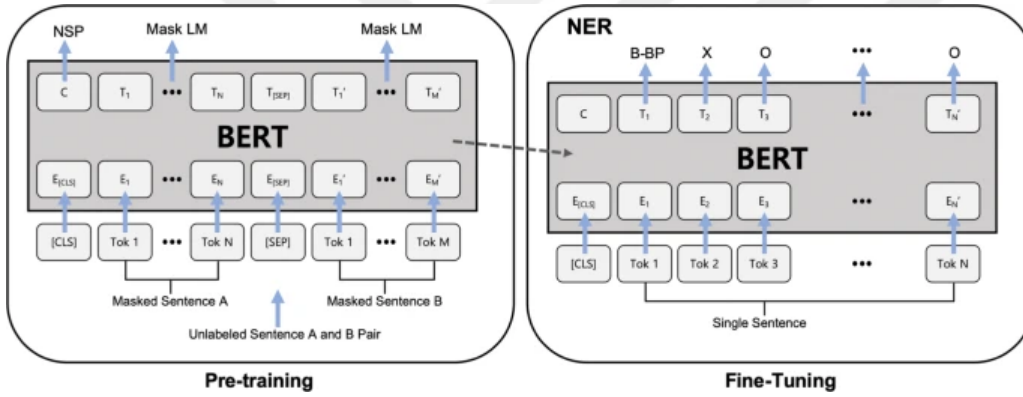
BERT modelinin genel yapısı Şekil 5.2'de gösterilmiştir. Bu şekil incelendiğinde burada “En”, kelimelerin kodlanmış temsilini, “Trm” transformatör yapısını ve “Tn” ise eğitilmiş kelime vektörünü gösterir. BERT modeli, herhangi bir metni girdi olarak alır. BERT'in girdi kelime vektörü, üç vektör eklenerek elde edilir: belirteç gömme (token embedding), segment gömme (segment embedding) ve konum gömme (position embedding).

- **Belirteç gömme (token embedding):** Giriş sözcüklerini farklı belirteçlere (token) ayırır ve metin örneğinin başlangıcını ve cümlelerin sonunu belirtmek için [CLS] ve [SEP] olmak üzere iki özel sembol ekler.
- **Segment gömme (segment embedding):** İki cümleyi ayırt etmek için kullanılır.
- **Konum gömme (position embedding):** Bir kelimenin konum bilgisini temsil eder.

BERT Kurulumu: BERT, " $x_1, \dots, x_N$ " ve " $y_1, \dots, y_M$ " olmak üzere iki segmentin (belirteç dizileri) bir birleşimini girdi olarak alır. Segmentler genellikle birden fazla doğal cümleden oluşur. İki segment, onları sınırlayan özel belirteçlerle BERT'e tek bir giriş dizisi olarak sunulur: "[CLS], $x_1, \dots, x_N$, [SEP], $y_1, \dots, y_M$, [EOS]". M ve N, $M+N < T$ olacak şekilde sınırlandırılır, burada T, eğitim sırasında maksimum dizi uzunluğunu kontrol eden bir parametredir.

BERT Mimarisi: BERT (base) modeli toplamda 12 transformer bloğu 768 gizli katman 12 dikkat bloğu ve 110 milyon parametreden oluşur.

İnce ayar (fine tune): Temel olarak Şekil 5.3'de gösterildiği gibi BERT dil modeli ince ayarlanarak doğal dil işlemede soru cevaplama, adlandırılmış varlık tanıma, ilişki çıkarımı ve daha birçok belirli görevler için kullanılabilir. İnce ayar önceden eğitilmiş sinir ağının ağırlıklarının ayarlanması anlamına gelmektedir.



Şekil 5.3. Adlandırılmış varlık etiketleme için ince ayarlanmış (fine tuned) BERT modeli

Batch boyutu (Batch_size): Bir ağ üzerine yayılacak örnek sayısını tanımlar. Örneğin, 1050 eğitim örneğiniz olduğunu ve 100'e eşit bir batch boyutu ayarlamak istediğinizi varsayalım. Algoritma, eğitim veri kümesinden ilk 100 örneği (1'den 100'e kadar) alır ve ağı eğitir. Ardından, ikinci 100 örneği (101'den 200'e) alır ve ağı yeniden eğitir. Tüm örnekleri ağ üzerinden yayana kadar bu prosedürü yapmaya devam eder. Son örnek seti 100 ile kalansız bölünemeyen 1050 sayısı ise yalnızca son 50 örneği alır ve ağı eğitir.

Batch boyutu < Tüm örneklerin sayısından küçük olmasının avantajları;

- **Daha az bellek:** Daha az bellek gerektirir. Ağı daha az örnek kullanarak eğittiğiniz için, genel eğitim prosedürü daha az bellek gerektirir.
- **Daha hızlı çalışma:** Genellikle ağlar mini partilerle daha hızlı çalışır. Bunun nedeni, her yayılımdan sonra ağırlıkları güncellemekten kaynaklanır.

Örneğin 11 grup oluşturduk (10 tanesi 100 örnek için ve 1 tanesi 50 örnek içermektedir) ve her birinden sonra ağımızın parametrelerini güncelledik. Yayılma sırasında tüm örnekleri kullansaydık, ağın parametresi için yalnızca 1 güncelleme yapardık.

BERT için standart yapılandırma ayarları şu şekildedir:

Batch boyutu : {16, 32}

Öğrenme oranı: $\{2^{-5}, 3^{-5}, 5^{-5}\}$

Öğrenme epogu: {2, 3, 4}

BERT birçok alana uyarlanmış ve DistilBERT (Sanh et al., 2019), RoBERT (Liu et al., 2019), BioBERT (Zhang et al., 2021), SciBERT (Beltagy, Lo and Cohan, 2019) ve ClinicalBERT (Huang et al., 2019) gibi farklı versiyonları çıkmıştır. Bu önceden eğitilmiş modeller, çok çeşitli NLP görevlerinde başarılı sonuçlar elde etmesine rağmen, yine de ince ayar gerektirmektedir. Bu versiyonlardan bazılarını incelersek;

RoBERTa: Önceden eğitilmiş dil modeli olan BERT modelindeki temel hiper parametreleri değiştirerek daha iyi performans elde etmeye yardımcı olan maskeli dil modelleme hedefini geliştirmesine olanak tanır. RoBERTa, BERT'den daha fazla veri ve daha uzun bir süre boyunca eğitilmiştir. BERT modelinden 4 kat daha fazla veri miktarı olan 160Gb metin ile eğitilmiştir. RoBERTa öğrenme epogunu 10 yapmanın dışında BERT ile aynı hiperparametreleri kullanır. RoBERTa modeli toplamda 12 transformer bloğu 768 gizli katman 12 dikkat bloğu ve 125 milyon parametreden oluşur.

BioBERT: BERT dil modelinin eğitim verisinin üstüne PubMed ve PMC derlemlerinin eklenmesiyle eğitilmiş bir dil modelidir. Özellikle biyomedikal alanda yapılan çalışmalar için kullanılan BERT'in bir versiyonudur.

ClinicalBERT: MIMIC-III veri tabanından alınan klinik metinler üzerinde eğitilmiş bir modeldir.

SCIBERT: Semantic Scholar külliyyatından alınan 1.14 milyon biyomedikal ve bilgisayar bilimi makalesinin tam metni konusunda eğitilmiştir.

5.5 Bitki ilaç etkileşimi ile ilgili Yapılan Çalışmalar

Biyomedikal literatürden bitki ve ilaç etkileşimlerini çıkarmak için bir takım mevcut çalışmalar bulunmaktadır. Ancak bu çalışmaların çoğunluğu gıda üzerine yapılmış çalışmalardan oluşmaktadır. Bitkiler kısmen gıdalar ile örtüşsede farklılıkları mevcuttur. Bu çalışmalar şöyledir.

Thierry Hamon ve ark. (2017) POMELO adını verdikleri “ilaç-gıda etkileşimi” derlemini geliştirmişlerdir. Bunun için 639 adet makale özetini MESH terimlerini aratarak biyomedikal literatürden çıkarmışlardır. Bu makalelerden ilaç ve gıda varlıklarını etiketlemişlerdir. Aralarında ilişki olabilecek önemli anahtar kelimeleri çıkartarak ilişki tespit çalışması yapmışlardır.

Gorjan Popovski ve ark. (2019) FoodBase adını verdikleri bir derlem geliştirmişlerdir. Bu derlemde içinde gıda varlıklarının yer aldığı 1000 adet öğünü kahvaltı, öğle yemeği, akşam yemeği ve içecekler olarak etiketlemişlerdir. Kural tabanlı bir AVT modeli geliştirerek verisetinden gıdaları otomatik olarak çıkarabilecek bir sistem tasarlamışlardır.

Riste ve ark. (2021) FoodNER adını verdikleri bir çalışma gerçekleştirmiştir. Bu çalışma Gorjan ve ark. yaptıkları kural tabanlı geliştirdikleri AVT modelinin gelişmiş bir çalışmasıdır. Bu çalışmada FoodBase derlemi BERT tabanlı AVT modeli ile ince ayarlanmıştır (fine tuned) ve %94.31 f-score (f-skor) başarı göstermiştir.

NA Deng ve ark. (2021) geleneksel Çin ilaçlarının (bitkisel ilaç ve ek gıdalar) hastalık ile olan etkilerini bulabilecek BİLSTM ve CRF modelinin birleşiminden oluşan bir model geliştirmişlerdir. Bu çalışma bitki adları, hastalık adı, sntomplar, tedavi edici özelliklerini bulmaya yönelik bir çalışmadan oluşmuştur ve %94,48 f-score başarı göstermiştir.

Kim ve ark. (2022) gıda ve ilaçları biyomedikal literatürden bulacak bir çalışma gerçekleştirmişlerdir. Çalışma daha çok gıda ve gıda bileşenlerini yakalamaya ve bunların ilaçlarla olan etkileşimlerini bulmaya yöneliktir.

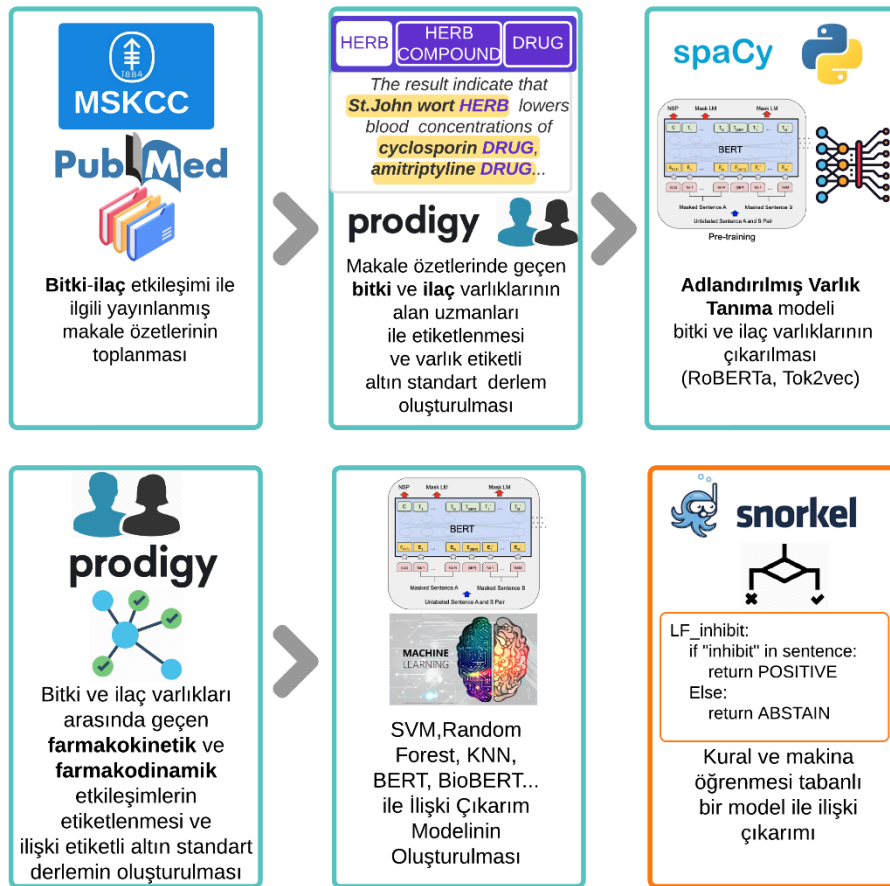
Trinh ve ark. (2018) arkadaşları PubMed biyomedikal literatürden bitki-ilâç etkileşimlerini tahmin edebilecek bir sistem geliştirmişlerdir. Bu çalışmada bitki verilerini TCM Database@Taiwan ve VietHerb sitelerinden, ilaçları ise Drugbank

veri kaynağından almışlardır. TCM Database@Taiwan ve VietHerb'de bulunan 10.061 bitkisel ilaç profilini 48.584 eş anlamlı ile birlikte çıkarmışlardır. İlaç ve bitki verileriyle ilgili makaleleri elde etmek için kural tabanlı bir yöntem kullanmışlardır.



6. MATERYAL ve YÖNTEMLER

Bu tez çalışmasının bu bölümünde biyomedikal metinlerde geçen bitki ve ilaç varlıklarını ve bu varlıklar arasında geçen farmakokinetik ve farmakodinamik etkileşimlerin makina öğrenmesi algoritmaları ile otomatik olarak yüksek doğruluk ve verimlilikle yakalanabilmesi ve çıkarılabilmesi amacıyla literatürün en büyük eksikliklerinden biri olan altın standart varlık ve ilişki etiketli derlemler hazırlanmıştır. Daha sonra bu derlemlerin eğitim veri kümesi olarak kullanılarak adlandırılmış varlık tanıma ve ilişki çıkarımı modelleri gerçekleştirme aşamaları detaylı bir biçimde anlatılmıştır. Şekil 6.1 bu boru hattının (pipeline) detaylı bir gösterimidir.



Şekil 6.1. Bitki ilaç etkileşimlerinin biyomedikal literatürden çıkarılması için oluşturulan boru hattı (pipeline) ve ayrıca kural tabanlı bir modelin oluşturulması

Çalışmanın ilk bölümü literatürde bitki-ilaç etkileşimlerine (BİE) ait yapılan çalışmaların biraraya getirilme sürecinden oluşmaktadır ve aşağıdaki iki aşamadan oluşmaktadır.

- İlk aşamada BİE'ye ait bilgilerin yer aldığı MSKCC veri kaynağında bulunan çeşitli bitki ve ilaç etkileşim bilgileri ve verilen PubMed referanslarından oluşan bitki ilaç etkileşimlerine ait makale özetleri çıkarılarak bir excel tablosu oluşturulmuştur. Bu tablo genel olarak bitki varlığı (terimi), ilaç varlığı ve bu varlıklar arasında geçen ilişki bilgisi ve bu etkileşimlerin referans olarak alındığı PubMed makale özetlerinden oluşmuştur.
- İkinci adımda bu tablo alan uzmanları tarafından filtrelenerek ilgisiz verilerin tablodan çıkarılması sağlanmıştır ve veriler ontoloji ve varlık bağlama düzeyinde analiz edilmiştir.

Çalışmanın ikinci bölümü adlandırılmış varlık tanıma sistemi için altın standart derlem ve model oluşturma sürecidir ve iki aşamadan oluşmuştur.

- İlk adımda tabloda yer alan yapılandırılmamış makale özetlerinden bitki (HERB) ve ilaç (DRUG) varlıkları iki alan uzmanı küratör tarafından aktif öğrenme tabanlı bir araç ile etiketlenerek varlık tanıma sistemleri için altın standart bir derlem oluşturma sürecinden oluşmuştur.
- İkinci adım ise bu altın standart derlemin son teknoloji doğal dil işleme yöntemlerinden önceden eğitilmiş dil modelleriyle (tok2vec, BERT) ince ayarlanarak yapılandırılmamış biyomedikal literatürden ilaç ve bitki varlıklarını otomatik olarak yakalayabilecek yüksek doğrulukta bir NER sistemi oluşturma sürecinden oluşmuştur.

Çalışmanın üçüncü bölümü ise ilişki çıkarım sürecidir ve iki aşamadan oluşmuştur.

- İlk aşama olarak bitki ve ilaçlar arasındaki ilişkiler için iki adet derlem hazırlanmıştır. İlk derlem bir bitki ve ilaç arasında bir ilişkinin olup olmaması ile ilgili etiketli verilerden oluşurken diğer derlem bitki ve ilaçlar arasındaki farmakokinetik ve farmakodinamik etkileşim tiplerinin etiketlendiği derlemden oluşmuştur.
- İkinci aşamada ise bu derlemler makina öğrenmesi algoritmaları (SVM, Rastgele orman ve KNN) ve doğal dil işlemenin son teknoloji yöntemlerinden BERT ve BioBERT ile ince ayarlanarak ilişkilerin otomatik olarak metinden çıkarılabilmesini sağlayan modeller üretilmiştir.

Çalışmanın dördüncü bölümü ise derlem oluşturmaktan daha farklı bir yol izlenerek ilişkilerin otomatik olarak etiketlenmesi için zayıf bir denetimli öğrenme çerçevesi kullanılmıştır. Bunun için kurallar oluşturulmuştur ve bu kurallar öznitelik olarak makina öğrenmesi modellerine verilerek ilişkilerin otomatik olarak etiketlenebilmesini sağlama çalışmaları yapılmıştır.

Son olarak bu modeller birbiri ile karşılaştırılarak en iyi sonucu veren model temel (baseline) model olarak ele alınmıştır.

6.1 Bitki -İlaç Etkileşimleri İçin Veri Seti

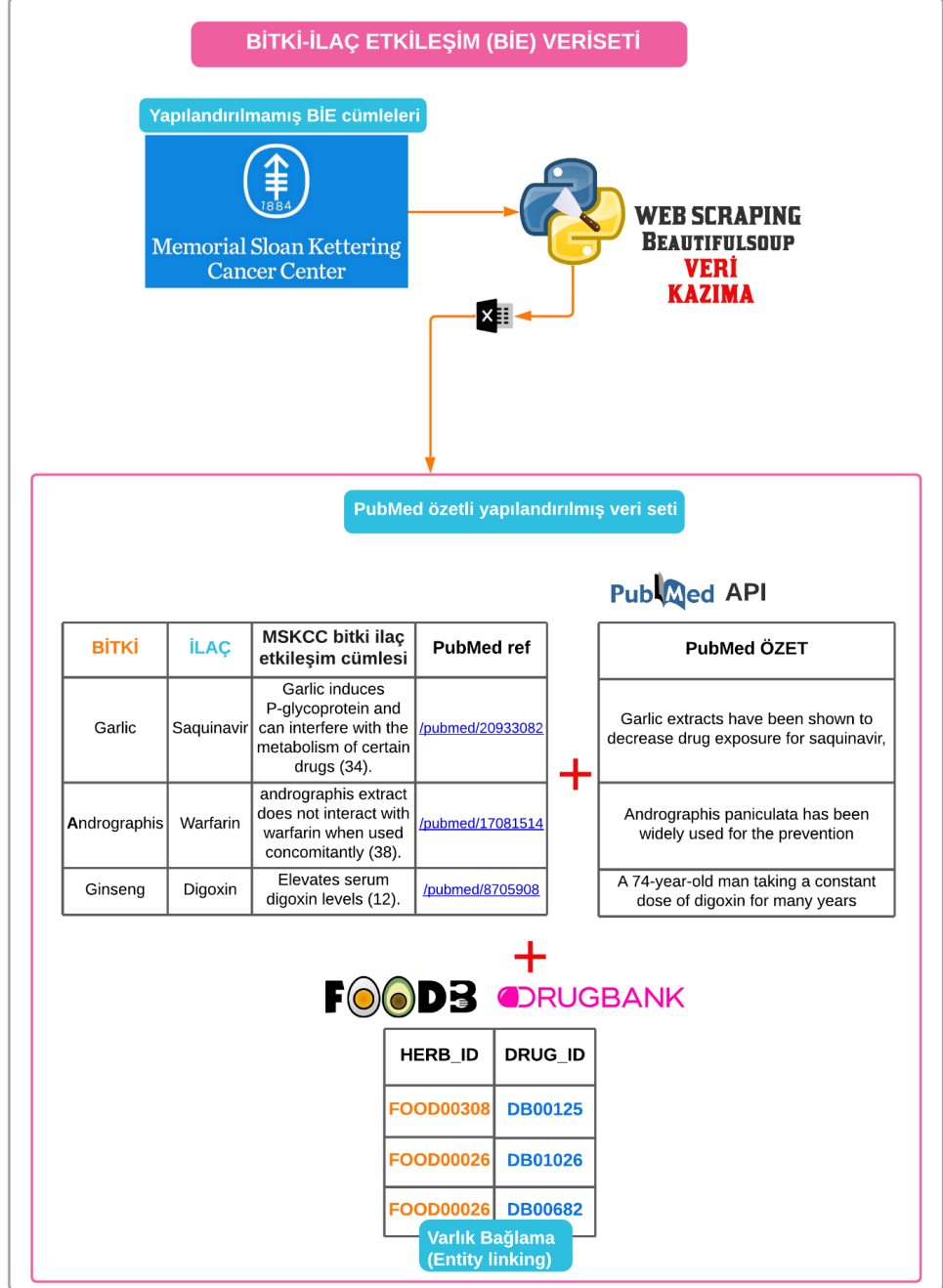
Derlem oluşturma süreci için bitki-ilaç etkileşimi ile ilgili bilimsel makaleler iki kaynaktan alınmıştır;

- I. Memorial Sloan Cancer Kettering (MSKCC)
- II. PubMed veri tabanı

MSKCC veri kaynağı 279 adet bitkiye ait bilimsel adı, fayda verdiği hastalık bilgisi, yan etki bilgisi, çalışma mekanizması, advers reaksiyonları, etkileştikleri ilaç bilgisi ve bilgilerin elde edildiği PubMed makale referansları gibi önemli bilgilerin yer aldığı bir kaynaktır.

Bu tez çalışması için MSKCC veri kaynağından alınan gerekli bilgiler BİE bilgisi ve bu etkileşim bilgilerinin yer aldığı referanslardır. Şekil 6.2’de MSKCC veri kaynağından bu bilgilerin nasıl alındığına dair detaylı bir tasarım mevcuttur. Bu bilgileri elde etmek için ilk olarak Python programlama diline ait bir kütüphane olan Beautiful Soup (*Beautifulsoup4*, 2022) veri kazıma (web scraping) yöntemi kullanılmıştır. Beautiful Soup kütüphanesi HTML veya XML dosyalarını işlemek için oluşturulmuş oldukça güçlü ve hızlı bir kütüphanedir. Bu kütüphane kullanılarak toplam 819 adet veri çıkarılmıştır. Bu verinin her satırında bitki adı, etkileştiği ilaç adı, iki varlık arasındaki MSKCC uzmanlarının manuel olarak yazdıkları etkileşim bilgisi ve bu etkileşim bilgisine atıfta bulunan referans bilgisi çıkarılarak bir tablo oluşturulmuştur. Etkileşim bilgilerinin yer aldığı referanslar PubMed makale referanslarından oluşmuştur. Bu referansların verdiği makale özetlerini çıkarmak için referanslardan elde edilen URL bilgisi kullanılarak beautifulsoup kütüphanesi ile makale özetleri de çıkarılmıştır ve tabloya eklenmiştir.

Elde edilen tablo iki alan uzmanı (Ege Üniversitesi biyokimya mühendisliğinden Prof. Dr. Figen Zihnioğlu ve Araş. Gör. Kerem Tok) tarafından incelenmiştir. Alan uzmanları ile incelenen MSKCC veri kaynağında Çizelge 6.1’de **Vitamin A, Vitamin C, Vitamin D, Vitamin B12, Vitamin, Magnezyum, Demir, Maitake, Reishi Mantar** ve buna benzer bazı varlıklar bitkisel ilaç kategorisine göre ele alındıkları gözlemlenmiştir. Alan uzmanları bu tür bitkisel ilaç olmayan varlıkları tablodan filtrelemiştir.



Şekil 6.2. MSKCC verisetinden bitkileri, ilaçları, bitki-ilaç etkileşim bilgisini ve PubMed referanslarını ve bu referanslara ait PubMed özetlerinin bir araya getirildiği veriseti. Bu veri setinde yer alan bitki ve ilaç varlıkları bilgi tabanları ile eşleştirilmiştir.

Çizelge 6.1. Alan uzmanları tarafından filtrelenen varlıklar

Vitamine A	Resveratrol	Proteolytic enzymes	Metabolic therapies	Hoxsey herbal therapy
Vitamine B12	Royal jelly	Quercetin	MGN-3	Hydrazine sulphate
Vitamine B6	Taurine	Reishi mushroom	MICOM	Indole-3-carbinol
Vitamine C	Tian xian	Sun farms vegetable soup	N-acetylcysteine	Inositol hexaphosphate
Vitamine D	Transfer factor	Superoxide dismutase	Nattokinase	Insulin potentiation therapy
Vitamine E	Usnic acid	SAM-e	Nicotinamide	Coenzyme Q-10
Vitamine O	Viacreme	Scorpion venom	Omega-3	Colloidal silver

MSKCC Tablo Analizi: Tablo haline getirilen veri setinde yer alan ilaç ve bitki varlıklarını otomatik olarak yakalayabilecek ontoloji ve bilgi tabanı araştırması yapılmıştır. Bunun için bilgi tabanı seviyesinde ve ontoloji seviyesinde araştırmalar yapılmıştır. Bilgi tabanı seviyesinde iki veri tabanı kullanılmıştır;

- **FOODB (bitkiler için):** FooDB veritabanı, gıda bileşenleri, kimya ve biyoloji konusunda kapsamlı bir kaynaktır. Gıdalara ait bilimsel ad, tanım, ID, gıda grubu, Wikipedia ID ve bileşikleri hakkında detaylı bilgilerin olduğu bir kaynaktır. Ayrıca gıdalara lezzetini, rengini, tadını, dokusunu ve aromasını veren birçok bileşen de dahil olmak üzere hem makro besinler hem de mikro besinler hakkında da bilgi sağlar(FooDB, 2022).
- **DrugBank (ilaçlar için):** DrugBank veritabanı ilaçlar (kimyasal, farmakolojik ve farmasötik) , ilaç hedefleri ve proteinler hakkında ayrıntılı bilgi sağlayan kapsamlı, erişimi ücretsiz, çevrim içi bir veri tabanıdır (DrugBank , 2022).

Tablomuzda yer alan bitki isimleri FooDB ve ilaç isimleri ise DrugBank bilgi tabanları ile eşleştirilmiştir. Eşleştirilen veriler incelendiğinde tablomuzda bulunan birçok bitki ve ilaç ismi ile yeterince eşleşme olmadığı gözlemlenmiştir. Bunun nedenleri:

- FooDB’de bulunan varlıkların daha çok gıda temelli olduğu gözlemlenmiştir.
- Özellikle metin içerisinde geçen iki veya daha fazla kelimedenden oluşan bitki ve ilaç isimlerinin bilgi tabanları ile eşleşmediği gözlemlenmiştir.
- Bir diğer neden bir bitkiye ait makale içinde bilimsel adlandırmalar veya eş anlamlılar kullanıldığından bu adlandırmalar yakalanamamıştır. Örneğin Danshen yerine *Salvia miltiorrhiza*’nın kullanılması
- Bazı bitki adlandırmaları için kısaltmalar kullanılmıştır ve bu nedenle eşleşme olmamıştır. Örneğin *Phyllanthus emblica* yerine *P.emblica* gibi bir kısaltma kullanılmıştır.

Bunun dışında ontoloji seviyesinde de araştırmalara yapılmıştır ve bioportal ontolojileri (örneğin SNOMED, FOODON vb.) araştırılmıştır. Bioportal birçok medikal ontolojinin yer aldığı bir sitedir. Ancak bu ontolojilerin de bitki ve ilaç varlıkları ile başarılı bir şekilde eşleşmedikleri gözlemlenmiştir. Hem ontoloji seviyesinde hem de bilgi tabanı seviyesinde olan araştırmalar bir excel tablosuna detaylı bir şekilde eklenmiştir.

Bilgi tabanlarının ve ontolojilerin bitki ve ilaç varlıkları ile eşleşme problemini çözebilecek makina öğrenmesi tabanlı akıllı sistemlere ihtiyaç duyulmuştur. Bu akıllı sistemler için eğitim verisi olarak kullanılacak derlem gerekmektedir.

Bu nedenle bu çalışmada MSKCC veri kaynağının verdiği PubMed referanslardan elde edilen bitki ilaç etkileşimi ile ilgili makale özetleri (içindeki bitki ve ilaç varlıklarının ve bu varlıklar arasındaki ilişkilerin etiketlenmesi için) etiketlenmemiş ham derlem olarak belirlenmiştir. Ayrıca ham derlemi zenginleştirmek amacıyla PubMed veri kaynağından bitki ilaç etkileşim sorgusu yapılarak derlem zenginleştirilmiştir. Böylece 835 makale özetinden oluşan bir ham derlem elde edilmiştir.

6.2 BiE İçin Varlık Etiketli Altın Standart Derlem ve AVT Modeli Geliştirme Süreçleri

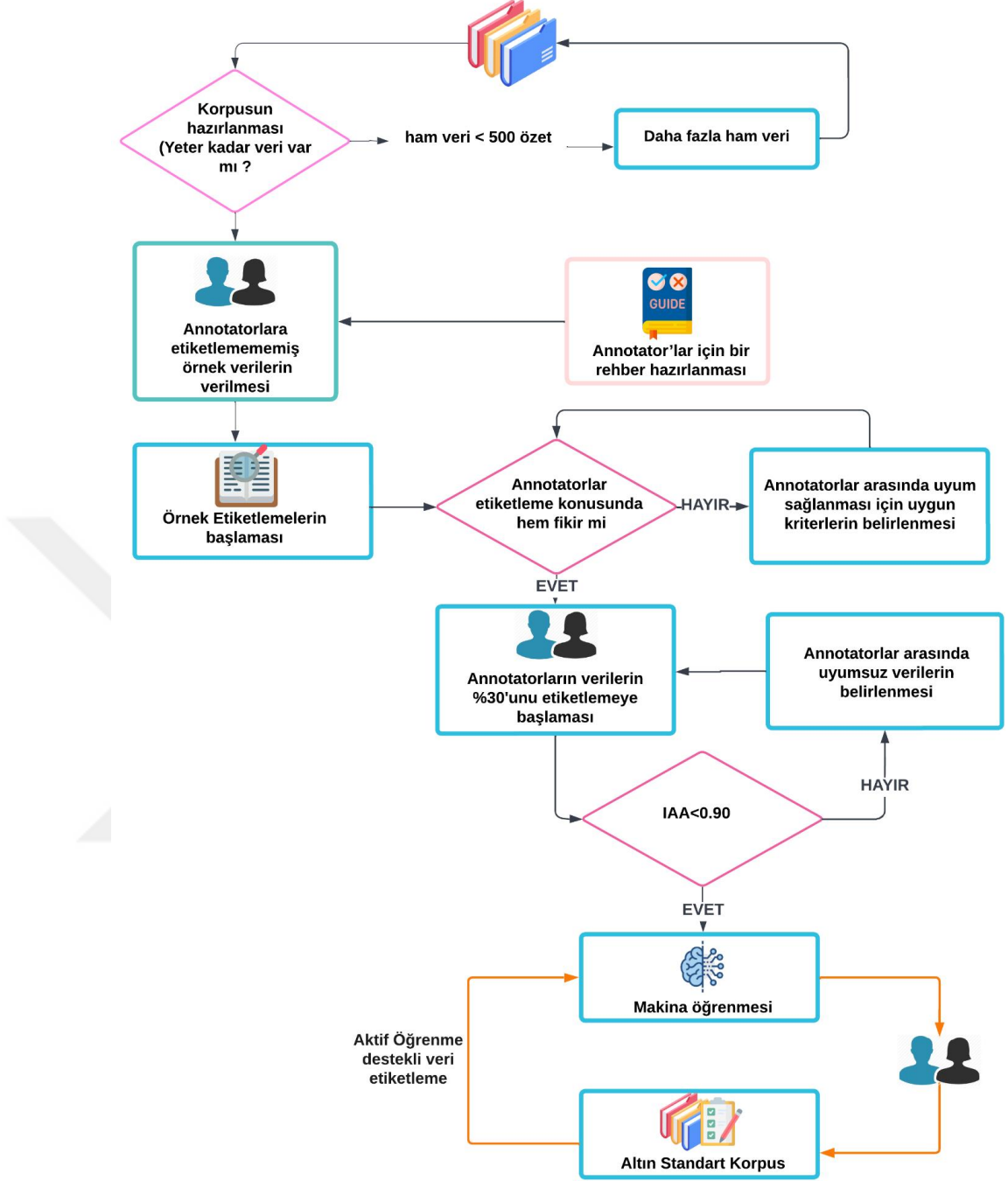
Bu bölüm iki aşamadan oluşmuştur. İlk bölümde altın standart derlem geliştirme süreci anlatılırken ikinci bölümde ise bu derlem kullanılarak makina öğrenmesi ve derin öğrenme tabanlı sistemler bir AVT modelinin oluşturulma süreci detaylı bir şekilde anlatılmıştır.

6.2.1 BİE varlıkları için altın standart (gold standart) derlem geliştirme süreci

Altın standart bir derlem oluşturma süreci pahalı, zaman alıcı ve uzman bilgisi gerektiren zahmetli bir süreçtir. Bu derlemleri geliştirebilmek için kitle kaynak platformları (Amazon MTURK) veya alan uzmanları ile birlikte çalışılmaktadır. Kitle kaynak platformları veri etiketleme sürecini küçük görevler halinde bir topluluğa veren sistemlerdir (bkz. bölüm 3.3). Ancak bu sistemler her ne kadar başarılı olsa da alan uzmanları kadar kaliteli olmadığı kanısı vardır. Bu nedenle bu çalışmada kaliteli bir altın standart derlem elde etmek için alan uzmanları seviyesinde bir etiketleme süreci gerçekleştirilmiştir. Ayrıca etiketleme sürecini maksimum derecede kolaylaştıracak ve meydana gelebilecek hataları minimum seviyeye indirecek bir etiketleme protokolü hazırlanmıştır. Böylece daha az vakit alan ve daha kaliteli bir etiketleme görevi gerçekleştirme amaçlanmıştır. Etiketleme için kullanılacak araç aktif öğrenme desteği verdiği için protokol bu araca göre düzenlenmiştir.

Etiketleme protokolünün detayları Şekil 6.3’de verilmiştir.

1. Etiketleme işlemi için etiketlenmemiş makale özeti sayısı 500’den az ise veri setini arttır.
2. Alan uzmanı küratörlere (annotator) etiketlemeyi nasıl yapacaklarına dair bir rehber hazırla.
3. Küratörlerin her birine birbirlerini göremeyeceği bir şekilde aynı sayıda (verinin %5’i kadar) etiketlemeleri için örnek veri ver.
4. Alan uzmanı küratörlerin etiketledikleri örnek veriler birbiri ile uyumlu ise veri setini etiketlemeye başla veriler birbiri ile uyumlu değil ise aralarındaki uyumsuzluk nedenini belirle.
5. Alan uzmanı küratörler aralarındaki uyumsuzluk nedeni belirlendi ise veri setinin %30’unu etiketlemeye başla.
6. Küratörlerin etiketledikleri veriler arası uyum skoru (inter annotator agreement (IAA)) yüzde 90 üzeri ise ortak uyumlu verilerden oluşan veri setini bir makina öğrenmesi modeli ile eğit veriler yüzde 90 altında ise uyumsuz verileri bir bilgi tabanı ile incele ve aradaki uyumsuzluğu minimize et ve uyum skorunu tekrar ölç.
7. Eğitilen modeli kullanarak kalan tüm veri setini eğit (aktif öğrenme destekli bir araç ile eğitilerek etiketleme işleminin hızlanması) ve yeteri etiketlemeden sonra ortak uyumlu verilerden oluşan altın standart derlem oluştur.



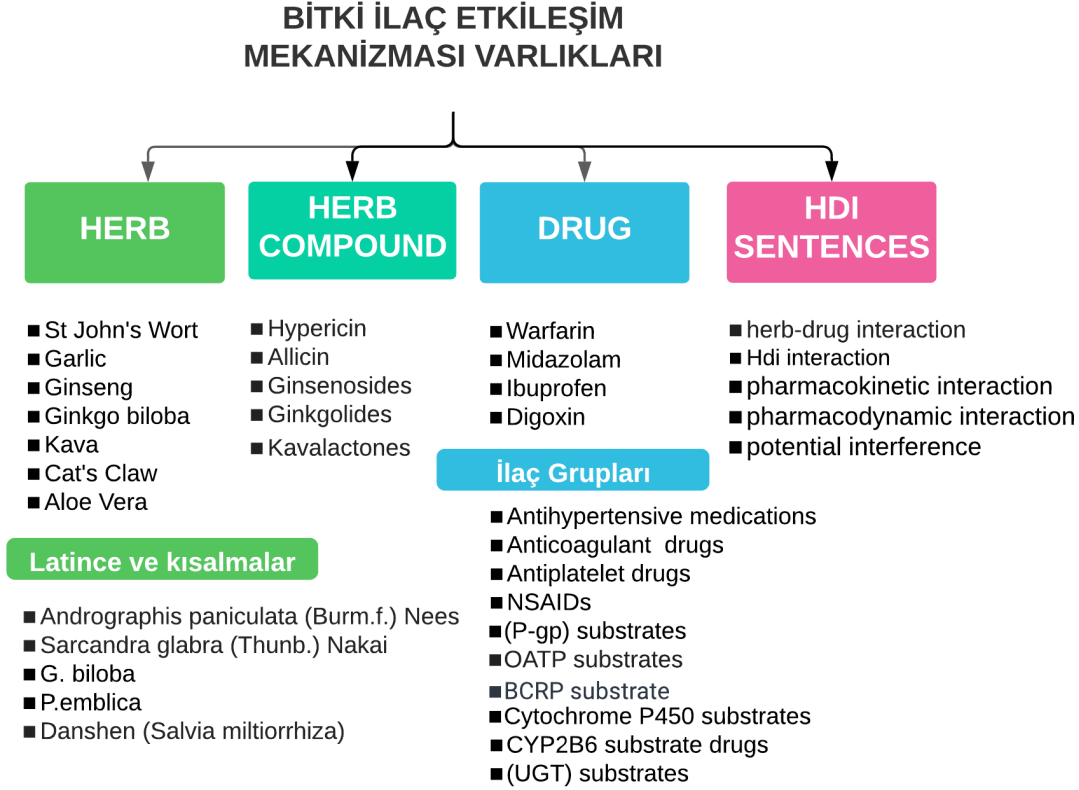
Şekil 6.3. Aktif öğrenme destekli altın standart derlem oluşturma protokolü

6.2.1.1 Uzman kişi belirleme

Uzman kişi olarak Ege Üniversitesi Biyokimya mühendisliği bölümünden bitki, ilaç ve kimya alanında alan uzmanı Prof. Dr. Figen Zihnioğlu ve aynı bölümünden Araş. Gör. Kerem Tok belirlenmiştir.

6.2.1.2 Rehber hazırlama ve kategori belirleme

Etiketleme süreci için bitki ve ilaç etkileşimleri mekanizmasına göre 4 varlık türü belirlenmiştir. Bu varlık türleri bitkiler için HERB, bitki bileşenleri için HERB_COMPOUND, ilaçlar için DRUG ve bitki ve ilaç etkileşimlerinde kilit rol oynayan bazı kelime ve kelime öbekleri için HDI_SENTENCES belirlenmiştir.



Şekil 6.4. Bitki-ilaç etkileşim mekanizmasına göre 4 kategoriden oluşan varlık etiketleme kategorileri ve bazı örnek veriler

Şekil 6.4'te bu etiketleme kategorileri ve örnekleri detaylı bir şekilde görülmektedir. Bu varlık türlerini etiketlemek için alan uzmanlarına bir takım kurallar bildirilmiştir. Bu kurallar bir rehber olarak etiketleme süreci başlamadan önce verilmiştir. Bu kurallar şöyledir.

HERB Varlık Türü: Makale özetlerinde geçen bitki varlıkları HERB olarak etiketlenecektir. Makale özetleri içinde geçen bitki adlandırmaları, ortak adlar, Latince adları, kısaltmalar gibi çeşitli şekillerde yazılabilir. Bu farklılıklar maddeler şeklinde belirtilirse;

1-Bilimsel makalelerde bitki adlandırmaları genel olarak tek kelimedenden oluşabilir birden fazla kelime öbeğinden de oluşabilir. Bitki kaç kelimedenden oluşuyor ise o kelimelerin tümü tek çatı altında etiketlenmesi istenmiştir.

Örneğin; “**garlic**” ve “**gingseng**” gibi tek kelimedenden oluşmakta “**St John Worth** (sarı kantaron)” ise üç kelimedenden oluşabilir.

2- Bazı bitkiler Latince olarak adlandırılmış olabilir veya bitki adı bitki aile ismi verilerek de yazılabilir. Bu tarz adlandırmalarında etiketleme sürecine dahil edilmesi belirtilmiştir. Örneğin aşağıdaki cümle incelenirse;

*“**Panax notoginseng (Burk.) F. H. Chen (Araliaceae)** is traditionally used for its hemostatic and cardiovascular effects when raw and as a tonic when steamed (Lau et al., 2009).”*

Yukarıdaki cümlede “**Panax notoginseng (Burk.) F. H. Chen (Araliaceae)**” bitkisi aile ismi verilerek adlandırılmıştır. Bu adlandırmanın hepsi bir bütün olarak etiketlenecektir.

3-Bazı bitki isimleri hem geleneksel hem de Latince isimlerle birlikte kullanılmaktadır. Bu iki adlandırmanın da ayrı ayrı etiketleme sürecine dahil olması belirtilmiştir. Örneğin;

*“Plausible cases of herb-drug interactions include: bleeding when warfarin is combined with **ginkgo (Ginkgo biloba)**, **garlic (Allium sativum)**, **dong quai (Angelica sinensis)**, or **danshen (Salvia miltiorrhiza)**”*

4-Bazı paragraflarda, paragrafın ilk cümlesinde bitki adları tam olarak yazılırken, paragrafın devamında bitki adının anlamlı bir kısaltması kullanılmıştır. Örneğin aşağıdaki cümle incelenirse;

*“**Ginkgo biloba** is available as an over-the-counter drug and reported to cause haemorrhage when coadministered with other antiplatelet agents. We set out to study the interactions of **G. biloba** with cilostazol and clopidogrel”(Wang et al., 2016).*

- Phyllanthus emblica yerine **P.emblica**
- Terminalia bellerica yerine **T.bellerica**

Yukarıdaki cümle ve örnekler incelendiğinde Ginkgo biloba, Phyllanthus emblica ve Terminalia bellerica ilk olarak tam adıyla yazılırken cümlelerin devamında kısaltma kullanılmıştır. Bu tarz adlandırmaların etiketleme sürecine dahil olması gerektiği belirtilmiştir.

Ayrıca yukarıda yazılan adlandırma formatlarının dışında da çeşitli formatlar bulunabilir bu tarz adlandırmalarında etiketlemeye dahil olacağı alan uzmanlarına belirtilmiştir. Ancak bazı makale özetlerinde bitki adlarında karmaşıklık oluşturabilecek kısaltmalar yapılmıştır. Bu şekilde yazılan kısaltmalar çoğunlukla bitki adlarının ilk karakterlerinin kısaltmalarıdır. Bunlar etiketlemeye dahil olmaması belirtilmiştir. Örneğin aşağıdaki kelimeler incelendiğinde;

- Tribulus Terrestris yerine **TT**
- Panax Ginseng yerine **PG**

Tribulus Terrestris yerine “**TT**” ve Panax Ginseng yerine “**PG**” gibi anlamsız kısaltmalar kullanılmıştır. Bunlar etiketleme sürecine dahil olmamıştır.

HERB_COMPOUND varlık türü: Bitki ilaç etkileşim olarak geçen etkileşim türlerinin bazıları bitki bileşikler düzeyinde oldukları gözlemlenmiştir. Bu varlıklar sayıca az olmasına rağmen ayrı olarak kategorize edilmiştir. Çizelge 6.2’de yaygın geçen bitki bileşenlerinin bir listesi verilmiştir.

Çizelge 6.2. Bilimsel makalelerde sıklıkla geçen bitki ve bitki bileşenlerine ait örnekler

HERB (bitki)	HERB COMPOUND (Bitki bileşenleri)	Referans
Ginkgo	Ginkgolides	(Wang <i>et al.</i> , 2016)
St. John's wort	Hypericin	(Barnes, Anderson and Phillipson, 2001)
Garlic	Allicin	(Patel <i>et al.</i> , 2004)
Ginseng	Ginsenosides	(Hao <i>et al.</i> , 2011)
Kava	Kavalactones	(Yamazaki <i>et al.</i> , 2008)

DRUG varlık türü: Makalede geçen ilaçların ve ilaç gruplarının etiketlenmesi için belirlenen kategoridir. Bu kategorideki varlıklar aşağıdaki şekilde etiketlenmiştir;

1- İlaçlarda aynı bitkiler gibi tek kelimedenden veya birçok kelimedenden oluşabilirler. Alan uzmanlarına bu yapıları tek bir bütün halinde etiketlenmesi gerektiği bildirilmiştir. Örneğin aşağıdaki cümlede koyu olarak etiketlenmiş tek kelimedenden oluşan aspirin ve warfarin ilacı görülmektedir.

*“Potential risks of bleeding due to the use of Ginkgo biloba, garlic or ginseng with **aspirin** or **warfarin** was the most reported herb-drug interaction (Agbabiaka et al. 2017). “*

2-İlaç adlandırmaları sınıf isimleri olabilir ve bir bitki ilaç etkileşimi sınıf seviyesinde verilmiş olabilir. Bu sınıflarda etiketleme işlemine dahil edilmesi gerektiği belirtilmiştir. Aşağıdaki cümlede koyu olarak gösterilenler ilaç sınıflarını göstermektedir.

*“Prescription medicines most combined with HMPs were **antihypertensive drugs, β -blockers, diuretics, antihyperlipidemic agents, anticoagulants, analgesics, antihistamines, antidiabetics, antidepressants and statins** (Agbabiaka et al. 2017). ”*

3- Bazı makalelerde geçen ilaç sınıfları araya konan bir “and” bağlacı yada “/” işareti ile birleştirilmiştir. Bu adlandırmalar bir bütün olarak etiketlenmesi gerektiği bildirilmiştir. Örneğin aşağıda koyu olarak belirtilen ilaç sınıfları bir bütün olarak etiketlenmesi belirtilmiştir.

- *“**Ginkgo biloba** extract, advertised as improving cognitive functioning, has been reported to cause spontaneous bleeding, and it may interact with **anticoagulants and antiplatelet agents** (Cupp, 1999). ”*
- *“Nowadays, patients with thrombus presiding to **anticoagulant/antiplatelet drugs** prefer taking TCM. However, an increasing number of studies on herb-drug interactions have been shown (Li et al. 2019). ”*

4- Bazı makalelerde geçen adlandırmalar ise enzim türlerini inhibe ve indükleyen ilaç sınıfları olabilir.

*“In conclusion, SLE was a strong P-gp inhibitor, which indicated a potential herb-drug interaction when SLE was co-administered with **P-gp substrate drugs** (Liang et al. 2013). ”*

HDI_SENTENCES varlık türü: Bu kategori bitki ve ilaç etkileşimi için kritik öneme sahip kelimelerin etiketlendiği bir kategoridir. Örneğin;

- herb-drug interaction / hdi interaction
- pharmacokinetic herb drug interaction / pharmacokinetic hdi
- pharmacodynamic herb drug interaction/ pharmacodynamic hdi
- potential to interact

6.2.1.3 Etiketleme aracını belirleme

Ham veride bulunan bitki ve ilaç varlıklarını etiketlemek için doğal dil işleme üzerine oldukça ünlü bir kütüphane olan Spacy kurucularının ürettiği özelleştirilebilir bir etiketleme aracı olan “**Prodi.gy Annotation Tool**” (Prodigy , 2022) kullanılmıştır.

Prodigy Etiketleme Aracı: Prodigy, makine öğrenimi modelleri için eğitim ve değerlendirme verileri oluşturmayı sağlayan bir açıklama aracıdır. Web arayüzünde çalışan kolay kullanımlı bir arayüzü vardır ve ayrıca aktif öğrenme desteği sağlamaktadır. Prodigy etiketleme aracı varlık etiketleme, metin kategori etiketleme, ilişki etiketleme, görüntü etiketleme, ses etiketleme gibi birçok görev sağlamaktadır.

Prdigy etiketleme aracını çalıştırma için komut satırına aşağıdaki şekil 6.5’te yer alan kod yazılır. Bu komutlara göre;

```
python -m prodigy ner.manual [dataset-name] blank:en
[sentence-file-path] -l [labels-file-path]
```

Şekil 6.5. Prodigy etiketleme aracı çalıştırma komutları

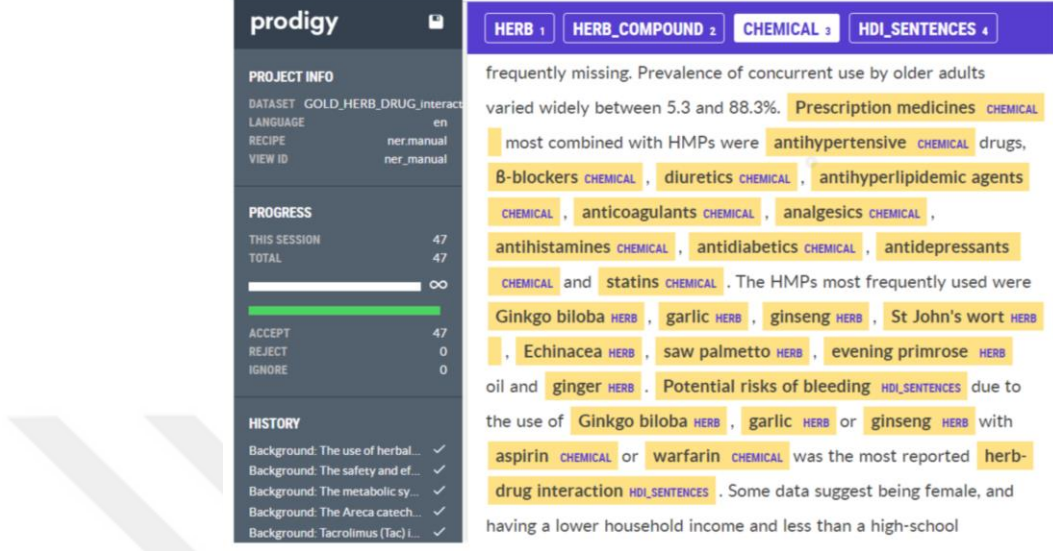
python -m prodigy: Prodigy uygulamasını çalıştırmak için yazılan komut

ner.manual: Prodigy aracına yapılacak işin varlık adlandırma görevi olduğunu belirten komut. Bu satırda rel.manual komutuda mevcuttur. rel. manual komutu ise ilişki etiketleme görevi sağlar.

[dataset-name]: Etiketlenecek olan verisetinin kaydedileceği veri tabanı adı

[blank: en]: Bu komut eğer var ise önceden eğitilmiş bir model üzerine etiketlemenin sağlanması sağlar

[sentences-file-path]: etiketlenecek olan verisetinin konumu ve adı
 -l ve [labels- fike- path]: verisetinde belirlenecek olan etiket kategorilerinin belirlendiği yer



Şekil 6.6. Prodigy etiketleme aracının bir görseli (Chemical olarak etiketlenenler ilaçlardır. Daha sonra bu kategori DRUG olarak değiştirilmiştir)

6.2.1.4 Etiketleme işleminin başlaması

Uzman kişilerin belirlenmesi, rehberin hazırlanması ve etiketleme aracının belirlenmesinden sonra etiketleme süreci başlamıştır. Prodigy etiketleme aracı yerel bilgisayarda çalışan bir uygulamadır. Bu çalışma hazırlanırken Covid19'un ölümcül tehlikesi bulunmakta ve çalışmalar çevrim içi yapılmaktaydı. Bu nedenle etiketleme süreci uzaktan bağlanma desteği sağlayan "ANYDESK" (AnyDesk: The Fast Remote Desktop Application, 2022) programı yüklenerek sağlanmıştır. Alan uzmanları bu program aracılığıyla verileri uzaktan ayrı ayrı etiketlemişlerdir.

```
>> python -m prodigy ner.manual GOLD_HERB_DRUG_interaction_v1
en_core_web_sm ./mskcc.json
--label HERB, HERB_COMPOUND, DRUG, HDI_SENTENCES
```

Şekil 6.7. BİE etkileşimi ile ilgili varlıkların Prodigy etiketleme aracı ile başlatılması

Etiketleme sürecinin başlaması için Şekil 6.7'de gösterilen komut satırı yazılmıştır. Bu komut satırına göre;

- **[ner.manual]** : programın varlık etiketleme görevi yapılacağını belirten komut satırını
- **[GOLD_HERB_DRUG_interaction_v1]**: etiketli verilerin kaydedileceği SQL veri tabanı
- **[en_core_web_sm]** : Spacy programına ait tokenleştirme işlemini sağlayan önceden tanımlı dil modelini
- **[./mskcc.json]** : BİE ile ilgili 40 adet örnek makale özetinden oluşan ham derlem
- **[--label HERB, HERB_COMPOUND, DRUG, HDI_SENTENCES]** : Dört adet kategorinin belirlendiği komut satırını göstermektedir

Echinacea **HERB**, ephedra **HERB**,
garlic **HERB**, ginkgo **HERB**,
ginseng **HERB**, kava **HERB**,
St John's wort **HERB**, and
valerian **HERB** are commonly used
herbal medications

Şekil 6.8. Yaygın kullanılan bir ve birden fazla kelimedenden oluşan bitkilerin etiketlenmesi (Ang-Lee, Moss and Yuan, 2001)

Etiketleme aracı çalıştıktan sonra protokole uygun olarak etiketleme işlemi şu adımlar sağlanarak gerçekleşmiştir:

Birinci Adım: Alan uzmanı küratörlere detaylı hazırlanan etiketleme rehberi eşliğinde rastgele seçilen 40 (derlemin yüzde 5'i) adet örnek makaleyi HERB, HERB_COMPOUND, DRUG ve HDI_SENTENCES kategorilerine göre etiketlemeleri istenmiştir. Alan uzmanları bu adımda birbirlerini kesinlikle görmemiştir ve birbirlerinden etkilenmemeleri için farklı zaman dilimleri seçilmiştir. Şekil 6.8 bir makale özetindeki bir kesitin küratörler tarafından etiketlenmesini göstermektedir.

İkinci Adım: Küratörlerin örnek etiketlemeleri sağladıktan sonra, aralarında hemfikir olmadıkları varlık örnekleri hakkında istişarede bulundular. Bu iki küratörün etiketleme sürecinde tutarlı olmasını sağlamak amacıyla gerçekleştirilmiştir.

Üçüncü Adım: Küratörlerin daha sonra 80 (derlemin % 10'u) adet makaleyi etiketlemeleri sağlanmıştır. Bu işlem her küratör için yaklaşık 5 saat sürmüştür.

Dördüncü Adım: 80 adet makale özeti etiketlendikten sonra Kappa skoru kullanılarak küratörler arasındaki tutarlılık ölçülmüştür. Bu adımın detayları aşağıdaki gibidir.

Uyum Skorunun Hesaplanması: Bu skoru hesaplamak için özet cümledeki her bir kelimeye 0,1 ve 2 den oluşan vektörler atanmıştır.

- Derlemde geçen cümlelerde etiketlenmeyen kelimelerin temsili için : “0”
- Derlemde geçen cümlelerde HERB olarak etiketlenen kelimeler için : “1”
- Derlemde geçen cümlelerde DRUG olarak etiketlenen kelimeler için: “2”

etiketi belirlenmiştir. Örneğin aşağıdaki cümle incelenirse;

“Bitter melon[HERB] and garlic[HERB] interact with vinblastine [DRUG]”

bu durumda yukarıdaki cümle;

[“1111111111110001111110000000000022222222222”] (6.1)

olarak vektörize edilecektir. Bu cümlede “Bitter melon” ve “garlic ” kelimeleri “HERB” olduğu için iki kelimeye 1 vektörü atanmıştır. Son kelimedeki “vinblastine” bir “DRUG” olduğu için 2 vektörü atanmıştır. İki küratörün vektörleri birbiri ile eşleştirilerek böylece aralarındaki tutarlılık ölçülmüştür. Daha sonra **python scikit learn** kütüphanesi kullanılarak kappa skoru ölçülmüştür. İki küratör arasındaki oran $0 < \text{değer} < 1$ aralığında bir değer döndürmektedir. Bu değer tablosu incelendiğinde 0.81 üzeri doğruluk yüksek başarı olarak değerlendirilebilir. Bu çalışmada küratörler arasında uyum skoru 0.92 çıkmıştır.

Beşinci Adım: Bu adımda etiketlenmiş 80 adet makale özeti %20 test %80 eğitim verisi olacak şekilde Spacy kütüphanesine ait önceden eğitilmiş bir derin öğrenme modeli olan ve CNN (evrimsel sinir ağı, convolutional neural network) alt yapısını kullanan Toc2vec (*Tok2Vec spaCy*, 2022) modeli ile eğitilerek bir taslak model üretilmiştir. Bu taslak model, metindeki varlıkları belirlenen kategorilere ve modelin başarısına göre önceden işaretler. Küratörler yalnızca bu etiketlenen varlıkların doğruluğunu onaylar. Eğer hatalı veya eksik etiketleme var ise bunu düzelterek model tarafından işaretlenmiş bir sonraki makaleyi etiketlemeye yönelmişlerdir. Prodigy tarafından etiketlenen veriler veri tabanında aşağıdaki gösterilen JSONL formatında tutulur.

“{

```

"text": "bleeding when warfarin is combined with ginkgo.",
"tokens": [
  { "text": "bleeding", "start": 0, "end": 8, "id": 0 },
  { "text": "when ", "start": 9, "end": 13, "id": 1 },
  { "text": "warfarin ", "start": 14, "end": 21, "id": 2 },
  { "text": "is ", "start": 22, "end": 24, "id": 3 },
  { "text": "combined ", "start": 25, "end": 33, "id": 4 },
  { "text": "with ", "start": 34, "end": 38, "id": 5 },
  { "text": "ginkgo", "start": 39, "end": 45, "id": 6 },
  { "text": ".", "start": 46, "end": 47, "id": 7 }
],
"spans": [
  {
    "text": "warfarin",
    "start": 14,
    "end": 21,
    "token_start": 2,
    "token_end": 2,
    "label": "DRUG"
  },
  {
    "text": "ginkgo",
    "start": 39,
    "end": 45,
    "token_start": 5,
    "token_end": 5,
    "label": "HERB"
  }
]
}

```

Text: Her bir girdi için ham metni belirtir

Tokens: Cümlelerin en küçük birimi olan kelimeleri ve her kelimenin indeksinin tutulduğu liste

Spans: Küratörlerce etiketlenen varlıkların başlangıç(start) ve son (end) karakter indekslerini tutan liste

token_start: Varlıkların cümle içinde kaçınıcı kelime ile başladığını ve kaç kelimedenden oluştuğunu belirtir

token_end: Varlıkların cümle içinde kaçınıcı kelime ile sonlandığını belirtir varlık tek kelimedenden oluşuyorsa token_start ile aynı değeri alır.

label: Varlıkların etiketlendiği kategorileri tutar

Altıncı Adım: Bu adımda model kullanılarak yeterli derecede etiketleme yapılmıştır. Verisetinin kaliteli ve tutarlı olması açısından iki alan uzmanı farklı zaman dilimlerinde ve farklı aralıklarla birbirlerini görmeyecek bir şekilde

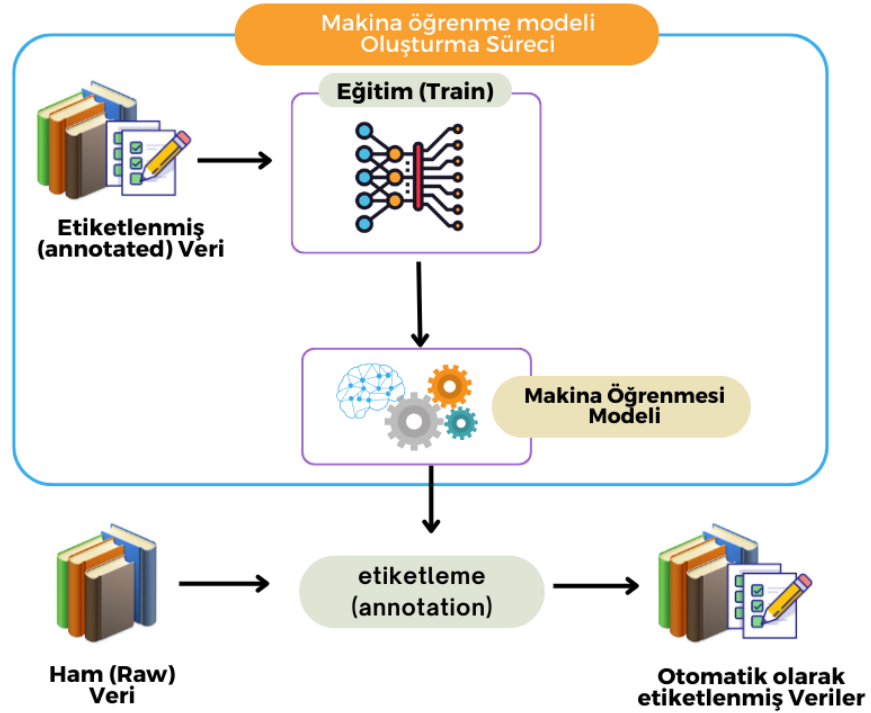
etiketleme işlemini tamamlamışlardır. Etiketleme işlemi her küratör için yaklaşık 7 saat sürmüştür.

6.2.1.5 Varlık tanıma için altın standart derlem

Tüm mmakale özetleri etiketlendikten sonra küratörlerin etiketledikleri JSONL formatındaki veri setleri arasındaki tutarlılığı ölçmek için uyum skoru (kappa skoru) tekrar ölçülmüştür. Uyum skoru %90 üzeri tatmin edici bir skor çıkmıştır. Küratörlerce ortak olarak etkilenen varlıkların kesişimi alınarak bu veri seti altın standart derlem olarak literatüre kazandırılmıştır. Çalışmanın detayları sonuç bölümünde verilmiştir.

6.2.2 AVT (NER) modeli oluşturma süreci

Tez çalışmasının bu bölümünde biyomedikal literatürde geçen ilaç ve bitki varlıklarını otomatik olarak yakalayabilecek en iyi performansa sahip derlem tabanlı bir AVT (NER) modeli geliştirilmiştir. Şekil 6.9’da görüldüğü üzere derlem tabanlı bir AVT modeli için en önemli bileşen, en iyi performansı verebilecek bir modelin seçimi ve bu modelleri eğitecek eğitim verisidir. Eğitilen model daha sonra ham veriler üzerinde bir tahmin yaparak otomatik bir şekilde verileri etiketleyebilecektir.



Şekil 6.9. RobertA Dil modeli ince ayarlanarak (fine tune) biyomedikal literatürde yer alan bitki ve ilaç terimlerinin otomatik olarak etiketlenmesi

Model Seçimi: Python'un spaCy (2022) kütüphanesi AVT ve ilişki çıkarımı gibi çeşitli doğal dil işleme görevleri için oldukça popüler bir kütüphanedir. Metinde bitki ve ilaç terimlerini tanımlayabilecek en başarılı model seçimi için için spacy kütüphanesine ait Tok2vec (tok2vec, 2022) ve RoBERTa dönüştürücü (Transformer) modelleri kullanılmıştır ve sonuçları birbiri ile karşılaştırılmıştır.

Eğitim verisi: Önceden eğitilmiş dil modelleri belirli görevler (AVT, ilişki çıkarımı vb) için eğitmek istenirse bunun için etiketlenmiş altın standart veri setine ihtiyaç duymaktadırlar. Bu bakımdan bir önceki bölümde BİE için oluşturulan varlık etiketli altın standart derlem bu modeller için eğitim verisi olarak belirlenmiştir.

6.2.2.1 Tok2Vec tabanlı AVT modelinin kurulumu

Spacy Tok2vec evrişimli sını ağı (CNN) tabanlı bir ön eğitilmiş dil modelidir. Tok2vec modelini eğitmek için Prodigy alt yapısı kullanılmıştır. Bunun için aşağıdaki komut dizini çalıştırılmıştır.

```
>> prodigy train ner GOLD_HERB_DRUG_interaction_v1 en_vectors_web_lg --
init-tok2vec ./tok2vec_cd8_model289.bin --output ./temp_model --eval-split 0.2
```

- prodigy train ner= Bir ner görevinin eğitilme bileşeninin ifade eder.
- GOLD_HERB_DRUG_interaction_v1= Bitki ve ilaç varlık etiketli altın standart verisetinin saklandığı veri tabanı, eğitim veriseti olarak kullanılmaktadır.
- en_vectors_web_lg= Spacy kütüphanesine ait temel bir dil modeli. Tok2vec katmanını eğitmek için kullanılmaktadır.
- --init-tok2vec ./tok2vec_cd8_model289.bin= Önceden eğitilmiş dil modeli, AVT modelini başlatmak (initialize) için kullanılır.
- --output ./temp_model = Ortaya çıkan model dosyalarının konumunu ifade eder (yaklaşık 1 GB)
- --eval-split 0.2= Eğitim ve test bölme oranını ifade eder. Bu çalışmada veriseti %20 si test %80'i eğitim olacak şekilde bölünmüştür.

Toplam 680 makaleden özetinden 544'si eğitim 126 tanesi test için bölünmüştür. Çıkan sonuçlar bir sonraki bölümde (bkz. 7.3) verilmiştir. Bu modeli eğitmek için 12 çekirdekli (CPU) 3,6 Ghz hızına ve 24 Gb Ddi Ram özelliklerine sahip bir bilgisayar kullanılmıştır.

6.2.2.2 **Transformer (RoBERTa) tabanlı AVT modelinin kurulumu**

Bir diğer model olan RoBERTa (Liu *et al.*, 2019) transformer modeli ise BERT mimarisine sahip bir modeldir. BERT'ten daha fazla sayıda veriseti ile önceden eğitilmiş bir dil modelidir. Bu modeli eğitmek için Spacy kütüphanesinin son versiyonu olan Spacy v3 ve ayrıca ortam olarak da Google Colab (2022) kullanılmıştır. Modelin eğitimi için gerekli olan aşamalar şu şekilde gerçekleştirilmiştir.

1-Eğitim verilerinin SpaCy'nin ikili biçimine (.spacy) dönüştürülmesi: Bu yeni biçim, serileştirilmiş bir DocBin nesnesinden oluşur. Özellikle birden fazla belgeyi bir arada paketlerken, depolamada verim sağlamak için kullanılmaktadır (*Doc · spaCy API documentation*, 2022). Bunun için şu adımlar gerçekleştirilmiştir.

- Altın standart derlem Prodigy veri tabanından JSONL formatı ile dışarı çıkarılmıştır.
- JSONL formatında bulunan etiketli derlem Python programlama dili kullanılarak Spacy'nin belirlediği bir format ile RoBERTa dil modeli ile eğitilebilecek bir yapıya getirilmiştir.
- Son olarak %90 eğitim ve %10 validasyon olmak üzere spacy formatında iki dosya oluşturulur.

2-Bir config.cfg dosyası ile tüm ayarları ve hiper parametreleri tanımlama ve model oluşturma: Spacy transformers dil modelinin eğitilebilmesi için yapılandırma dosyalarına ihtiyaç vardır. Bu dosyalar için Spacy kütüphanesinden yapılandırma dosyaları alınmıştır (*Training pipelines & models · spaCy usage documentation*, 2022).

```
"!python -m spacy train config.cfg --verbose --output
ner_demo/training/ --paths.train train.spacy --paths.dev
valid.spacy --gpu-id 0"
```

Oluşturulan yapılandırma dosyası ile yukarıdaki kod satırı yazılarak bir önceki aşamada oluşturulan eğitim (train.spacy) ve validasyon (valid.spacy) veri setleri üzerinde model oluşturulur.

6.2.2.3 Değerlendirme

Tok2Vec ve RoBERTa modellerinin ham verideki varlıkları doğru bir şekilde etiketleyebilme başarılarının değerlendirilmesi gereklidir. Bu değerlendirme şu şekilde gerçekleştirilir;

Gerçek Pozitif (True Positive): Model, etiketlenmiş derlemde var olan bir etiket (açıklama) sağlar;

Gerçek Negatif (True Negative): Bir etiketin olmaması, etiketlenmiş derleme göre doğrudur;

Yanlış Pozitif (False Positive): Model, etiketlenmiş derlemde bulunmayan bir etiket sağlar;

Yanlış Negatif (False Negative): Model, etiketlenmiş derlemde bulunan bir etiket sağlamaz.

AVT ve ilişki çıkarımını değerlendirmek için sıklıkla kullanılan bir hata ölçütü, kesinlik ve duyarlılık (recall) ölçütlerinin bir kombinasyonu olan F-skorudur. Bu ölçütler 0 (en kötü) ve 1 (en iyi) arasında değerler alır.

$$\text{Kesinlik} = \frac{\text{Doğru tespit edilen varlıklar}}{\text{Tespit edilen toplam varlıklar}} \quad (6.2)$$

$$\text{Kesinlik_formül} = \frac{\text{Gerçek Pozitif}}{\text{Gerçek Pozitif} + \text{Yanlış Pozitif}} \quad (6.3)$$

$$\text{Duyarlılık} = \frac{\text{Doğru tespit edilen varlıklar}}{\text{Korpusdaki tüm varlıklar}} \quad (6.4)$$

$$\text{Duyarlılık_formül} = \frac{\text{Gerçek Pozitif}}{\text{Gerçek Pozitif} + \text{Negatif}} \quad (6.5)$$

$$F - \text{Skor} = 2 \times \frac{\text{Kesinlik} \times \text{Duyarlılık}}{\text{Kesinlik} + \text{Duyarlılık}} \quad (6.6)$$

Kesinlik (precision): Bulunan tüm varlık adlarında geçen gerçek varlık adlarının yüzdesini ifade eder ve ayrıca özgüllük olarak da adlanır.

Duyarlılık (recall): Tüm gerçek varlık adlarına kıyasla bulunan gerçek varlık adlarının yüzdesini ifade eder.

6.3 İlişki Etiketli Altın Standart Derlem ve İlişki Çıkarım Modeli Geliştirme Süreçleri

Bu tez çalışmasının bu bölümünün amacı bitki ve ilaç varlıkları arasındaki geçen ilişkileri için bir altın standart derlem geliştirmek ve daha sonra yapılandırılmamış biyomedikal literatürden bu ilişkileri otomatik olarak yakalayabilen modeller geliştirmektir.

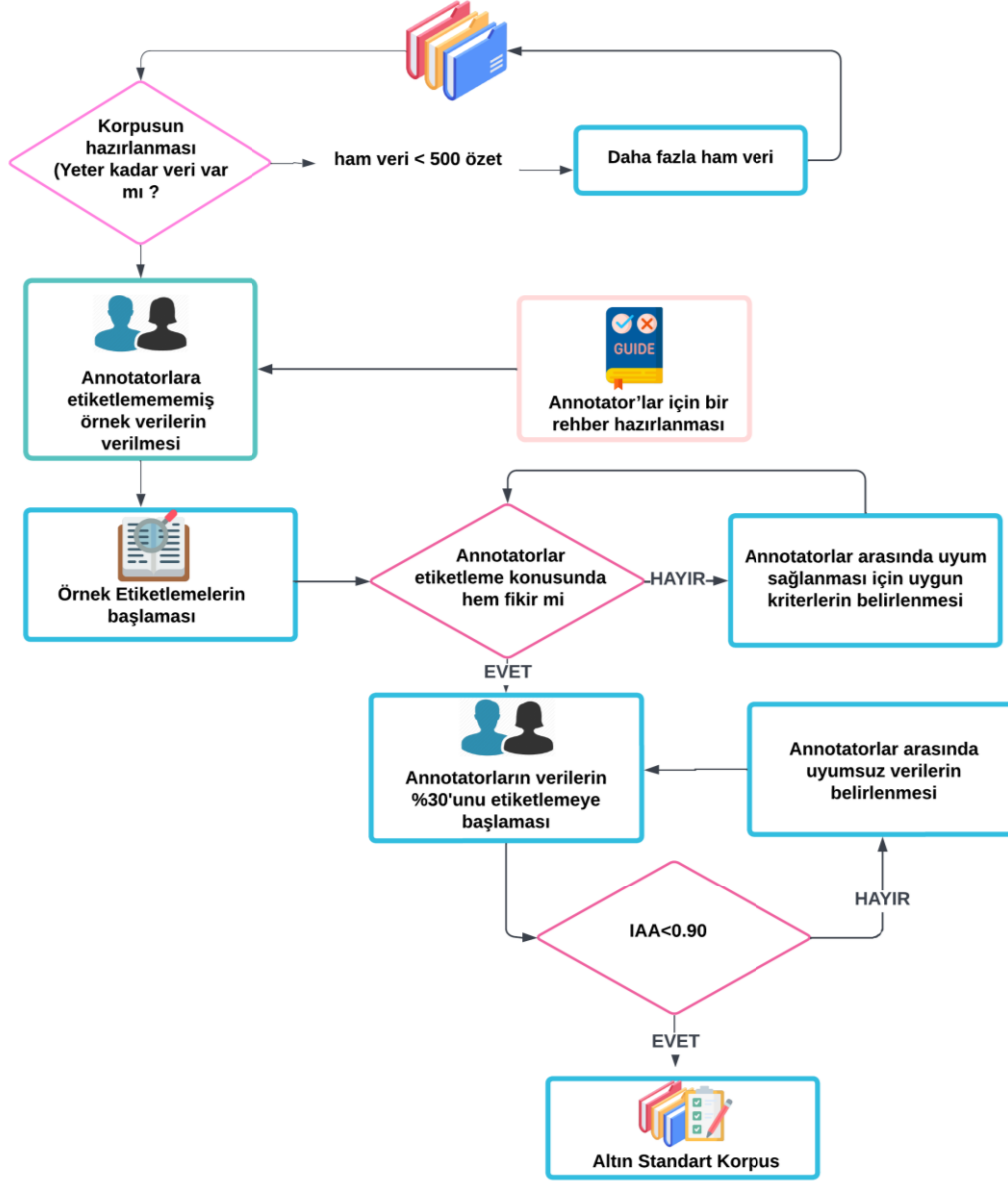
Bu tez çalışmasında bitki ve ilaç varlıkları arasında sadece sınıflandırma için kullanılacak olan altın standart ilişki tipi (ADMET, Potantiative, Summative, Adverse, Antagonistic/Agonistic, Additive, No interaction) derlemi ve ilişki tespit ve sınıflandırma için kullanılacak (POSITIVE, NEGATIVE, UNCERTAIN) altın standart ilişki tespit derlemi geliştirilmiştir. Daha sonra aşağıdaki modeller geliştirilerek ilişki tespiti ve ilişki türünün otomatik olarak biyomedikal literatürden çıkarılabilmesi sağlanmıştır.

- Makina öğrenmesi tabanlı (SVM, Random Forest, KNN)
- Dil modeli tabanlı (BERT, BioBERT)
- makina öğrenmesi ve kural tabanlı modellerin bir birleşimi olan zayıf denetimli öğrenme tabanlı (SNORKEL)

Yukarıdaki modellerden en iyi performansa sahip olanı temel ilişki modeli olarak belirlenmiştir.

6.3.1 İlişki tespiti etiketli altın standart derlem etiketleme süreci

İlişki etiketli bir derlem oluşturma süreci varlık etiketli derlem oluşturma sürecine göre daha zahmetli, daha çok vakit alan ve daha çok uzmanlık bilgisi gerektiren bir süreçtir. Bu süreç içinde ilk olarak ilk olarak Şekil 6.10'da görüldüğü üzere bir protokol hazırlanmıştır. Bu protokole bir önceki varlık etiketleme protokolünün bir benzeridir ancak bu protokoldeki fark ilişki etiketlemede aktif öğrenme destekli etiketleme yönteminin kullanılmamasıdır.



Şekil 6.10. İlişki etiketleme protokolü

Etiketleme süreci adımları şu şekilde gerçekleştirilmiştir:

1.Adım: İlk olarak etiketleme işlemi için bir önceki bölümde AVT derlemini etiketleyen alan uzmanları küratör olarak belirlenmiştir.

2.Adım: İkinci olarak etiketleme aracı belirlenmiştir. Bu işlem için ilişki tipinde etiketleme özelliği bulunan bir önceki AVT görevi için kullanılan Prodigy etiketleme aracı kullanılmıştır. İlişki etiketleme görevi Prodigy 'de iki şekilde olmaktadır:

- Uygun komutlar ile etiketleme aracı başlatılır ve belirlenen veri setinde ilk olarak varlık aday çiftleri etiketlenir ve daha sonra aralarındaki ilişki kategorilerden hangisine giriyorsa o kategori ile etiketlenir.
- İkinci yol ise elde varlık etiketli bir derlem var ise bu durumda yalnızca ilişkiler belirlenir ve küratörler yalnızca ilişki etiketleme görevini gerçekleştirir.

Bu süreçte elimizde daha önce var olan varlık etiketli derlemi kullanarak küratörlerin yalnızca ilişkileri etiketlemesi sağlanmıştır.

3.Adım: Bu adımda ilk olarak aday bitki ve ilaç varlıkları arasında ilişki kategorileri belirlenmiştir. Bunun için ilk olarak 3 kategori belirlenmiştir Çizelge 6.3’de bu kategoriler ile ilgili örnekler verilmiştir. Bunlar;

- **POSITIVE:** Aday varlıkların bulunduğu cümlede farmakokinetik, farmakodinamik veya tavsiye niteliğinde bir etkileşim söz konusu ise bu kategori ile etiketlenmesi gerektiği belirtilmiştir.
- **NEGATIVE:** Aday varlıkların bulunduğu cümlede herhangi bir etkileşimin tespit edilmediği bir durum var ise ilişkilerin bu kategori etiketlenmesi gerektiği belirtilmiştir.
- **UNCERTAIN:** İki aday varlık arasında bir etkileşim olup olmadığı araştırılıyorsa veya herhangi bir ilaç ve bitki varlığının birlikte geçtiği araştırmaya bir durum söz konusu ise bu kategori ile etiketlenmesi belirtilmiştir.

Çizelge 6.3. Bir cümlede geçen aday bitki ve ilaçlar arasında meydana gelen etkileşimler

Dokümanda Geçen Cümle	Aday İlaç	Aday Bitki	Etkileşim Sınıfı
“boldo-fenugreek and increased bleeding time in patients treated with warfarin.”	warfarin	boldo, fenugreek	POSITIVE
“ginkgo and ginger did not affect the apparent volumes of distribution or protein binding of either s-warfarin or r-warfarin.”	ginkgo, ginger	s-warfarin, r-warfarin	NEGATIVE
“this study aimed to investigate the pharmacokinetics of fa and potential action with honghua and clopidogrel...”	honghua	clopidogrel	UNCERTAIN

4.Adım: Etiketleme işleminin başlatılması için etiketleme aracı şu komutlarla başlatılmıştır.

```
>> python -m prodigy rel.manual REL_EXTRACTION_HDI_v1
Mskcc_PubMed.jsonl -label POSITIVE, NEGATIVE, UNCERTAIN, COREF --
span-label HERB, HERB_COMPOUND, DRUG
```

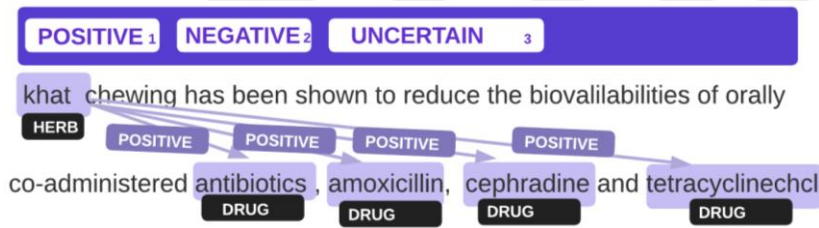
python -m rel.manual: Bir ilişki etiketleme görevi yapılacağını belirtmek için yazılır.

REL_EXTRACTION_HDI_v1: İlişki etiketli veri setinin kaydedileceği veri tabanı ismini belirtmektedir.

Mskcc_PubMed.jsonl: İlişkilerin etiketleneceği makale özetlerinin yer aldığı dosya ismidir.

-label: İlişki kategorilerinin belirlendiği dizindir.

-span-label: Varlık kategorilerinin belirlendiği dizindir.



Şekil 6.11. Prodigy etiketleme aracı ile ilişki var-yok seviyesinde etiketleme işlemi

5.Adım: Etiketleme süreci bir önceki bölümde olduğu gibi iki alan uzmanı birbirini görmeyecek bir şekilde ve protokole uygun bir biçimde ilk olarak yaklaşık 10 adet örnek veri ile başlamıştır. Küratörler yerel bilgisayara ANYDESK uygulamasından bağlanarak etiketleme işlemini uzaktan gerçekleştirmişlerdir (bkz Şekil 6.11).

6-Adım Örnek veriler üzerinden küratörler arası yapılan istişareden sonra etiketleme işlemi iki küratör için yaklaşık 30 adet makale özeti ile başlamıştır.

7-Adım: Protokoldeki adımlar gerçekleştirildikten sonra küratörler arası tutarlılık oranı ölçülmüştür. İki küratör arası uyum skoru %92 çıkmıştır. Bu skor kappa skoruna göre (%82'den yukarı) başarılı olarak değerlendirildiğinden geriye kalan makale özetleri iki küratör için ayrı olarak etiketlenmiştir.

Son olarak iki küratör tarafından ayrı ayrı etiketlenen veri setleri bir araya getirilerek birleştirilmiştir. Böylece toplamda 155 adet makale özetinde bulunan aday bitki ve ilaç varlıklar arasındaki ilişkiler “POSITIVE, NEGATIVE ve UNCERTAIN” olarak etiketlenerek altın standart bir derlem geliştirilmiştir. Bu derlem daha sonra jsonl formatında dışarı çıkartılmıştır.

6.3.2 İlişki türü etiketli altın standart derlemlerin geliştirilme süreci

Bu derlem ise ilişkilerin türlere göre ayrılması için alan uzmanlarının yaklaşık 100 adet makaleyi etiketledikleri bir derlem olarak gerçekleştirilmiştir. İlişki türünü etiketlemek ilişki tespitine göre daha karmaşık, daha çok zaman alan ve sadece alan uzmanı bilgisi gerektiren bir süreçten oluşmuştur. Bu derlemde bir önceki bölümdeki protokole göre hazırlanmıştır. İlişki türleri şu şekilde kararlaştırılmıştır.

ADMET: Bitki ve ilaçlar arasında meydana gelen Emilim (Absorbation), Dağılım (Distributon), Metabolizma (Metabolism), Eliminasyon (Elimination) ve Tokistite ilişkileri bu kategori altında toplanmıştır.

POTANTIATIVE: Bitki ve ilaçlar arasında potansiyel ilişkiler bu kategori altında toplanmıştır.

ADVERSE: istenmeyen etkileşimleri belirtir.

SYNERGISTIC: $1+1>2$ ilişkisi

SUMMATIVE: $1+1=2$ ilişkisi

ANTAGONIST: $1+1<2$ ilişkisi

ADDITIVE : $1+1=2$ ilişkisi

NO_INTERACTION: Bir bitki ve bir ilaç arasında herhangi bir ilişkisi yok ise bu kategori kullanılmıştır.

Protokoldeki adımlar gerçekleştirildikten sonra küratörler arası tutarlılık oranı ölçülmüştür. Bu skor %82 gibi bir sonuç vermiştir. Bu da iki alan uzmanı tarafından yüksek tutarlılık oranı olduğu için etiketlenen 100 adet makale özeti altın standart olarak belirlenmiştir.

6.4. İlişkilerin Çıkarım Modelinin Geliştirilme Süreci

Tez çalışmasının bu bölümünde biyomedikal literatürde geçen aday ilaç ve bitki varlıkları arasında bir ilişkinin olup olmaması ve ilişkinin türünü bulabilecek ve otomatik olarak yakalayabilecek en iyi performansa sahip derlem tabanlı bir ilişki çıkarım (RE) modelinin geliştirilmesi amaçlanmıştır. Bunun için ilişki tespiti ve ilişki türü bulma modelleri geliştirilmiştir.

1-İlişki tespit modeli: İlişkinin var olup olmasını yakalamak için ilk olarak ilişki tespit derlemi eğitim veri seti olarak ayarlanmıştır. İlişki çıkarım modeli için makina öğrenmesi ve dil modelleri performansları birbiri ile karşılaştırılmıştır. Makina öğrenmesi için KNN, SVM ve Random Forest algoritmaları kullanılırken son teknoloji dil modellerinden ise BERT ve biyomedikal literatür üzerinde eğitilmiş bir model olan BioBERT kullanılmıştır. Çıkan sonuçlar birbiri ile karşılaştırılarak en iyi performansı göstermesi belirlenen model temel model olarak ele alınmıştır.

2- İlişki Türü modeli: İlişki türünün otomatik olarak yakalanabilmesi için İlişki türü etiketli derlem eğitim veri seti olarak kullanılmıştır ve BioBERT modeli ile ince ayarlanarak temel bir model geliştirilmiştir.

Bu modellerin yapılandırılması ve çalıştırılması şu şekilde olmuştur.

6.4.1 SVM, KNN ve Random Forest ve TF-IDF ile model oluşturma

Bu çalışmada ilk olarak ilişki tespit derlemi eğitim veri seti olarak kullanılmıştır. Daha sonra bu derlemdeki cümleler ilk olarak Tf-Idf vektörleri ile vektörize edilmiştir. Derlem %90 eğitim %10 test olacak şekilde ayrılmıştır. Eğitim veri seti SVM, KNN ve Rastgele Orman ile eğitilerek ayrı ayrı 3 model oluşturularak ilişkiler bulunmaya çalışılmıştır.

6.4.2 BERT ve BioBERT modelleri ile model oluşturma

BERT ve BioBERT dil modelleri için Google Colab ortamı kullanılmıştır. İlk olarak veri seti BERT ve BioBERT ile eğitilebilmek için bu modeller uygun hale getirilmiştir. Derlem %90 eğitim %10 test olacak şekilde ayrılmıştır. Eğitim sırasında öğrenme oranı için $0.00001(1e-5)$ ile $0.005(5e-3)$ arasındaki değerler test edilmiştir. İki model içinde ‘adam’ optimizier kullanılmıştır.

6.4.3 İlişki Çıkarım modellerinin değerlendirilmesi

İlişki çıkarımı modellerinin performansını ölçmek için Kesinlik (Precision), Duyarlılık (Recall) ve F-skor (F-score) ölçütleri kullanılmıştır.

Gerçek Pozitif (True Positive): Model, etiketlenmiş derlemde var olan bir ilişki bulur.

Gerçek Negatif (True Negative): Bir ilişki etiketin olmaması, etiketlenmiş derleme göre doğrudur;

Yanlış Pozitif (False Positive): Model, etiketlenmiş derlemde bulunmayan bir ilişki etiketi sağlar;

Yanlış Negatif (False Negative): Model, etiketlenmiş derlemde bulunan bir ilişki etiketi sağlamaz.

$$\text{Kesinlik} = \frac{\text{Doğru tespit edilen ilişkiler}}{\text{Tespit edilen toplam ilişkiler}} \quad (6.7)$$

$$\text{Kesinlik} = \frac{\text{Gerçek Pozitif}}{\text{Gerçek Pozitif} + \text{Yanlış Pozitif}} \quad (6.8)$$

$$\text{Duyarlılık} = \frac{\text{Doğru tespit edilen ilişkiler}}{\text{Korpusdaki tüm ilişkiler}} \quad (6.9)$$

$$\text{Duyarlılık} = \frac{\text{Gerçek Pozitif}}{\text{Gerçek Pozitif} + \text{Negatif}} \quad (6.10)$$

$$\text{F – Skor} = 2 \times \frac{\text{Kesinlik} \times \text{Duyarlılık}}{\text{Kesinlik} + \text{Duyarlılık}} \quad (6.11)$$

6.5 Zayıf Denetimli Öğrenme (SNORKEL) Modeli İle İlişki Çıkarımı

Çalışmanın bu bölümünde kural tabanlı ve makina öğrenmesi tabanlı bir modelin birleşiminden oluşan zayıf denetimli öğrenme modeli gerçekleştirilerek bir model oluşturulmuştur. Bunun için Snorkel (2022) tabanlı bir uygulama gerçekleştirilmiştir. Snorkel, büyük eğitim veri kümeleri oluşturmak ve metinden varlık ve ilişkiler için tipik doğal dil işleme ilişki çıkarım görevleri sağlayan bir çerçevedir. Snorkel'in ana avantajı, bir görev veya şema değiştiğinde yeniden

etiketlenmesi, genişletilmesi veya göz ardı edilmesi gerekebilecek elle etiketlenmiş veri kümeleri gerektirmemesidir.

Snorkel boru hattı Şekil 6.12'de gösterilmiş ve aşağıda açıklanmıştır. Özetle, uygulama girdi olarak metinler bütünü alır ve bazı ilişki veya ilişki tanımlarına uyan varlık çiftlerini çıkarmaya çalışır. Çalışmanın adımları şu şekildedir:

1.Adım veri setinin belirlenmesi: İlişki var yok derlemi veri seti olarak belirlenmiştir.

2.Adım aday varlıkların belirlenmesi: RoBERTa dil modeli ile eğitilen AVT modeli aday cümleleri belirlemek için kullanılmıştır.

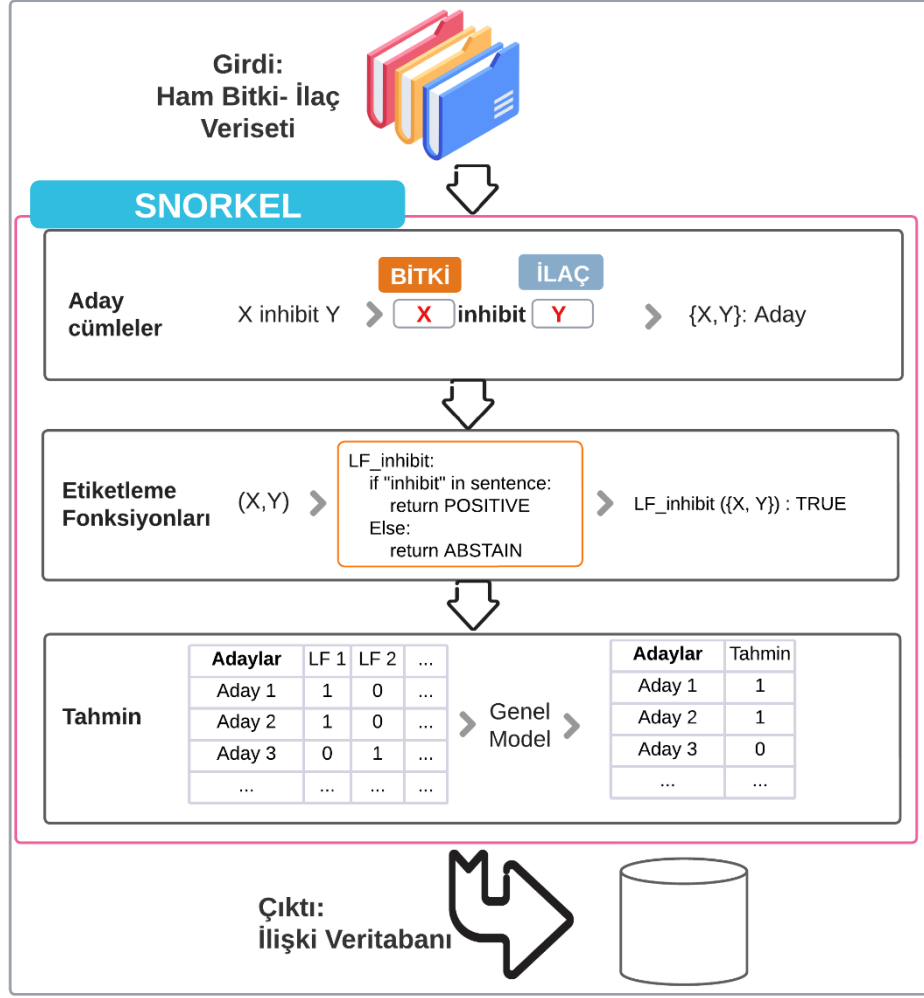
3.Model İlişki Tiplerinin belirlenmesi: Bu model ile 3 tür ilişki belirlenmesi amaçlanmıştır:

ABSTAIN = -1 (belirsiz veya ilişki ile ilgili bir kelime grubunun geçmediği durumlar)

POSITIVE = 1 (Bir ilişkinin tespiti durumu)

NEGATIVE = 0 (Herhangi bir ilişkinin olmama durumu)

“ABSTAIN” iki varlık arasında herhangi bir ilişkinin olmadığı belirsiz ilişki tipidir. Bir önceki derlemde yer alan UNCERTAIN ilişki tipidir. “POSITIVE” ilişki ise farmakodinamik ve farmakokinetik bir ilişkinin olduğu ilişki türünü belirtir. “NEGATIVE” ise iki varlık arasında yapılan araştırmalar sonucunda herhangi bir ilişkinin tespit edilmediği etkileşimleri belirlemek için kullanılmıştır.



Şekil 6.12. Zayıf denetimli öğrenme çerçevesi SNORKEL modeli ile ilişki çıkarımı

4.Etiket Fonksiyonlarının (Labeling Function) Belirlenmesi: Bu fonksiyonlar kelime ve kelime öbeklerinden oluşan kurallardan oluşur. Her etiket tipi için kurallar yazılır ve bu kurallar makina öğrenmesi algoritmalarının ilişki tahmini için öznitelik olarak kullanılır. İlişkileri belirlemek için alan uzmanı küratörler ile birlikte çalışılmıştır. Bir cümlede geçen iki aday varlık arasında olası bir ilişkinin olmasını sağlayacak toplam 21 adet etiket fonksiyonu (Labeling Function) POSITIVE ve ABSTAIN olarak programlanmıştır. Çizelge 6.4’de POSITIVE ilişki türü için 21 adet kuraldan (etiketleme fonksiyonu) 6’sına ait örnekler verilmiştir. Bu örnekler daha detaylı olarak aşağıda açıklanmıştır:

- **def interaction_lf(x):** Eğer iki aday varlığın bulunduğu cümlede “interaction” kelimesi geçiyor ise POSITIVE geçmiyor ise ABSTAIN olarak değerlendirilir.

- **def side_effect_lf(x):** Eğer iki aday varlığın bulunduğu cümlede “side” ve “effect” veya “side effect” kelimeleri birlikte geçiyor ise bu kurala göre POSITIVE geçmiyor ise ABSTAIN olarak değerlendirilir.
- **def synergistic_lf(x):** Eğer iki aday varlığın bulunduğu cümlede “synergistic” kelimesi geçiyor ise bu kurala göre POSITIVE geçmiyor ise ABSTAIN olarak değerlendirilir.
- **def potential_lf(x):** Eğer iki aday varlığın bulunduğu cümlede “potential” kelimesi geçiyor ise POSITIVE ilişki geçmiyor ise ABSTAIN ilişki türü olarak değerlendirilir.

Çizelge 6.4. POSITIVE ve ABSTAIN için belirlenen kurallardan bazıları

İLİŞKİ TÜRÜ	ETİKETLEME FONKSİYONU (LABELING FUNCTION)
POSITIVE-ABSTAIN	<pre>@labeling_function() def interaction_lf(x): matches = re.findall(r'interaction[s]*', x.text) if len(matches) > 0: return POSITIVE return ABSTAIN lfs.append(interaction_lf)</pre>
	<pre>@labeling_function() def side_effect_lf(x): matches = re.findall(r'side[-]*effect[s]*', x.text) if len(matches) > 0: return POSITIVE return ABSTAIN lfs.append(side_effect_lf)</pre>
	<pre>@labeling_function() def synergistic_lf(x): matches = re.findall(r'synergistic', x.text) if len(matches) > 0: return POSITIVE return ABSTAIN lfs.append(synergistic_lf)</pre>
	<pre>@labeling_function() def inhibit_lf(x): matches = re.findall(r'inhibit', x.text) if len(matches) > 0: return POSITIVE return ABSTAIN lfs.append(inhibit_lf)</pre>
	<pre>@labeling_function() def additive_effect_lf(x): matches = re.findall(r'additive[-]*effect[s]*', x.text) if len(matches) > 0: return POSITIVE return ABSTAIN</pre>

	lfs.append(side_effect_lf)
	<pre>@labeling_function() def interfere_lf(x): matches = re.findall(r'interfere[s]*', x.text) if len(matches) > 0: return POSITIVE return ABSTAIN lfs.append(interfere_lf)</pre>

NEGATIVE ilişki türü için ise Çizelge 6.5'te verilen kurallar yazılmıştır. Bu kurallardan bazıları şu şekilde açıklanmıştır.

def no__observed_lf(x): Aday varlıkları arasında “no” ve “observe” kelimeleri birlikte geçiyor ise NEGATIVE ilişki tipi değil ise ABSTAIN ilişki tipi olarak değerlendirilir.

def did__not_lf(x): Aday varlıkları arasında “did” ve “not” varlıkları birlikte geçiyor ise NEGATIVE ilişki tipi değil ise ABSTAIN ilişki tipi olarak değerlendirilir.

Çizelge 6.5. NEGATIVE ve ABSTAIN için belirlenen kurallardan bazıları

İLİŞKİ TÜRÜ	ETİKETLEME FONKSİYONU (LABELING FUNCTION)
NEGATIVE-ABSTAIN	<pre>@labeling_function() def no__observed_lf(x): matches = re.findall(r'no.*observe', x.text) if len(matches) > 0: return NEGATIVE return ABSTAIN lfs.append(no__observed_lf)</pre>
	<pre>@labeling_function() def did__not_lf(x): matches = re.findall(r'did.*not', x.text) if len(matches) > 0: return NEGATIVE return ABSTAIN lfs.append(did__not_lf)</pre>
	<pre>@labeling_function() def aim_of_the__(x): matches = re.findall(r'aim.*of.*the', x.text) if len(matches) > 0: return NEGATIVE return ABSTAIN lfs.append(aim_of_the__)</pre>

6.Adım: Bu adımda ilişki çıkarımı için iki model uygulanır. Birincisi “MajorityLabelVoter” ikincisi ise “Label Model”. Etiketleme fonksiyonları yazıldıktan sonra amaç etiketleme fonksiyonlarından (LF) gelen etiketleri, güven ağırlıklı etikete dönüştürmektir. Bunu yapmak için en temel görev veri noktası bazında çoğunluk oyu almaktır: Örneğin bir cümle için POSITIVE etiketten daha fazla NEGATIVE etikete oy verilmiş ise bu cümle NEGATIVE olarak etiketlenir (veya tersi). Bunun için MajorityLabelVoter temel modeli kullanılır ve test edilir.

LABEL model ise daha karmaşık ve “LF” doğruluklarını öğrenmek ve çıktı etiketlerini birleştiren bir modeldir. Bu sınıf, etiketleme fonksiyonlarının gerçek (gözlemlenmemiş) Y , $P(L_f | Y)$ etiketinin çıktısını almanın koşullu olasılıklarının bir modelini öğrenir ve bu öğrenilen modeli çıktı etiketlerini yeniden ağırlıklandırmak ve birleştirmek için kullanır. Eğitim sürecinde altın standart etiket kullanılmamaktadır. Bu yöntemde ihtiyaç olan tek bilgi, eğitim setindeki LF'lerin çıktısını içeren etiket matrisidir. Label Model, girdi olarak yalnızca etiket matrisini kullanarak etiketleme işlevleri için ağırlıkları öğrenebilir

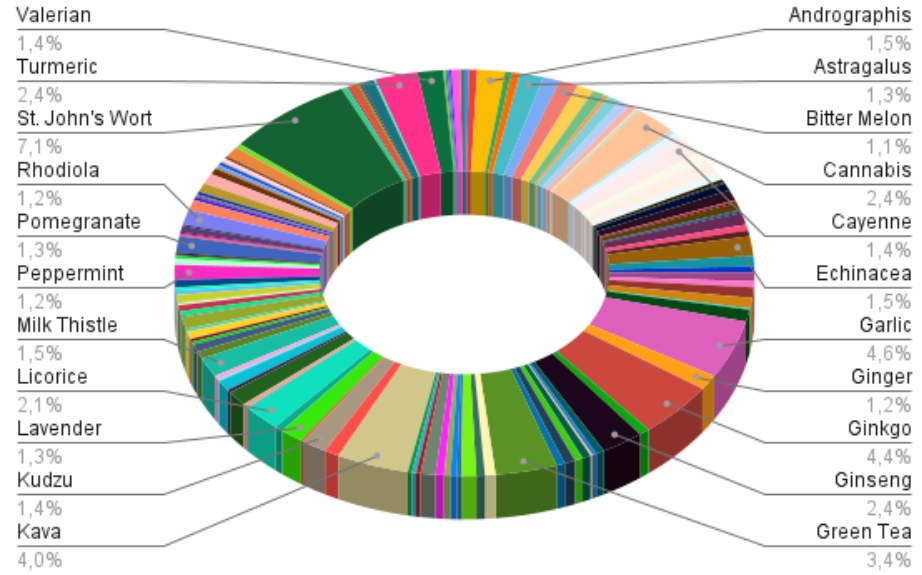
7.Adım: Bu adımda ise ilişki tespiti olarak etiketli altın standart ilişki tespit derlemi test seti olarak kullanılmıştır ve “majority” ve “label” modellerinin doğrulukları ölçülmüştür. Değerlendirme için doğruluk (accuracy) ölçütü kullanılmıştır.

7.DENEYSEL BULGULAR

Bu bölüm 5 aşamada ele alınmıştır. MSKCC veri kaynağından alınan veri kaynağı ile ilgili sonuçlar, varlık etiketli altın standart derlem etiketleme süreci ile ilgili sonuçlar, AVT modeli sonuçları, ilişki çıkarımı için etiketli altın standart derlem süreci ile ilgili sonuçlar ve ilişki etiketleme modeli geliştirme süreci ile ilgili sonuçlardan oluşmaktadır.

7.1 MSKCC Veri Kaynağından Elde Edilen Sonuçlar

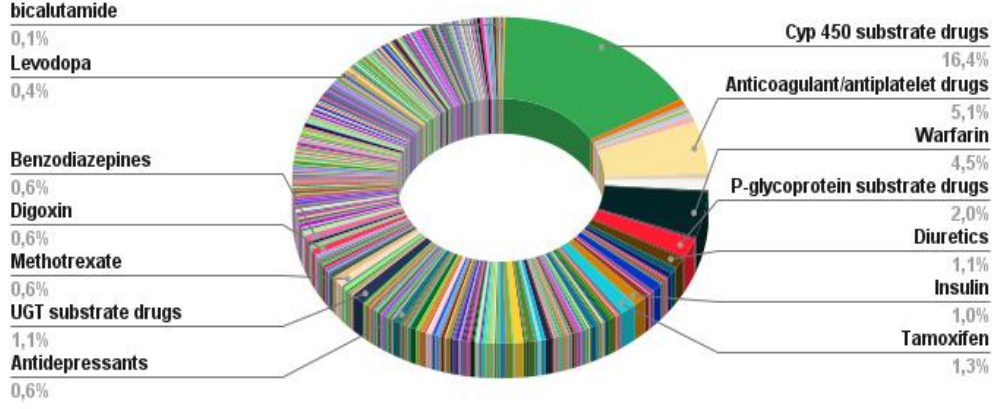
MSKCC veri kaynağı ile ilgili yaygın olarak ele alınan bitki ve ilaç terimleri ile ilgili grafik düzeyinde bir çalışma yapılmıştır. Bu analiz etkileşime giren en sık ilaç ve bitki varlıkları ile ilgili bir fikir oluşturmayı amaçlamıştır. Buna göre Şekil 7.1’de görüldüğü üzere üzerinde en çok çalışma yapılan bitkiler sırasıyla “St. John Wort (sarı kantaron), Garlic (sarımsak), Ginkgo (Mabet ağacı), Kava, Green tea (yeşil çay), Ginseng, Cannabis (hint kenevir), Turmeric (zerdeçal) ve licorice” bitkileridir.



Şekil 7.1. İlaçlarla en çok etkileşime giren bitki terimleri

Mskcc veri kaynağında elde edilen ve bitkiler ile etkileşime giren en yaygın ilaçlar üzerine yapılan analize bakıldığında ise daha çok ilaç sınıfı düzeyinde etkileşim bilgisi vardır. En çok etkileşime girenler ilaçlar sırasıyla Şekil 7.2’de

görüldüğü üzere “Cyp 450 substrate drug (Cyp 450 enzim metabolize ilaçları, Anticoagulant/Antiplatelet drugs (Pıhtılaşma önleyici ilaçlar), Warfarin (aslında pıhtılaşma önleyici bir ilaçtır), UGT substrate drugs (UGT enzimini metabolize eden ilaçlar), Tamoxifen ve Diuretic” gibi ilaç sınıfları olarak belirtilmiştir.



Şekil 7.2. Bitkiler ile en çok etkileşime giren ilaç ve sınıfları

7.2 BİE için Varlık Etiketli Altın Standart Derlemde Etiketlenen Varlık Sayıları ve Uyum Skoru Sonuçları

Bu çalışmada toplam 680 adet makale özeti DRUG, HERB, HERB_COMPOUND ve HDI_SENTENCES olarak etiketlenmiştir. Çizelge 7.1 bu açıklamalı varlıkların sayısını ve türünü gösterir.

Çizelge 7.1. BİE altın standart derleminde etiketlenen varlık sayısı

Varlık Türü	Toplam Etiketleme Sayısı	Yüzdesi
DRUG	4291	%34,8
HERB	4382	%35,6
HERB COMPOUND	1053	%8,5
HDI_SENTENCES	2599	%21,1
TOTAL	12325	

Yukarıdaki çizelgeye göre küratörlerin en çok açıklama ekledikleri (etiketledikleri) varlık tipleri sırasıyla HERB (4382) ve DRUG (4291) varlıklarıdır. Açıklama eklenen en az sayıdaki varlık tipi ise 1053 adet ile HERB_COMPOUND (bitki bileşenleri) olmuştur. Tüm varlık sayısı içinde yüzdelere bakıldığında ise HERB %35,6 ve DRUG %34,8 ile birbirlerine oldukça yakın oranda oldukları gözlemlenmiştir.

Çizelge 7.2’de ise altın standart derlemi hazırlayan iki alan uzmanı küratörlerin birbiriyle tutarlılığını gösteren uyum skoru (IAA-inter anotator agreement) sonuçları yer almaktadır. Bu sonuçlara HERB_COMPOUND ve HDI_SENTENCES varlıkları dahil edilmemiştir. Bu varlıklar derleme alan uzmanları tarafından ortak olarak eklenen varlık tipleridir.

Çizelge 7.2. Varlık etiketli altın standart derlem oluşturma sürecinde küratörler arası uyum oranı

Varlık Türü	Uyum Oranı
DRUG	0,91
HERB	0,95
TOPLAM	0.93

7.3 AVT Model Sonuçları

Biyomedikal literatürden bitki ve ilaç varlıklarını (terimlerini) yakalayabilecek en iyi model için Tok2vec ve RoBERTa modellerinin başarı performansları karşılaştırılmıştır.

Çizelge 7.3 incelendiğinde Tok2Vec modelinin kesinlik (precision) ve geri çağırma (recall) ölçütlerinin harmonik ortalaması olan F-Score başarısı %85,472 oranında çıkmıştır. HERB varlığı kesinlik (precision) ve geri çağırma ve F-score performansı bakımından en iyi sonucu verirken. HERB_COMPOUND varlığı ise bu 3 ölçütte en düşük başarı performansı gösteren varlık türü olmuştur. DRUG varlığı ise F-score bakımından %81,786 başarı göstermiştir.

Çizelge 7.3. Tok2vec Modelinin AVT sonuçları

Varlık Türü	Precision (Kesinlik)	Recall (Geri çağırma)	F-Score
DRUG	83,154	80,462	81,786
HERB	91,622	92,617	92,117
HERB COMPOUND	76,531	68,807	72,464
HDI SENTENCES	81,993	87,992	84,887
ORTALAMA F-SCORE			85,472

Çizelge 7.4’de ise RoBERTa modelinin kesinlik, geri çağırma ve f-score sonuçları verilmiştir. Bu sonuçlara göre ’de en yüksek değer %95,345 ile HERB varlığı olmuştur.

Çizelge 7.4. RoBERTa Modelinin AVT sonuçları

Varlık Türü	Precision (Kesinlik)	Recall (Geri çağırma)	F-Score
DRUG	92,378	89,168	90,744
HERB	94,350	96,361	95,345
HERB COMPOUND	87,260	81,146	84,092
HDI SENTENCES	88,962	92,341	90,620
ORTALAMA	90,730	90,402	90,566

7.4 BİE için İlişki Etiketli Altın Standart Derlem

Çizelge 7.5’te aday varlıklar arasında ilişki tespit altın standart derlemi için etiketlenen ilişki türlerinin sayısı görülmektedir. Buna göre en çok etiketlenen ilişki türü 589 adet ile POSITIVE ilişki türü olurken, en az etiketlenen ilişki türü 53 adet

ile NEGATIVE olmuştur. Yüzdelik olarak bakıldığında etiketlenen POSITIVE %48,35, UNCERTAIN %47,29 etiketleme sayısı ile birbirine yakın iken NEGATIVE ise %4,35 ile çok daha az bir etiketleme sayısına sahiptir.

Çizelge 7.5. BİE altın standart ilişki tespit derlemde etiketlenen varlık sayısı

İlişki Türü	Toplam Etiketleme Sayısı	Yüzdesi
POSITIVE	589	%48,35
UNCERTAIN	576	%47,29
NEGATIVE	53	%4,35
TOPLAM	1218	

7.5 BİE İçin Tespit İlişki Etiketli Model Performansları

Dönüştürücü (Transformer) tabanlı BioBERT modeli f-skor olarak %4-%15'in üzerinde iyileştirme ile geleneksel makine öğrenimi modellerinden daha iyi performans göstermiştir. En düşük performansı ise %75 ile en düşük F-skor puanına sahip KNN, vermiştir. Toplam 5 modelin performansları Çizelge 7.6'da verilmiştir.

Çizelge 7.6. İlişki etiketli model performansları

Model	Doğruluk (Accuracy)	Precision (Kesinlik)	Recall (Geri çağırma)	F1
Knn	83%	79%	72%	75%
Svm	91%	95%	78%	82%
Random Forest	92%	92%	84%	86%
Bert	82%	96%	84%	89%
BioBERT	90%	92%	89%	90%

7.6 BİE için İlişki Türü Etiketli Altın Standart Derlem

Çizelge 7.7 ‘de aday varlıklar arasında ilişki türü seviyesinde altın standart derlem için etiketlenen ilişki türlerinin sayısı görülmektedir. Buna göre en çok etiketlenen ilişki türü farmakokinetik ilişki olan 438 adet ile ADMET iken, ANTAGONIST, POTANTİATIVE, SUMMATIVE ve ADDITIVE ilişki türlerinde ise oldukça az sayıda ilişki türü etiketlenmiştir.

Çizelge 7.7. BİE altın standart derleminde etiketlenen varlık sayısı

İlişki Türü	Toplam Etiketleme Sayısı	Yüzdesi
ADMET	438	67,5%
ADVERSE	114	17,6%
SYNERGISTIC	56	8,6%
NO_INTERACTION	26	4,0%
ANTAGONIST	5	0,8%
POTANTİATIVE	3	0,5%
SUMMATIVE	4	0,6%
ADDITIVE	3	0,5%
TOPLAM	646	

7.7 BİE İçin Tür İlişkisi etiketli Altın Standart Derlem

Tip etiketli veri setinin model başarısı BioBERT modeli üzerinde ölçülmüştür. Çizelge 7.8’de görüldüğü üzere BioBERT %72,22 f-skor ile BERT modelinden daha başarılı bir sonuç verdiği gözlemlenmiştir.

Çizelge 7.8. Tür İlişkisi etiketli model performansı

Model	Accuracy (Doğruluk)	Precision (Kesinlik)	Recall	F1
BERT	77,08%	69,65%	75,83%	70,65%
BioBERT	78,12%	70,49%	77,77%	72,22%

7.8 Zayıf Denetimli Öğrenme (SNORKEL) Sonuçları

Snorkel için iki model kullanılmıştır: Label Model ve Majority Model. Çizelge 7.9’da iki modelin başarısı görülmektedir. İki modelinde başarısının oldukça az olduğu gözlemlenmiştir.

Çizelge 7.9. SNORKEL modellerinin performansı

Model	Doğruluk (Accuracy)
Label Model	49.4%
Majority Model	39.2%

8.TARTIŞMA

8.1 Mskcc Veri seti

Bu tez çalışmasında etiketli altın standart derlem için gerekli olan BİE'lere ait makale özetlerinin çoğunluğu MSKCC veri kaynağının verdiği PubMed referanslardan elde edilmiştir. Bu makale özetlerinin çoğu PubMed sorgusu yapılarak da çıkarılabilirdi. Ancak aşağıda yapılan PubMed sorgulamaları;

- “Herb Drug interaction” ,
- “Food Drug interaction”,
- “Plant Drug interaction”

sonucunda toplamda yaklaşık 20 bin üzerinde indekslenmiş makale sonucu çıkmıştır. Yapılan bu aramalarda çıkan ilk makaleler çeşitlilik açısından tatmin edici sonuçlar vermemiştir. Bunun nedeni sorgulamalarda elde edilen ilk makale özetleri genel olarak birbirine benzer bitki ve ilaçlar arasındaki etkileşim bilgileri yer almıştır. Örneğin “St. John Worth” ve “warfarin” ilişkisi üzerine yaygın olarak çalışmalar yapılmıştır.

Çizelge 8.1. PubMed Sorgusu sonucu çıkan bitki-ilaç etkileşim ile ilgili çalışmalar

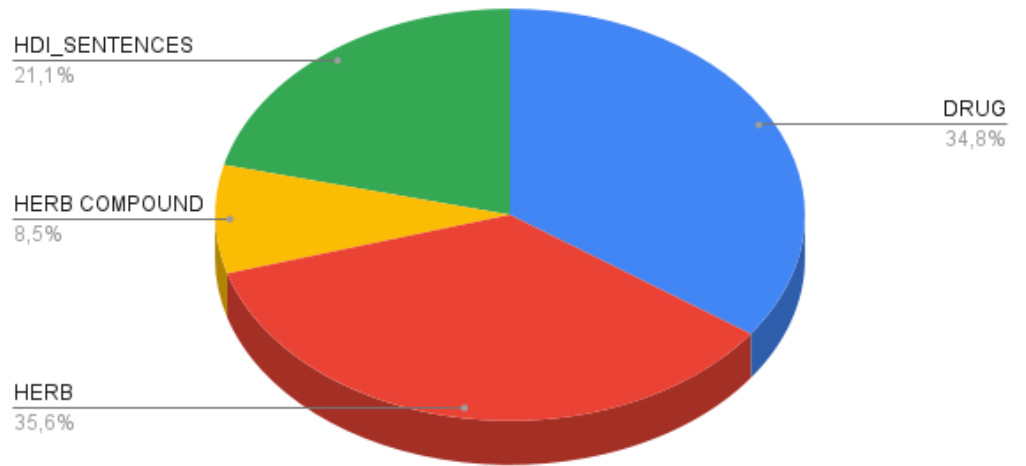
Pubmed İndeksi	PubMed (HERB DRUG interaction) Sorgu Sonucu	Makale Özetinde bulunan Aday Varlıklar
1	Herb-Drug Interactions and Hepatotoxicity.	(Parvez and Rishi, 2019)
2	Herb-drug interactions.	(Fugh-Berman, 2000)
3	Herb-Drug Interactions in Cancer Care.	(Yeung, Gubili and Mao, 2018)
4	Herb-drug interactions: an overview of systematic reviews.	(Posadzki, Watson and Ernst, 2013)
5	Herb-drug interactions: a literature review.	(Hu <i>et al.</i> , 2005)
6	PXR as a mediator of herb-drug interaction.	(Hogle <i>et al.</i> , 2018)
7	Common Herbal Dietary Supplement-Drug Interactions.	(Asher, Corbett and Hawke, 2017)
8	Herb-drug interaction studies of herbs used in treatment of cardiovascular disorders-A narrative review of preclinical and clinical studies	(Shaikh, Thomas and Chitlange, 2020)

9	Overview of Current Herb-Drug Interaction Databases.	(Zhang <i>et al.</i> , 2022)
10	Selected Food/Herb-Drug Interactions: Mechanisms and Clinical Relevance.	(Amadi and Mgbahurike, 2018)

Çizelge 8.1'e bakıldığında PubMed arama sonuçlarından çıkan ilk 10 sonuç görülmektedir. Bu sonuçlar incelendiğinde belirli bir bitki ve ilaç arasındaki etkileşimden ziyade daha genel amaçlı etkileşimlere odaklandıkları gözlemlenmiştir. Bu bağlamda MSKCC veri kaynağı ise "A'dan Z'ye" bir sözlük şeklinde bitkiler ve ilaçlar ile olan etkileşim bilgilerini tutan ve bu etkileşim verilerinin PubMed referanslarını içeren bir kaynak olduğu için bu referanslardan elde edilen makale özetleri etiketleme sürecine çeşitlilik kazandırmıştır. Böylece hem zaman açısından hem de çeşitlilik sağladığından daha kaliteli bir derlem oluşturmayı sağlamıştır.

8.2 Varlık Etiketli BİE Derlemi Üzerine Değerlendirme ve Hata Analizi

Bitki ve ilaç varlıklarının etiketlendiği varlık etiketli derlemin hazırlanması için alan uzmanları seviyesinde etiketleme işlemi gerçekleştirilmiştir. İki alan uzmanı toplam 680 adet makale özetinde bulunan bitki ve ilaç varlıklarını Çizelge 7.1'de görüldüğü üzere bitkileri %95 ilaçları ise %91 tutarlılıkta etiketlemişlerdir. Bu oranlar kappa skoruna göre mükemmel etiketleme olarak düşünülebilir.



Şekil 8.1. BİE derleminde etiketlenen varlıkların denge durumu

Ayrıca toplamda etiketlenen 12,325 adet varlık sayısı az sayıda etiketli veri seti ile oldukça başarılı sonuçlar veren BERT tabanlı dil modellerini eğitmek için yeterli bir sayı olduğu düşünülebilir. Ayrıca Şekil 8.1’de görüldüğü üzere derlemde küratörlerin etiketlediği DRUG (%34,8) ve HERB (%35,6) varlıkları arasında denge (balans) sorununun da olmadığı söylenebilir.

Etiketleme sürecinde alan uzmanı küratörlerin yaygın olarak yaptıkları hatalar aşağıdaki gibidir.

1-Makale özetlerinde geçen bitki ve ilaç varlıklarının farklı sayıda etiketlenmesi: Bu sorun etiketleme aracının aktif öğrenme tabanlı sistemiyle kısmende olsa başarılı bir şekilde çözülmüştür. Ancak özellikle sınıf isimlerinde bazı işaretlemelerin gözden kaçırıldığı gözlemlenmiştir.

2-Makale özetlerinde geçen bitki ve ilaçların hatalı etiketlenmesi: Etiketleme araçları kullanılırken yapılan hatalardan bir tanesi bu olmuştur. Örneğin makale özetlerinde geçen “garlic”bitkisi DRUG olarak ve “warfarin” ilacı ise HERB olarak etiketlenmiştir. Bu yanlışlıklar Prodigy aracının hata düzeltme özelliği sayesinde düzeltilmiştir.

3- Kelime kapsamından kaynaklı hatalar: Özellikle iki veya daha fazla kelimedenden oluşan ilaç ve bitki adlandırmalarında kapsamdan kaynaklı hatalar gözlemlenmiştir. Örneğin cümlede geçen “anticoagulant and antiplatelet drugs” bir küratör kelimeyi birebir etiketlerken bir diğeri ise yalnızca “antiplatelet drugs” olarak etiketleme yaptığı gözlemlenmiştir.

8.3 AVT Modeli İçin Bir Değerlendirme ve Hata Analizi

Biyomedikal literatürden ilaç ve bitki varlıklarını çıkarabilecek performansı en yüksek modeli seçmek için doğal dil işleme için oldukça popüler olan Spacy kütüphanesi aracılığıyla iki model karşılaştırılmıştır. Bu modellerden en başarılı sonucu BERT dil modelinin gelişmiş versiyonu olan RoBERTa vermiştir. RoBERTa her varlık tipi (DRUG, HERB, HERB_COMPOUND ve HDI_SENTENCES) için hem kesinlik hem geri çağırma hem de f-score (90,620) bakımından CNN tabanlı Tok2Vec modelinden daha başarılı performans göstermiştir. Bu nedenle ilişki çıkarma modeli içinde RoBERTa dil modeliyle eğitilen ve saklanan model temel (baseline) olarak kullanılmıştır.

8.4 İlişki Etiketleme Derlemi

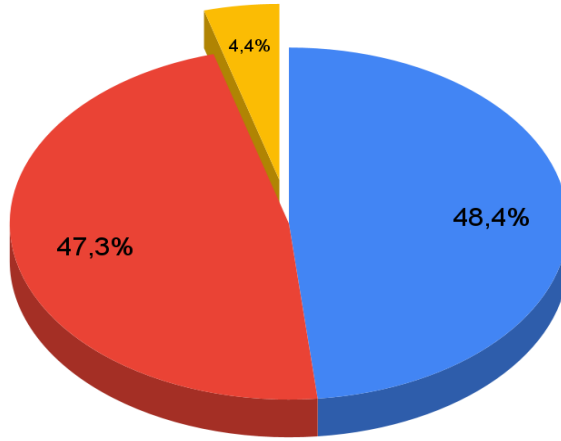
Bu çalışmada 2 adet ilişki etiketli derlem geliştirilmiştir. Bunlardan ilki ilişki tespit (POSITIVE, NEGATIVE ve UNCERATAIN) derlemi diğer ikinci derlem ise ilişki türü etiketlemekten oluşmuştur. İlişki türü derlemi ilişki tespit derlemine göre daha çok uzmanlık alan bilgisi gerektiren bir süreçten oluşmuştur ve bu nedenle daha çok zaman ve daha zahmet veren bir çalışma olmuştur.

İlişki Tespiti Derlemi: Şekil 8.2'ye bakıldığı zaman aday bir bitki ve aday bir ilaç arasında etkileşimin olduğu ilişki türü olan **POSITIVE** ve herhangi bir ilişkinin olmadığı sadece varlıkların birlikte geçtiği durumlar için kullanılan **UNCERTAIN** ilişki türü için bir denge söz konusudur. Aday varlıklar için araştırılan ve hiçbir olumlu ya da olumsuz ilişkinin bulunmadığı durumlar için kullanılan **NEGATIVE** ilişki tipi ise %4.4 gibi oldukça küçük kalmaktadır ve burada bir dengesizlik söz konusudur denilebilir. Ancak makalelerde genel olarak yapılan çalışmalarda herhangi bir ilişki yok ise;

- “birlikte kullanmayınız”,
- “dikkatli olunuz”

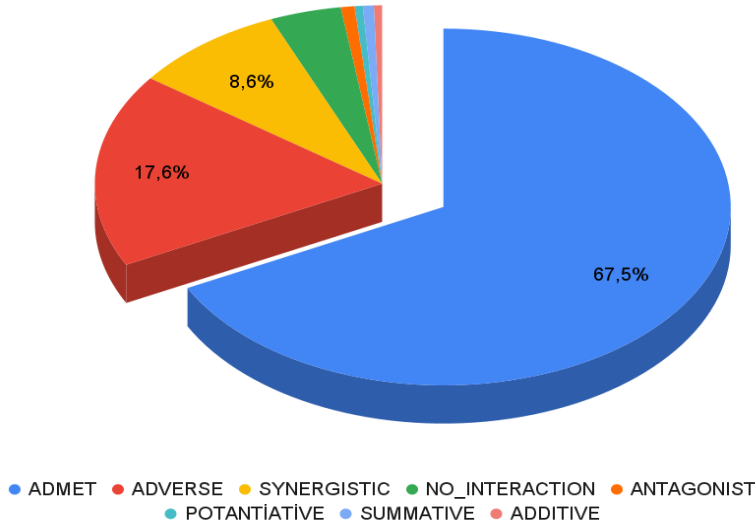
gibi tavsiye niteliğinde cümleler kullanılmıştır. Ancak tavsiye niteliğinde kullanılan cümlelerden bazıları da gerçekten etkileşim bulunan tipler içinde geçerli olduğundan tavsiyeler POSITIVE ilişki tipi içinde gösterilmiştir. Bu nedenle NEGATIVE ilişki türüne çok az çalışmada rastlanmıştır. Bir diğer ilişki türü olan UNCERTAIN ilişki türü ise iki ilaç arasında herhangi bir etkileşim cümlesinin geçmediği ya da speculative durumların olduğu tipler için kullanılmıştır. Örneğin aşağıdaki cümleler UNCERTAIN ilişki tipine girmiştir.

- The aim of this study was to evaluate...
- To study the effects of the...
- To investigate the ability of ...
- The interactions of natural products with drugs are discussed...
- To determine whether...
- To clarify whether



Şekil 8.2. POSITIVE (mavi), UNCERTAIN (kırmızı) ve NEGATIVE (sarı) ilişkilerin denge durumu

İlişki Türü Etiketli Derlem: Bitki ve ilaçlar arasında ilişki türü seviyesinde bir etiketleme gerçekleştirilmiştir. Bu ilişki türü ADMET, ADVERSE, SYNERGISTIC, NO_INTERACTION, ANTAGONIST, POTANTİATİVE, SUMMATIVE ve ADDITIVE olmak üzere 8 adet ilişki tipinden oluşan bir derlemden oluşmuştur.



Şekil 8.3. İlişki Türü Etiketli derlemin denge durumu

İlişki Türü Etiketli derlemin denge durumu Şekil 8.3’de verilmiştir. Şekilde görüldüğü üzere bitki ve ilaçlar arasında meydana gelen ilişkilerden %67,5 ile farmakokinetik seviyede olan ADMET (Emilim, dağılım, metabolizma, eliminasyon ve toksisite) ilişkileri olarak etiketlenmiştir. Farmakodinamik seviyede

ise oldukça az sayıda etiketleme yapılmıştır. Bu nedenle gelecekte yapılacak çalışmalarda farmakodinamik etkileşimler tek bir çatı altında toplanabilir.

8.5 İlişki Çıkarım Modeli

İlişki tespiti (var-yok) Modeli: İlişki çıkarım modeli için geleneksel makina öğrenmesi modellerinden Random Forest (86% F-score), SVM (82% F-score) ve KNN (75% F-score) algoritmaları ile son teknoloji doğal dil işleme tekniklerinden BERT (89% F-score) ve BioBERT (90% F-score) kullanılarak bir model oluşturulmuştur. Bu modellerden en yüksek sonucu BioBERT vermiştir. Bu beklenen bir skordur. Çünkü BİE derlemi bir biyomedikal literatür derlemi olduğu için ve BioBERT modeli biyomedikal literatür üzerinden eğitildiği için en başarılı performansı vermesi beklenen bir skor olmuştur.

İlişki Türü Modeli: İlişki türü etiketleme ise yalnızca BioBERT modeli üzerinde eğitilmiştir. Bir önceki çalışmada BioBERT (%81) en iyi skoru verdiği için bu çalışmada da BioBERT üzerinden veri seti ince ayarlanmıştır ve %81 f-skor puanına sahip bir model elde edilmiştir.

8.6 Zayıf Denetimli Öğrenme (SNORKEL) Modeli ve Hata Analizi

SNORKEL modeli için toplamda 27 adet kural geliştirilmiştir. Bu kurallardan 21 tanesi POSITIVE ve 6 tanesi ise NEGATİF sonuçlar için hazırlanmıştır. SNORKEL kural ve makina öğrenmesi modellerinin bir arada kullanıldığı ve az etiketleme ile başarılı sonuç vermesi beklenen bir zayıf denetimli öğrenme tekniğidir. Ancak Snorkel öğrenme modeli sonuçları karşılaştırıldığında BERT ve BioBERT modellerine göre oldukça yetersiz kaldığı gözlemlenmiştir.

Label Model: 49,4 % Doğruluk

Majority Model: 39,2 Doğruluk

SNORKEL 'in başarısının artırılabilmesi için kural sayısının artırılması ve veri sayısının artırılması gerekmektedir. Bu zahmet veren ve oldukça zaman alıcı bir süreçtir. Ancak BioBERT gibi son teknoloji yöntemler bu eksiklikleri giderdiği gözlemlenmiştir. Bu bağlamda BioBERT ile ince ayarlanmış model gösterdiği yüksek performans neticesi ile temel (baseline) model olarak kabul görmüştür.

9. SONUÇ VE ÖNERİLER

Bu tez çalışmasında literatürdeki büyük eksikliklerden biri olan bitki ilaç etkileşimi ile ilgili farmakokinetik ve farmakodinamik ilişkiler gibi faydalı bilgilerin her geçen gün artan milyonlarca makalenin yer aldığı biyomedikal literatürden çıkarılabilmesi için varlık ve ilişki etiketli altın standart derlemler hazırlanmıştır. Kaliteli bir altın standart derlem oluşturma süreci farklı disiplinlerin bir araya geldiği yoğun emek ve zaman isteyen bir çalışmaya dayanmaktadır. Bu tez çalışması da altın standart derlem oluşturma sürecinde bilgisayar ve biyokimya mühendisliği bölümlerinin ortaklaşa yaptıkları bir çalışma ile ortaya çıkmış olduğundan disiplinlerarası bir çalışmadır.

Bu derlemler kullanılarak adlandırılmış varlık tanıma ve ilişki çıkarım modelleri geliştirilmiştir. Bu modeller SVM, KNN, Rastgele orman algoritmaları ve son teknoloji BERT, BioBERT, RoBERTa ve Tok2vec modelleri ile eğitilerek performansı en yüksek model elde edilmeye çalışılmıştır. Adlandırılmış varlık tanıma için RoBERTa modeli % 90 başarı göstererek en iyi performansı göstermiştir. İlişki tespiti (POSITIVE, NEGATIVE) ve ilişki türü (farmakokinetik ilişkiler ve farmakodinamik ilişkiler) modelleri için ise BioBERT dil modeli sırasıyla %90 ve %72,22 ile en iyi performansı göstermiştir.

Ayrıca ilişki çıkarımı için makina öğrenmesi ve kural tabanlı yöntemlerin birleştiği bir sistem de geliştirilmiştir. Gelecek çalışma bu kural tabanlı modelin BERT tabanlı modeller ile birleştirilerek başarı performansının artırılması olabilir.

Son olarak çalışma ile ilgili geliştirilen derlemlere, modellere ve tablolara https://github.com/erkanotales/bert_herb_drug adresinden erişilebilmektedir.

KAYNAKLAR DİZİNİ

- Acibuca, V. and Budak, D.B.** (2018) ‘Dünya’da ve Türkiye’de tıbbi ve aromatik bitkilerin yeri ve önemi’, *Çukurova Tarım ve Gıda Bilimleri Dergisi*, 33(1), pp. 37–44.
- Adnan, K. and Akbar, R.** (2019) ‘Limitations of information extraction methods and techniques for heterogeneous unstructured big data’, *International Journal of Engineering Business Management*, 11, p. 1847979019890771.
- Akhtar, K., et al.** (2018). Use of alternative medicine is delaying health-seeking behavior by Bangladeshi breast cancer patients. *European journal of breast health*, 14(3), 166.
- Al-Ash, H.S. and Wibowo, W.C.** (2018) ‘Fake News Identification Characteristics Using Named Entity Recognition and Phrase Detection’, in *2018 10th International Conference on Information Technology and Electrical Engineering (ICITEE)*, pp. 12–17.
- Alshaikhdeeb, B. and Ahmad, K.** (2016). Biomedical named entity recognition: a review. *International Journal on Advanced Science, Engineering and Information Technology*, 6(6), 889-895.
- Amadi, C.N. and Mgbahurike, A.A.** (2018) ‘Selected Food/Herb-Drug Interactions: Mechanisms and Clinical Relevance’, *American journal of therapeutics*, 25(4), pp. e423–e433.
- Ang-Lee, M.K., Moss, J. and Yuan, C.S.** (2001) ‘Herbal medicines and perioperative care’, *JAMA: the journal of the American Medical Association*, 286(2), pp. 208–216.
- AnyDesk**, <https://anydesk.com/tr> (Erişim tarihi: 9 Ekim 2022).
- Asiimwe, J. B., et al.** (2021). Prevalence of the use of herbal medicines among patients with cancer: A systematic review and meta-analysis. *Evidence-Based Complementary and Alternative Medicine*, 2021.
- Asher, G.N., Corbett, A.H. and Hawke, R.L.** (2017) ‘Common Herbal Dietary Supplement-Drug Interactions’, *American family physician*, 96(2), pp. 101–107.
- Barnes, J., Anderson, L.A. and Phillipson, J.D.** (2001) ‘St John’s wort (*Hypericum perforatum* L.): a review of its chemistry, pharmacology and clinical properties’, *The Journal of pharmacy and pharmacology*, 53(5), pp. 583–600.
- Barthelmeh, M.** (2011) ‘Journeys: An interpretive critique of the Christchurch railway station’, *Spaces & flows: an international journal of urban and extraurban studies*, 1(1), pp. 37–54.
- Baum, L.E. and Petrie, T.** (1966) ‘Statistical Inference for Probabilistic Functions of Finite State Markov Chains’, *Annals of Mathematical Statistics*, 37(6), pp. 1554–1563.
- Beautifulsoup4**, <https://pypi.org/project/beautifulsoup4/> (Erişim tarihi: 5 Ekim 2022).
- Beltagy, I., Lo, K. and Cohan, A.** (2019) ‘SciBERT: A Pretrained Language Model for Scientific Text’, *arXiv [cs.CL]*. Available at: <http://arxiv.org/abs/1903.10676>.

KAYNAKLAR DİZİNİ (devam)

Bent, S. et al. (2005) ‘Spontaneous bleeding associated with ginkgo biloba: a case report and systematic review of the literature: a case report and systematic review of the literature’, *Journal of general internal medicine*, 20(7), pp. 657–661.

Bhagavatula, M., Gsk, S. and Varma, V. (2012) ‘Named entity recognition an aid to improve multilingual entity filling in language-independent approach’, in *Proceedings of the first workshop on Information and knowledge management for developing region*. New York, NY, USA: Association for Computing Machinery (IKM4DR ’12), pp. 3–10.

Brat annotation tool (2022), <https://brat.nlplab.org>, (Erişim tarihi: 09 Temmuz 2022)

Borse, S.P., Singh, D.P. and Nivsarkar, M. (2019) ‘Understanding the relevance of herb-drug interaction studies with special focus on interplays: a prerequisite for integrative medicine’, *Porto biomedical journal*, 4(2), p. e15.

Bose, P. et al. (2021) ‘A Survey on Recent Named Entity Recognition and Relationship Extraction Techniques on Clinical Texts’, *NATO Advanced Science Institutes series E: Applied sciences*, 11(18), p. 8319.

Budi, I. and Bressan, S. (2003) ‘Association rules mining for name entity recognition’, in *Proceedings of the Fourth International Conference on Web Information Systems Engineering, 2003. WISE 2003.*, pp. 325–328.

Buneman, P. (1997). Semistructured data, in ‘PODS’97: Proceedings of the 16th ACM SIGACTSIGMOD-SIGART Symposium on Principles of Database Systems’.

Bushra, R., Aslam, N. and Khan, A.Y. (2011) ‘Food-drug interactions’, *Oman medical journal*, 26(2), pp. 77–83.

Campos, D., Matos, S. and Oliveira, J. L. (2012). Biomedical named entity recognition: a survey of machine-learning tools. *Theory and Applications for Advanced Text Mining*, 11, 175-195.

Ceran (2008) ‘Antik Mısır ve eski Anadolu uygarlıklarında tıp’, Selçuk Üniversitesi Sosyal Bilimler Enstitüsü Yüksek Lisans Tezi.

Cherkassky, V. (1997) ‘The Nature Of Statistical Learning Theory~’, *IEEE Transactions on Neural Networks*, pp. 1564–1564. Available at: <https://doi.org/10.1109/tnn.1997.641482>.

Chinchor, N. and Robinson, P. (1997). MUC-7 named entity task definition. In *Proceedings of the 7th Conference on Message Understanding* (Vol. 29, pp. 1-21).

Chiticariu, L., Li, Y. and Reiss, F.R. (2013) ‘Rule-Based Information Extraction is Dead! Long Live Rule-Based Information Extraction Systems!’, in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA: Association for Computational Linguistics, pp. 827–832.

KAYNAKLAR DİZİNİ (devam)

- Cohen, K.B., Christiansen, T. and Hunter, L.E.** (2011) ‘Parenthetically speaking: classifying the contents of parentheses for text mining’, *AMIA ... Annual Symposium proceedings / AMIA Symposium*, 2011, pp. 267–272.
- Cohen, K.B. and Demner-Fushman, D.** (2014) *Biomedical Natural Language Processing*. John Benjamins Publishing Company.
- Cupp, M.J.** (1999) ‘Herbal remedies: adverse effects and drug interactions’, *American family physician*, 59(5), pp. 1239–1245.
- DataTurks annotation tool**, <https://docs.dataturks.com>, (Erişim tarihi: 16 Ağustos 2022).
- Devlin, J. et al.** (2018) ‘BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding’, *arXiv [cs.CL]*. Available at: <http://arxiv.org/abs/1810.04805>.
- Doc spaCy API documentation**, <https://spacy.io/api/doc> (Erişim tarihi: 16 Ekim 2022).
- Doccano Annotation tool**, <https://doccano.herokuapp.com>, (Erişim tarihi: 16 Ağustos 2022).
- DrugBank**, <https://go.drugbank.com/> (Erişim tarihi: 6 Ekim 2022).
- Ekor, M.** (2014) ‘The growing use of herbal medicines: issues relating to adverse reactions and challenges in monitoring safety’, *Frontiers in pharmacology*, 4, p. 177.
- Feldman, R. and Sanger, J.** (2007) *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press.
- Fiske, D.W.** (1859) *The Book of the first American Chess Congress: Containing the Proceedings of that celebrated Assemblage, held in New York, in the Year 1857, By Daniel Willard Fiske*. Rudd & Carleton.
- FoodDB** (2022), <https://foodb.ca/> (Erişim tarihi: 6 Ekim 2022).
- Fugh-Berman, A.** (2000) ‘Herb-drug interactions’, *The Lancet*, 355(9198), pp. 134–138.
- Furrer, L. et al.** (2019) ‘OGER++: hybrid multi-type entity recognition’, *Journal of cheminformatics*, 11(1), p. 7.
- Gharehchopogh, F.S. and Khalifelu, Z.A.** (2011) ‘Analysis and evaluation of unstructured data: text mining versus natural language processing’, in *2011 5th International Conference on Application of Information and Communication Technologies (AICT)*, pp. 1–4.
- Giannenas, I. et al.** (2020) ‘Chapter 1 - The history of herbs, medicinal and aromatic plants, and their extracts: Past, current situation and future perspectives’, in P. Florou-Paneri, E. Christaki, and I. Giannenas (eds) *Feed Additives*. Academic Press, pp. 1–18.
- Google Colab**, <https://colab.research.google.com>, (Erişim tarihi: 6 Ekim 2022).

KAYNAKLAR DİZİNİ (devam)

- Goyal, A. et al.** (2018) ‘Recent Named Entity Recognition and Classification techniques: A systematic review’, *Computer Science Review*, 29, pp. 21–43.
- Grishman, R.** (2015) ‘Information Extraction’, *IEEE intelligent systems*, 30(5), pp. 8–15.
- Haddadian, K. et al.** (2014) ‘A review of Plantago plant’, *Indian journal of traditional knowledge*, <http://nopr.niscpr.res.in/handle/123456789/29518> (Erişim tarihi: 6 Eylül 2022).
- Hahn, U. et al.** (2012) ‘Active Learning-based corpus annotation--the PathoJen experience’, *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2012, pp. 301–310.
- Hänig, C. et al.** (2010) ‘Comparison of structured vs. unstructured data for industrial quality analysis.’ In *Proceedings of the World Congress on Engineering and Computer Science (Vol. 1, pp. 20-22)*.
- Hao, M. et al.** (2011) ‘Deglycosylated ginsenosides are more potent inducers of CYP1A1, CYP1A2 and CYP3A4 expression in HepG2 cells than glycosylated ginsenosides’, *Drug metabolism and pharmacokinetics*, 26(2), pp. 201–205.
- HAYIRLIDAĞ, M.** (2021) Mısır Tıbbının Gizemi Papirüsler. *Akademik Tarih ve Araştırmalar Dergisi*, 4(4), 68-86.
- Health Organization, W.** (2007) *WHO guidelines for assessing quality of herbal medicines with reference to contaminants and residues*. apps.who.int. Available at: <https://apps.who.int/iris/bitstream/handle/10665/43510/?sequence=1> (Erişim tarihi: 1 Eylül 2022).
- Hodges, P.J. and Kam, P.C.A.** (2002) ‘The peri-operative implications of herbal medicines’, *Anaesthesia*, 57(9), pp. 889–899.
- Hogle, B.C. et al.** (2018) ‘PXR as a mediator of herb-drug interaction’, *Journal of Food and Drug Analysis*, 26(2S), pp. S26–S31.
- Howard, J. and Ruder, S.** (2018) ‘Universal Language Model Fine-tuning for Text Classification’, *arXiv [cs.CL]*. Available at: <http://arxiv.org/abs/1801.06146>.
- Huang, K. et al.** (2019) ‘ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission’, *arXiv [cs.CL]*. Available at: <http://arxiv.org/abs/1904.05342>.
- Hu, Z. et al.** (2005) ‘Herb-drug interactions: a literature review’, *Drugs*, 65(9), pp. 1239–1282.
- Jamshidi-Kia, F. et al.** (2018) ‘Medicinal plants: Past history and future perspective’, *Journal of Herbmed Pharmacology*, 7(1), pp. 1–7.
- Jones, F.A.** (1996) ‘Herbs--useful plants. Their role in history and today’, *European journal of gastroenterology & hepatology*, 8(12), pp. 1227–1231.
- Kaya** (2010) ‘Tıbbi bitkiler ve etnobotanik çalışmalar’, *Bitkilerle tedavi sempozyumu*.
- Khare, R. et al.** (2016) ‘Crowdsourcing in biomedicine: challenges and opportunities’, *Briefings in bioinformatics*, 17(1), pp. 23–32.

KAYNAKLAR DİZİNİ (devam)

- Kim, Y.-M. and Lee, T.-H.** (2020) ‘Korean clinical entity recognition from diagnosis text using BERT’, *BMC medical informatics and decision making*, 20(Suppl 7), p. 242.
- Klinger, R. et al.** (2008) ‘Detection of IUPAC and IUPAC-like chemical names’, *Bioinformatics*, 24(13), pp. i268–76.
- Lafferty, J. et al.** (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.
- Lau, A.-J. et al.** (2009) ‘Antiplatelet and anticoagulant effects of Panax notoginseng: comparison of raw and steamed Panax notoginseng with Panax ginseng and Panax quinquefolium’, *Journal of ethnopharmacology*, 125(3), pp. 380–386.
- Lauriola, I., Lavelli, A. and Aiolfi, F.** (2022) ‘An introduction to Deep Learning in Natural Language Processing: Models, techniques, and tools’, *Neurocomputing*, 470, pp. 443–456.
- Liu, Y. et al.** (2019) ‘RoBERTa: A Robustly Optimized BERT Pretraining Approach’, *arXiv [cs.CL]*. Available at: <http://arxiv.org/abs/1907.11692>.
- Mandel, M.A.** (2006) ‘Integrated Annotation of Biomedical Text : Creating the PennBioIE Corpus’, *Proceedings of the Workshop on Text Mining, Ontologies and Natural Language Processing in Biomedicine*, Manchester, UK, 2006 . Available at: <https://ci.nii.ac.jp/naid/10021372252/> (Erişim: 5 Ekim 2022).
- Mansouri, A. et al.** (2008). Named entity recognition approaches. *International Journal of Computer Science and Network Security*, 8(2), 339-344.
- McCallum, A. et al.** (2000). Maximum entropy Markov models for information extraction and segmentation. In *Icml* (Vol. 17, No. 2000, pp. 591-598).
- Mechler, I.** (2017) Amazon mechanical Turk: Start to make money online. North Charleston, SC: Createspace Independent Publishing Platform.
- Memorial Sloan Cancer Center**, <https://www.mskcc.org/cancer-care/diagnosis-treatment/symptom-management/integrative-medicine/herbs/search> (Erişim tarihi : 1 Eylül 2022)
- Mikolov, T. et al.** (2013) ‘Efficient Estimation of Word Representations in Vector Space’, *arXiv [cs.CL]*. Available at: <http://arxiv.org/abs/1301.3781>.
- Miller, F.P. et al.** (2010) *Foldit*. Alphascript Publishing.
- Monica, N. and Kumar, K. R.** (2013). Survey on big data by coordinating mapreduce to integrate variety of data. *International Journal of Science and Research (IJSR) ISSN (Online)*, 2319, 7064.
- Naderi, N. et al.** (2011) ‘OrganismTagger: detection, normalization and grounding of organism entities in biomedical documents’, *Bioinformatics*, 27(19), pp. 2721–2729.
- Narayanaswamy, m. Et al.** (2002) ‘A biological named entity recognizer’, in *biocomputing 2003*. World scientific, pp. 427–438.

KAYNAKLAR DİZİNİ (devam)

Niu, J., Straubinger, R.M. and Mager, D.E. (2019) ‘Pharmacodynamic Drug-Drug Interactions’, *Clinical pharmacology and therapeutics*, 105(6), pp. 1395–1406.

Oga, E.F. et al. (2016) ‘Pharmacokinetic Herb-Drug Interactions: Insight into Mechanisms and Consequences’, *European journal of drug metabolism and pharmacokinetics*, 41(2), pp. 93–108.

Ohta, T. et al. (2012) ‘Open-domain Anatomical Entity Mention Detection’, in *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*. Jeju Island, Korea: Association for Computational Linguistics, pp. 27–36.

Ong, C.E. and Pan, Y. (2014) ‘Drug-herb interactions: Mechanisms involved and clinical implications of five commonly and traditionally used herbs’, *CELLMED*, 4(3), pp. 17.1–17.8.

Özer, A. (2010) *Bazı tıbbi bitkilerin sıcak havalı kurutucuda kurutulması ve kurutma sıcaklıklarının ürün kalitesi üzerine etkileri*. Namık Kemal Üniversitesi. Available at: <http://acikerisim.nku.edu.tr/xmlui/handle/20.500.11776/538> (Erişim tarihi: 31 Ağustos 2022).

Parvez, M.K. and Rishi, V. (2019) ‘Herb-Drug Interactions and Hepatotoxicity’, *Current drug metabolism*, 20(4), pp. 275–282.

Patel, J. et al. (2004) ‘In vitro interaction of the HIV protease inhibitor ritonavir with herbal constituents: changes in P-gp and CYP3A4 activity’, *American journal of therapeutics*, 11(4), pp. 262–277.

Peng, J. et al. (2020) ‘Natural language processing (NLP) tools in extracting biomedical concepts from research articles: a case study on autism spectrum disorder’, *BMC medical informatics and decision making*, 20(Suppl 11), p. 322.

Perera, N. et al. (2020) ‘Named Entity Recognition and Relation Detection for Biomedical Information Extraction’, *Frontiers in cell and developmental biology*, 8, p. 673.

Posadzki, P. et al. (2013) ‘Herb-drug interactions: an overview of systematic reviews’, *British journal of clinical pharmacology*, 75(3), pp. 603–618.

Prabavathy, S.K. and Sumathi, P. (2014) ‘Dictionary and Gene Ontology Based Similarity for Named Entity Relationship Proteinprotein Interaction Prediction from Biotext Corpus’, *Research Journal of Applied Sciences, Engineering and Technology*, 8(22), pp. 2282–2289.

Prodigy · An annotation tool for AI, <https://prodi.gy/> (Erişim tarihi: 7 ekim 2022).

Quimbaya, A.P. et al. (2016) ‘Named Entity Recognition Over Electronic Health Records Through a Combined Dictionary-based Approach’, *Procedia computer science*, 100, pp. 55–61.

Rehman, S.U. et al. (2015) ‘Interactions between herbs and antidiabetics: an overview of the mechanisms, evidence, importance, and management’, *Archives of pharmacal research*, 38(7), pp. 1281–1298.

Ruokolainen, T. et al. (2020) ‘A Finnish news corpus for named entity recognition’, *Language resources and evaluation*, 54(1), pp. 247–272.

KAYNAKLAR DİZİNİ (devam)

- Sammartino, J. C. et al.** (2016). Annotation Process, Guidelines and Text Corpus of Small Non-Coding RNA Molecules: the MiNCor for MicroRNA Annotations. In SMBM (pp. 56-63).
- Sanh, V. et al.** (2019) ‘DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter’, *arXiv [cs.CL]*. Available at: <http://arxiv.org/abs/1910.01108>.
- Şarıßen, Ö. ve Çalışkan, D.** (2005). Fitoterapi: bitkilerle tedaviye dikkat (!). *Sted*, 14(8), 182-187.
- Schippmann, U. et al.** (2006) ‘A comparison of cultivation and wild collection of medicinal and aromatic plants under sustainability aspects’, *Frontis*, pp. 75–95.
- Scott, G.N. and Elmer, G.W.** (2002) ‘Update on natural product–drug interactions’, *American journal of health-system pharmacy: AJHP: official journal of the American Society of Health-System Pharmacists*. Available at: <https://academic.oup.com/ajhp/article-abstract/59/4/339/5157965>.
- Sewell and Rafieian-Kopaei** (2014) ‘The history and ups and downs of herbal medicines usage’, *Journal of Herbmed Pharmacology*. Available at: <http://eprints.skums.ac.ir/4716/1/14.pdf>.
- Shaalan, K.** (2010). Rule-based approach in Arabic natural language processing. *The International Journal on Information and Communication Technologies (IJICT)*, 3(3), 11-19.
- Shaikh, A.S. et al.** (2020) ‘Herb-drug interaction studies of herbs used in treatment of cardiovascular disorders-A narrative review of preclinical and clinical studies’, *Phytotherapy research: PTR*, 34(5), pp. 1008–1026.
- Song, H.-J. et al.** (2018) ‘Comparison of named entity recognition methodologies in biomedical documents’, *Biomedical engineering online*, 17(Suppl 2), p. 158.
- Song, M., Yu, H. and Han, W.-S.** (2015) ‘Developing a hybrid dictionary-based bio-entity recognition technique’, *BMC medical informatics and decision making*, 15 Suppl 1, p. S9.
- SpaCy industrial-strength Natural Language Processing in Python**, <https://spacy.io> (Erişim tarihi: 16 Ekim 2022).
- Stachelscheid, H. et al.** (2014) ‘CellFinder: a cell data repository’, *Nucleic acids research*, 42(Database issue), pp. D950–8.
- Sullivan, R. J.** (1989). England's “Age of Invention”: The acceleration of patents and patentable invention during the Industrial Revolution. *Explorations in Economic history*, 26(4), 424-452.
- Sutskever, I. et al.** (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.
- Swanston-Flatt, S.K. et al.** (1990) ‘Traditional plant treatments for diabetes. Studies in normal and streptozotocin diabetic mice’, *Diabetologia*, 33(8), pp. 462–464.

KAYNAKLAR DİZİNİ (devam)

Takeda, T. et al. (2017) ‘Predicting drug–drug interactions through drug structural similarities and interaction networks incorporating pharmacokinetics and pharmacodynamics knowledge’, *Journal of Cheminformatics*. Available at: <https://doi.org/10.1186/s13321-017-0200-8>.

Tanabe, L. et al. (2005) ‘GENETAG: a tagged corpus for gene/protein named entity recognition’, *BMC bioinformatics*, 6 Suppl 1, p. S3.

Titz, A. (2004). Policy, Research & Development and Commercialisation Strategies, Scope for Diversified and Sustainable Extraction. *Bangalore, India*, 72-80.

Tok2Vec, <https://spacy.io/api/tok2vec> (Erişim tarihi: 11 Ekim 2022).

Van Tongeren, J.M.Z. et al. (2020) ‘The Development of Practice Recommendations for Drug-Disease Interactions by Literature Review and Expert Opinion’, *Frontiers in pharmacology*, 11, p. 707.

SpaCy usage documentation, Training Pipelines & Models, <https://spacy.io/usage/training> (Erişim tarihi: 17 Ekim 2022).

Trinh, K., Pham, D. and Le, L. (2018) ‘Semantic Relation Extraction for Herb-Drug Interactions from the Biomedical Literature Using an Unsupervised Learning Approach’, in *2018 IEEE 18th International Conference on Bioinformatics and Bioengineering (BIBE)*, pp. 334–337.

Vajjala, S. et al. (2020) *Practical Natural Language Processing: A Comprehensive Guide to Building Real-world NLP Systems*. O’Reilly Media.

Vaswani, A. et al. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, H. et al. (2008) ‘Biomedical Named Entity Recognition Based on Classifiers Ensemble’, *Int. J. Comput. Sci. Appl.*, 5(2), pp. 1–11.

Wang, K. et al. (2013) ‘Prediction of drug-target interactions for drug repositioning only based on genomic expression similarity’, *PLoS computational biology*, 9(11), p. e1003315.

Wang, R. et al. (2016) ‘Effect of Ginkgo Leaf Tablets on the Pharmacokinetics of Amlodipine in Rats’, *European journal of drug metabolism and pharmacokinetics*, 41(6), pp. 825–833.

Wikipedia contributors (2022) *Mechanical Turk*, *Wikipedia, The Free Encyclopedia*. Available at: https://en.wikipedia.org/w/index.php?title=Mechanical_Turk&oldid=1111895458.

Wissler et al. (2014) ‘The Gold Standard in Corpus Annotation’, *IEEE GSC*, https://www.academia.edu/download/43756030/ieeegsc2014_submission_3.pdf.

Xintong, G. et al. (2014) ‘Brief survey of crowdsourcing for data mining’, *Expert systems with applications*, 41(17), pp. 7987–7994.

Xu, J. et al. (2019) ‘Potential Pharmacokinetic Herb-Drug Interactions: Have we Overlooked the Importance of Human Carboxylesterases 1 and 2?’, *Current drug metabolism*, 20(2), pp. 130–137.

KAYNAKLAR DİZİNİ (devam)

- Yamazaki, Y. et al.** (2008) ‘High dose of commercial products of kava (*Piper methysticum*) markedly enhanced hepatic cytochrome P450 1A1 mRNA expression with liver enlargement in rats’, *Food and chemical toxicology: an international journal published for the British Industrial Biological Research Association*, 46(12), pp. 3732–3738.
- Yeung, K.S. et al.** (2018) ‘Herb-Drug Interactions in Cancer Care’, *Oncology* , 32(10), pp. 516–520.
- Zhang, Y. et al.** (2018) ‘Drug-drug interaction extraction via hierarchical RNNs on sequence and shortest dependency paths’, *Bioinformatics* , 34(5), pp. 828–835.
- Zhang, Y. et al.** (2021) ‘Biomedical and clinical English model packages for the Stanza Python NLP library’, *Journal of the American Medical Informatics Association: JAMIA*, 28(9), pp. 1892–1899.
- Zhang, Y. et al.** (2022) ‘Overview of Current Herb-Drug Interaction Databases’, *Drug metabolism and disposition: the biological fate of chemicals*, 50(1), pp. 86–94.
- Zhou, Z. H.** (2018), A brief introduction to weakly supervised learning. *National science review*, 5(1), 44-53.
- Zhu, Q. et al.** (2018) ‘GRAM-CNN: a deep learning approach with local context for named entity recognition in biomedical text’, *Bioinformatics* , 34(9), pp. 1547–1554.
- Z. Msomi, N. and B.C. Simelane, M.** (2019) ‘Herbal Medicine’, in *Herbal Medicine*. IntechOpen.

TEŞEKKÜR

Tez çalışmam sırasında bana her konuda yardımcı olan değerli danışmanım Dr. Öğretim Üyesi Özgür Gümüş'e ve tezin her aşamasında yüksek katkıları olan Hollanda Maastricht Üniversitesinden Asistan Prof. Remzi Çelebi'ye teşekkürlerimi sunarım. Tez izleme komitesi toplantılarındakiengin bilgi, tecrübe ve önerileri ile tez çalışmasına katkıda bulunan Prof. Dr. Oğuz Dikenelli'ye ayrıca teşekkür ederim. Bu tez çalışmasının literatüre kazandırılması için öğrencisi Araş. Gör. Kerem Tok ile alan bilgisini kullanarak saatler ve günler süren veri etiketleme gibi zahmetli bir süreçte yardımlarını aldığım Ege Üniversitesi Biyokimya Mühendisliği Bölümünden Prof. Dr. Figen Zihnioğlu'na ve öğrencisi Araş.Gör. Kerem Tok'a teşekkürü borç bilirim. Ayrıca tez çalışması sırasında fikir alışverişinde bulunduğum ve bu konuda yardımcı olan doktora öğrencisi Vahap Mustafapour ve Araş. Gör. Hazal Türkmen'e de teşekkürlerimi sunarım. Son olarak da teknik destek vererek katkıda bulunan öğrencimiz Ensar Emir Erol'a teşekkürlerimi sunarım.

Bu tez çalışmasını 10 Kasım 2010'da kaybettiğim babam Mustafa YAŞAR'a atfediyorum. İlginçtir ki kendisi bitkilerin her türlü hastalığı tedavi edici özelliğine ikna olmuş birisi idi ve onu kaybetmemizin en büyük nedenlerinden bu inancı olmuştur ve yıllar sonra akademiye girmek için yıllarca desteklediği çocuğu bu tez çalışmasında tesadüfi olarak bitkilerin ilaçlar ile alınması sonucu ortaya çıkan potansiyel olumsuzlukları üzerine bir doktora çalışması yapmıştır. Umut ediyorum bu çalışma, bitki ilaç etkileşimlerinden kaynaklı gelecekte meydana gelebilecek birçok olumsuzluğu önlemeye yardımcı olacaktır.

Benim bilim üzerine çalışmamı ve insanlığa faydalı bir çalışma gerçekleştirmem için çocukluğumdan akademik yaşamına girene kadar her zaman yanımda olan annem Gülşen YAŞAR'a ve tez çalışmasının imla ve yazım kuralları hataları için destek veren kardeşim Zelal Yaşar'a ve desteklerinden dolayı eşim Rakhat YAŞAR'a da teşekkürlerimi sunarım.

ÖZGEÇMİŞ

2016'dan önce ağ benzerliği yöntemlerini kullanarak veri ağları üzerine çalışmaktaydım. 2016'dan sonra veri bilimi ve yapay zeka üzerine çalışmalar yapmaktayım. Verinin temizlenmesi, verinin analizi ve veriden faydalı bilgilerin çıkarılması için yapay zeka ve doğal dil işleme tekniklerini kullanan sistemler üzerine çalışmaktayım. 3 uluslararası SCI indeksli yayın, 2 uluslararası konferans yayını, 2 ulusal yayın ve 1 ulusal projem vardır. Ayrıca 2017'de Hack4Health'tin düzenlediği sağlıkta yapay zeka uygulamaları üzerine Dünya 3.lüğüm vardır. Akademik olarak da Python programlama dili ile algoritmaya giriş ve Veri bilimi derslerinin asistanlığını yaptım. Son yıllarda mobil uygulamalar üzerine yönelimim vardır. Sağlık üzerine bir mobil uygulama ekibinin bir parçasıyım. Ayrıca Grafik tasarımı, Rusça Eğitimliği ve Sosyal Medya danışmanlığını hobi olarak yapmaktayım.

EĞİTİM

- Ege Üniversitesi Bilgisayar Mühendiliği (Doktora) 2016 - 2022
- Ege Üniversitesi Bilgisayar Mühendiliği (Yüksek Lisans) 2012 - 2015
- Manas Üniversitesi Bilgisayar Mühendiliği (Lisans) 2005 - 2010

YABANCI DİLLER

- **RUSÇA :** ÇOK İYİ
- **İNGİLİZCE:** ÇOK İYİ
- **KIRGIZCA:** ÇOK İYİ

Uluslararası SCI İndeksli Yayınlar

1. Evaluation of knowledge graph embedding approaches for drug-drug interaction prediction in realistic settings. BMC Bioinform. (2019) [Alıntılanma sayısı: 53](#)
2. Classification of olive oils using chromatography, principal component analysis and artificial neural network modelling (2018) [Alıntılanma sayısı: 16](#)
3. Comparison of different classification algorithms to identify geographic origins of olive oils (2020) [Alıntılanma sayısı: 8](#)

Uluslararası Konferanslar

1. Evaluation of knowledge graph embedding approaches for drug-drug interaction prediction using Linked Open Data (2018) [Alıntılanma sayısı: 12](#)
2. Prediction of Drug-Drug Interactions Using Pharmacological Similarities of Drugs (2015) [Alıntılanma sayısı: 9](#)
3. Link Prediction for Drug-Drug Interaction Network (2014)

Ulusal Yayınlar

1. A combined approach of base and meta learners for hybrid system (2022)
2. Voting Combinations-Based Ensemble: A Hybrid Approach (2022)